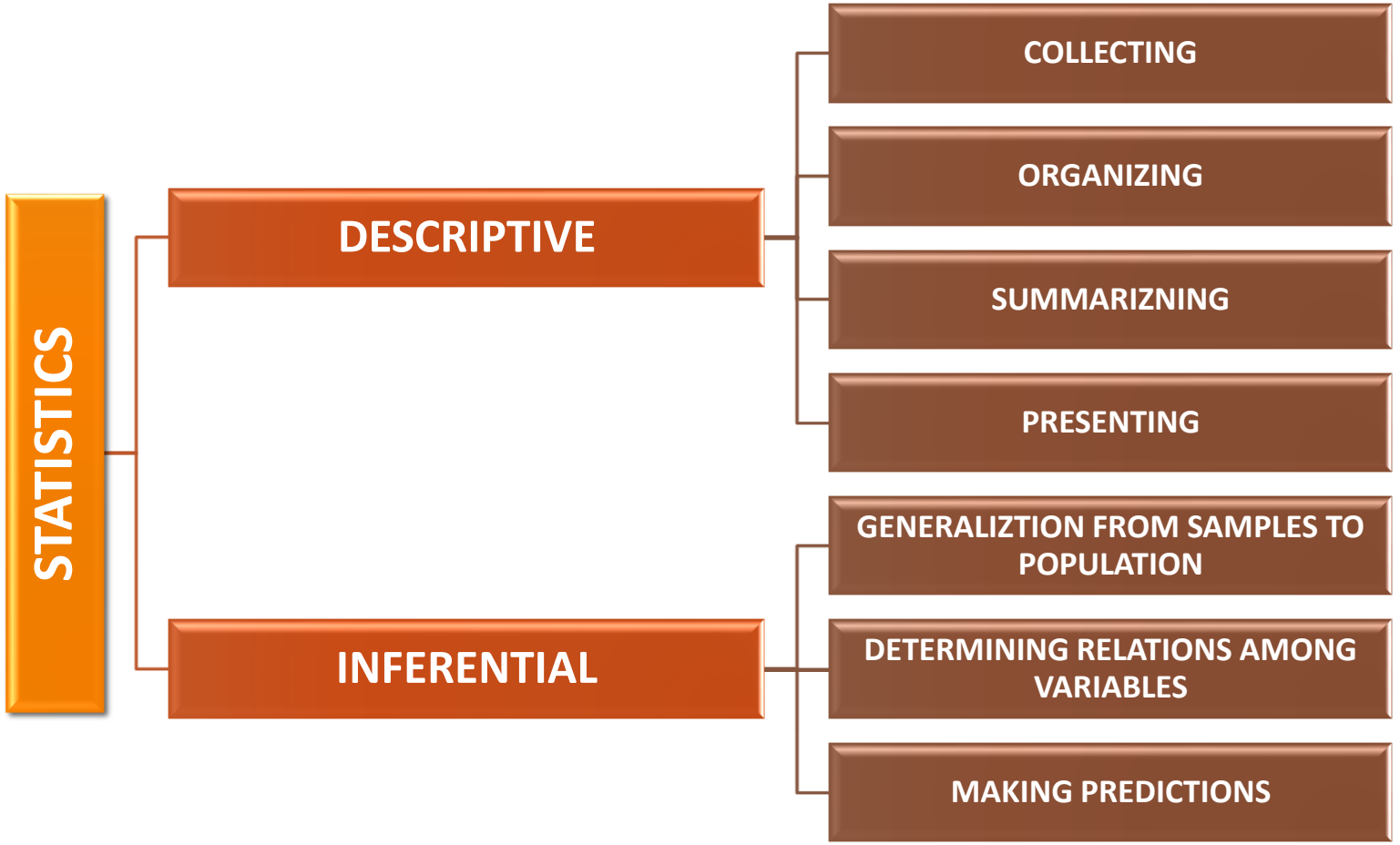
STATISTICAL ANALYSIS - LECTURE 2

Dr. Mahmoud Mounir

mahmoud.mounir@cis.asu.edu.eg

RE-CAP

INTRODUCTION



SAMPLING

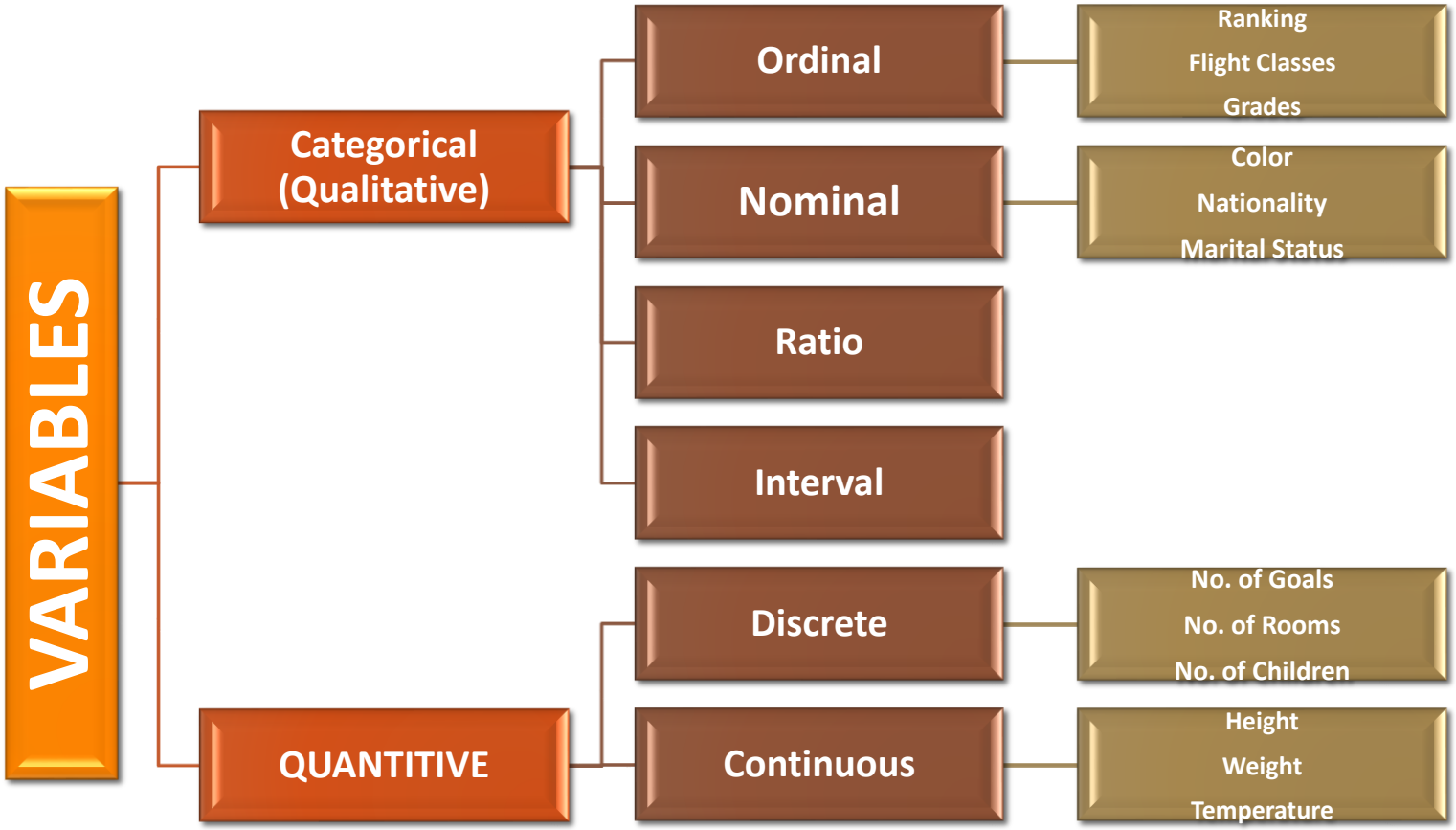
- **POPULATION SIZE = 30**

Street 1

Street 2

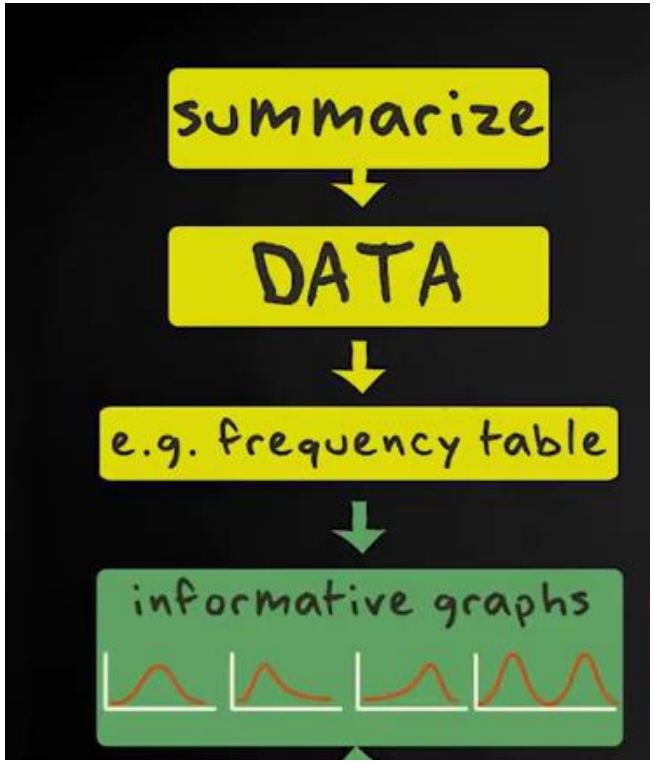
Street 3

LEVELS OF MEASUREMENTS



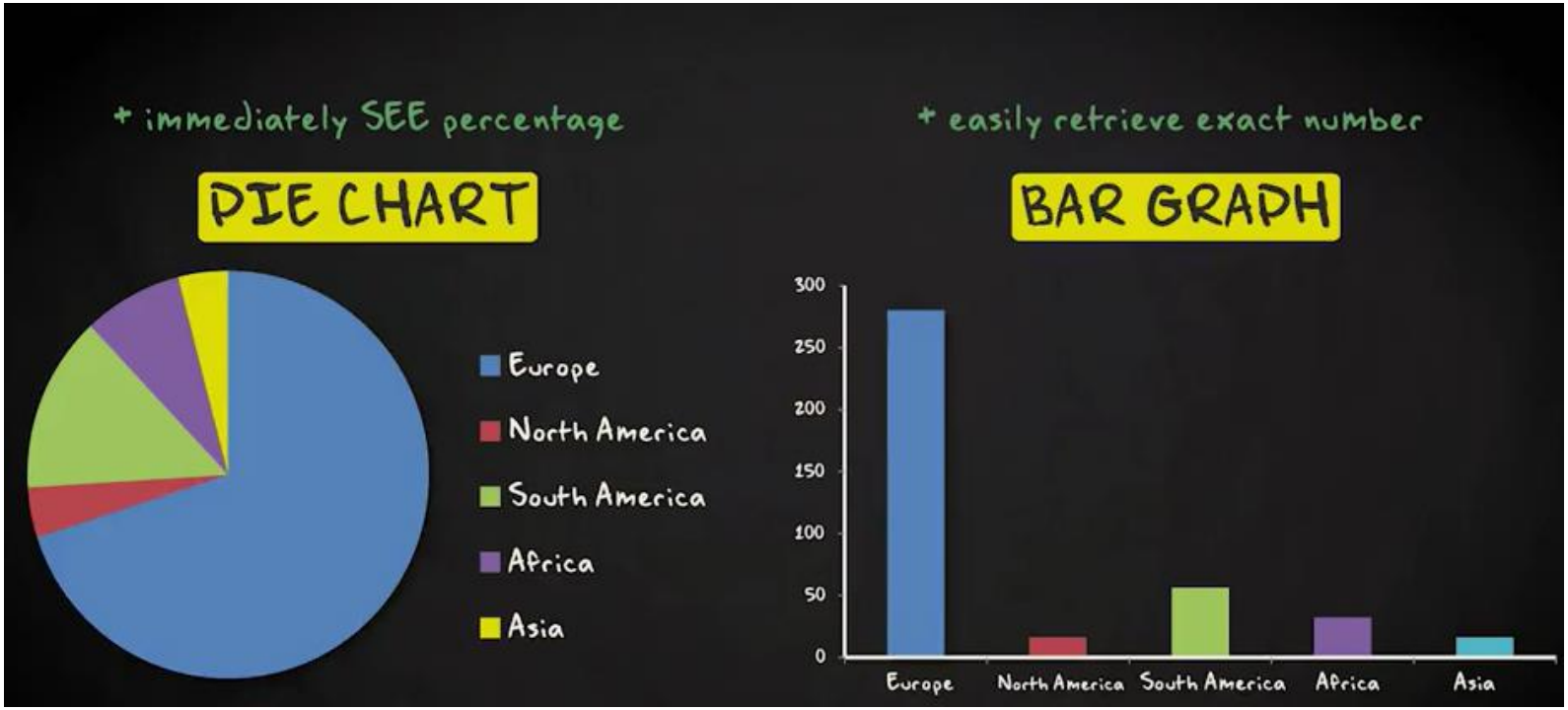
DATA MATRIX AND FREQUENCY TABLE

DATA MATRIX AND FREQUENCY TABLE



GRAPHS AND SHAPES OF DISTRIBUTIONS

GRAPHS AND SHAPES OF DISTRIBUTIONS



GRAPHS AND SHAPES OF DISTRIBUTIONS

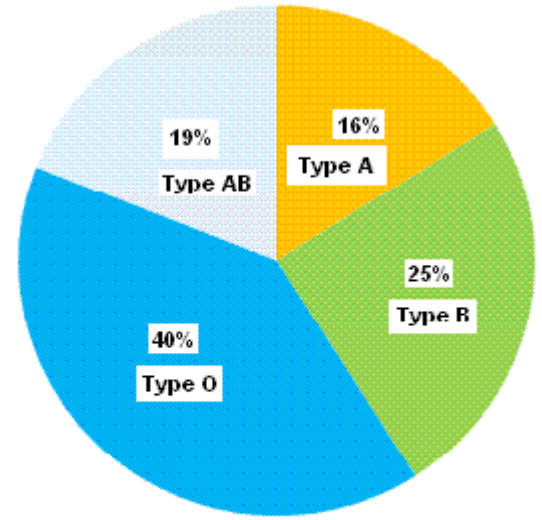
Example (1)

This pie chart below shows the percentages of blood types for a group of **200** people.

a) How many people, in this group, have blood type AB?

b) How many people, in this group, do not have blood type O?

c) How many people, in this group, have blood types A or B?



Blood Types for a group of 200 people

www.analyzemath.com

GRAPHS AND SHAPES OF DISTRIBUTIONS

Example (1) Solution

a) How many people, in this group, have blood type AB?

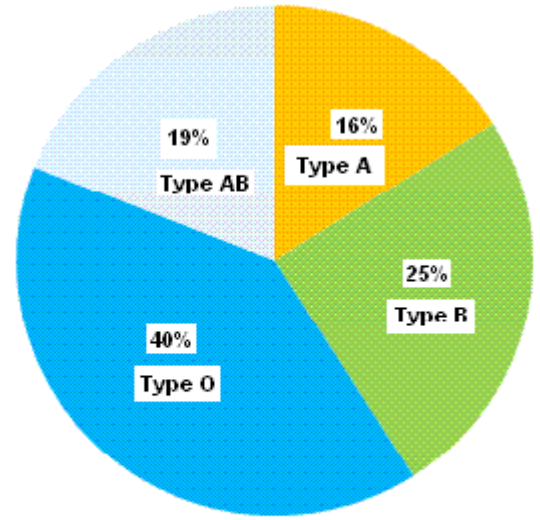
19% × 200 = 19 × 200 / 100 = 38 people

b) How many people, in this group, do not have blood type O?

(100% - 40%) × 200 = 60 × 200 / 100 = 120 people

c) How many people, in this group, have blood types A or B?

(16% + 25%) × 200 = 41 × 200 / 100 = 82 people



Blood Types for a group of 200 people

www.analyzemath.com

GRAPHS AND SHAPES OF DISTRIBUTIONS

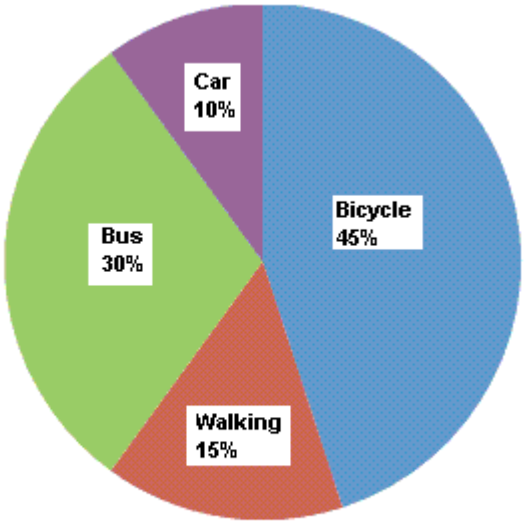
Example (2)

This pie chart shows the percentages of types of transportation used by **800** students to come to school.

a) How many students, in the school come to school by bicycle?

b) How many students do not walk to school?

c) How many students come to school by bus or in a car?



Types of Transportation

www.analyzemath.com

GRAPHS AND SHAPES OF DISTRIBUTIONS

Example (2) Solution

a) How many students, in the school, come to school by bicycle?

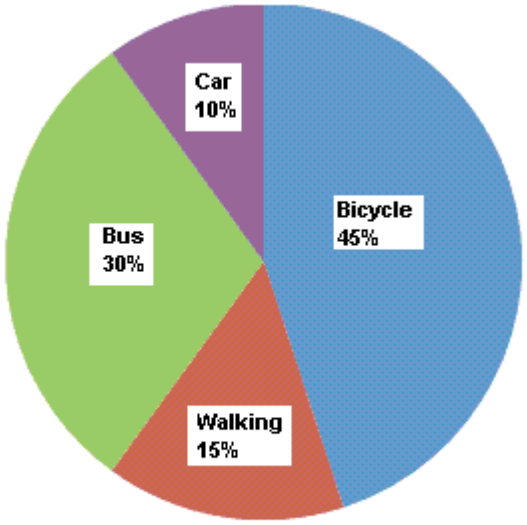
45% × 800 = 360 students

b) How many students do not walk to school?

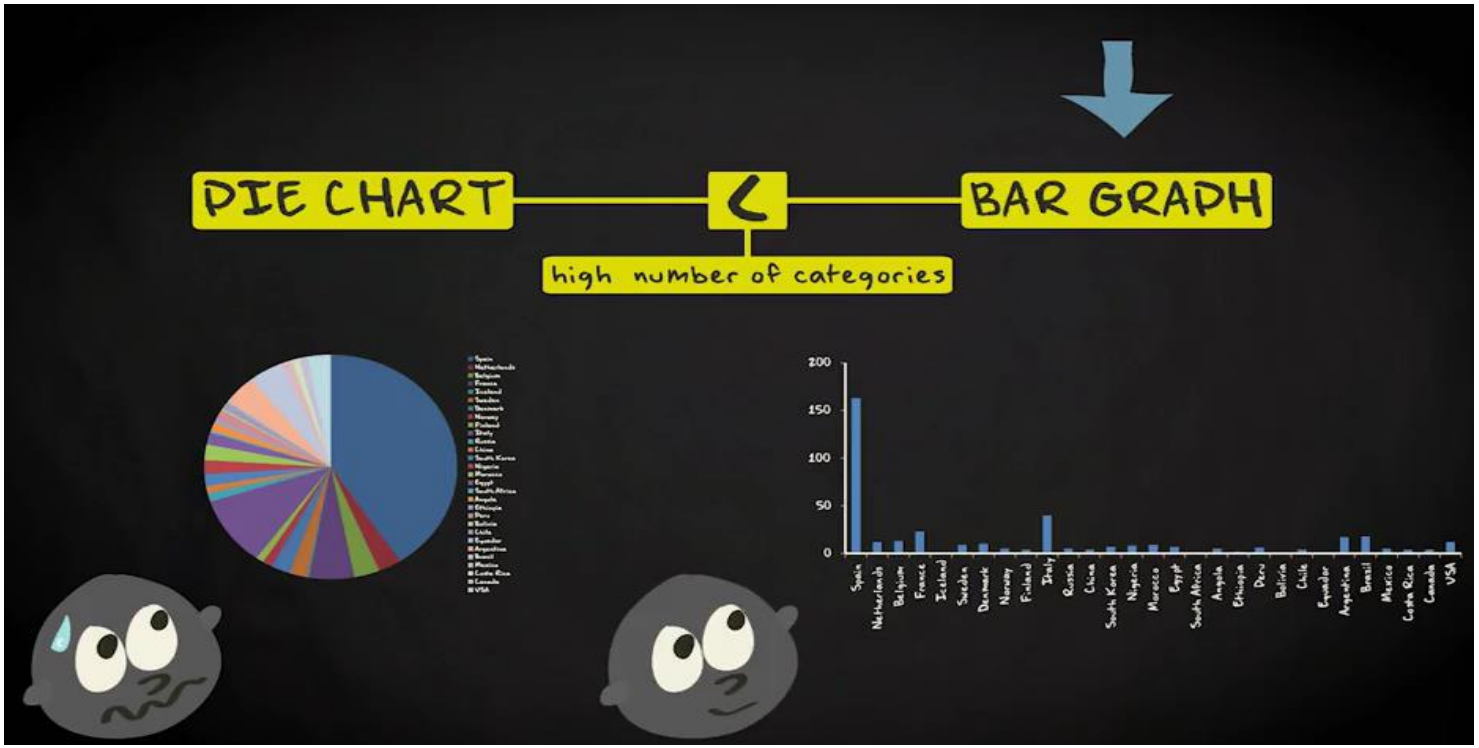
(100% - 15%) × 800 = 680 students

c) How many students come to school by bus or in a car?

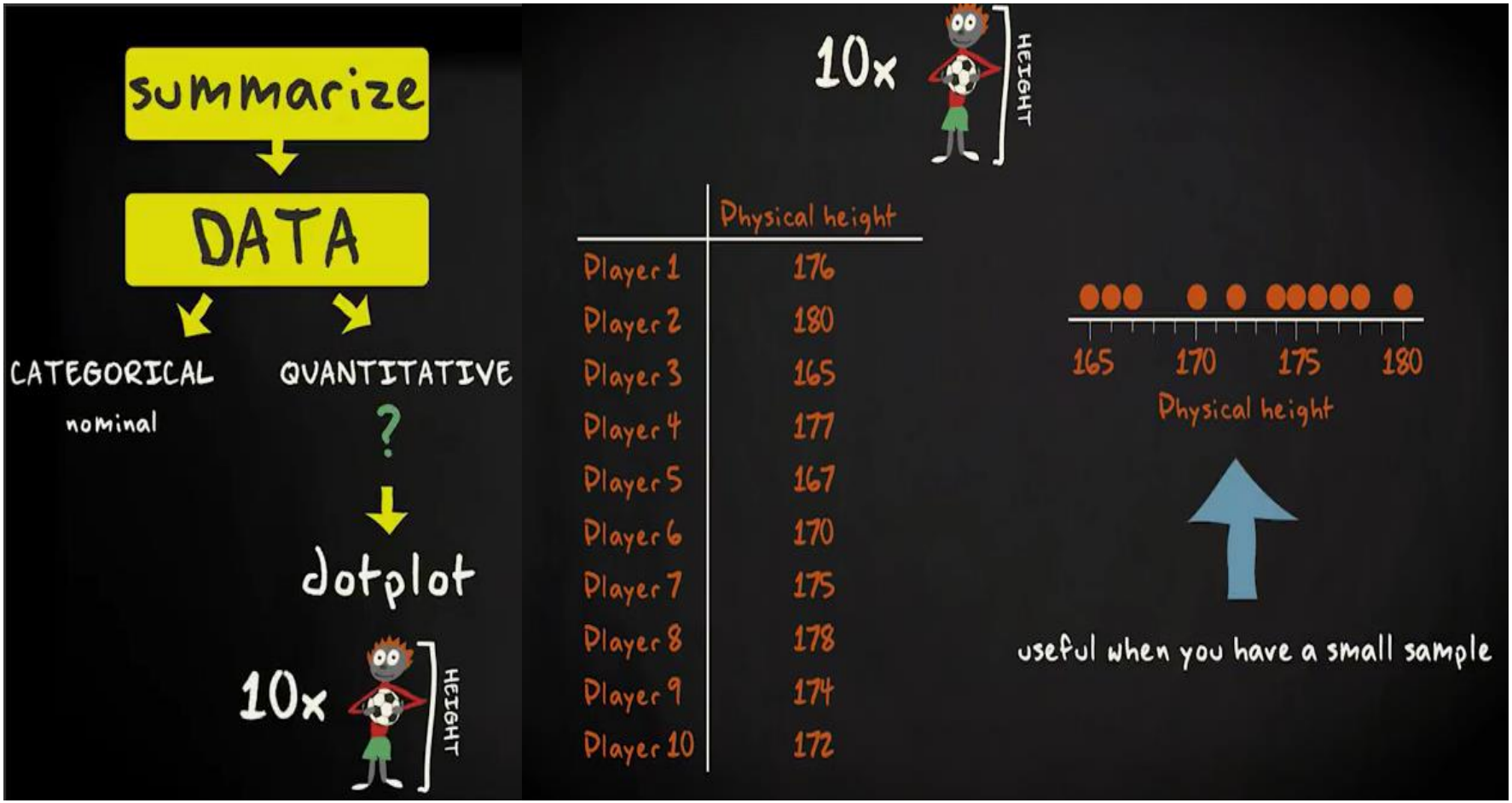
(30% + 10%) × 800 = 320 students



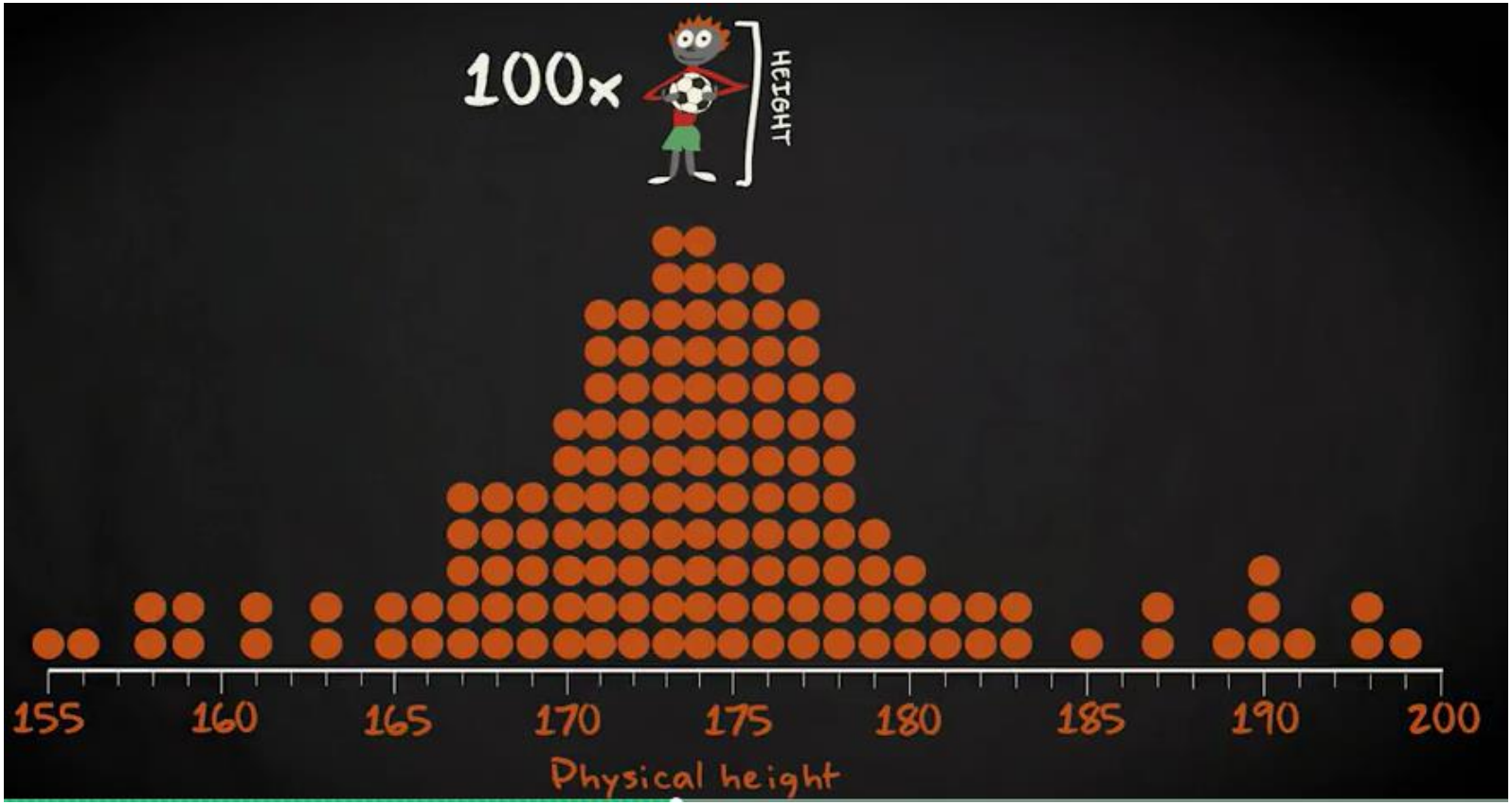
GRAPHS AND SHAPES OF DISTRIBUTIONS



GRAPHS AND SHAPES OF DISTRIBUTIONS



GRAPHS AND SHAPES OF DISTRIBUTIONS



GRAPHS AND SHAPES OF DISTRIBUTIONS

GROUPED FREQUENCY DISTRIBUTIONS

❑ Given the weight of 40 new born children measured in lbs.

58	118	92	108	132
32	140	138	96	161
120	86	115	118	95
83	112	128	127	124
123	134	94	67	124
155	105	100	112	141
104	132	98	146	132
93	85	94	116	113

- Min Value = 32
- Max Value = 161
- Range (R) = Max – Min
  = 161 – 32
  = 129

- No. of Classes = 1 + 3.3 log (n) = 1+ 3.3 log (40) = 6.29 ≈ 7

ROUND UP

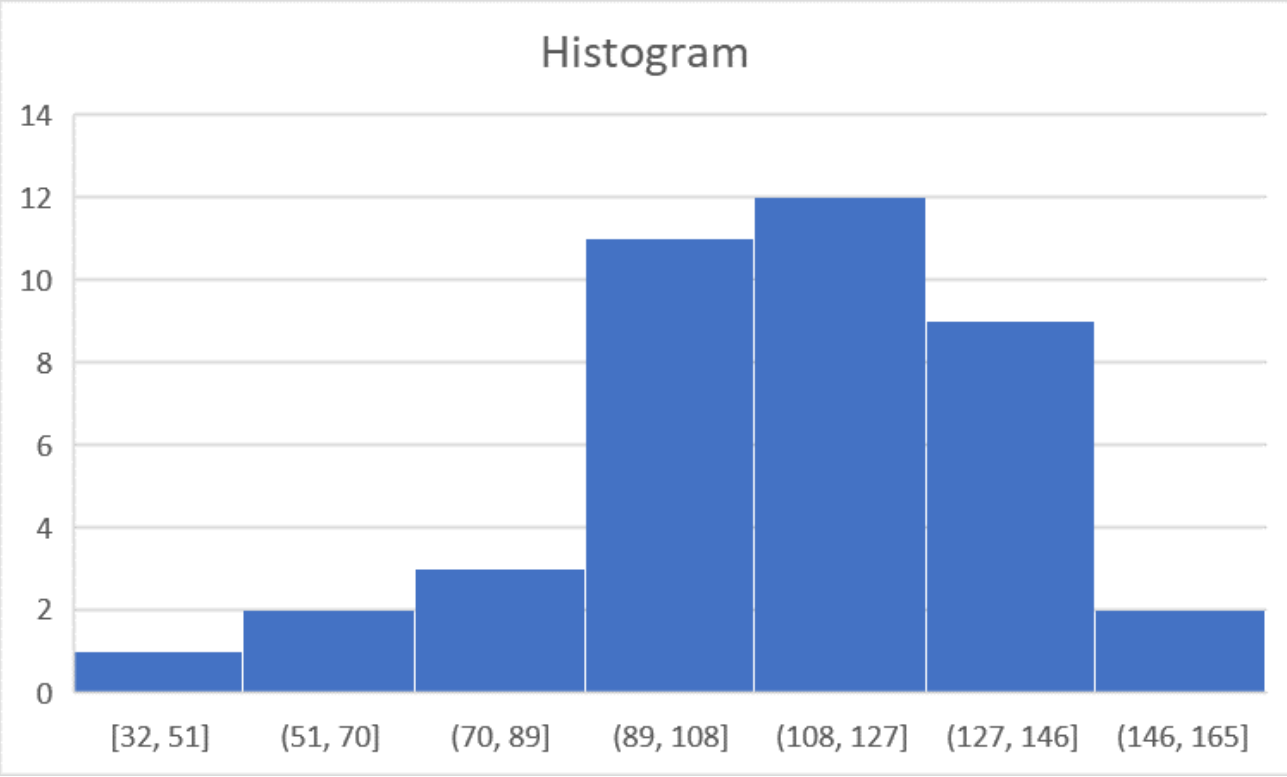
- OR No. of Classes = $\sqrt{n}$ = $\sqrt{40}$ = 6.32 = 7
- Class Width = Range / No. of Classes = 129 / 7 = 18.4 ≈ 19

GROUPED FREQUENCY DISTRIBUTIONS

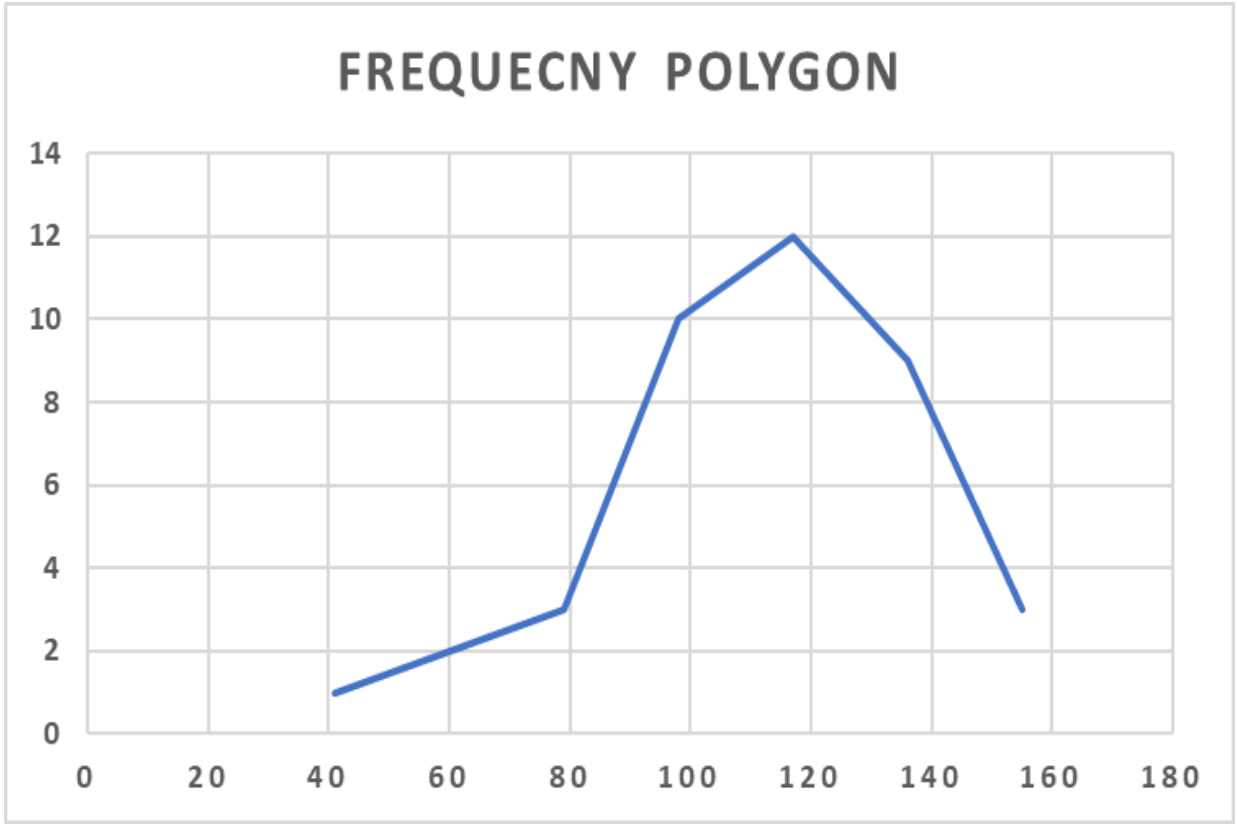
- Given the weight of 40 new born children measured in lbs.

Class Limits		Class Midpoint	Frequency	Relative Frequency
Lower Limit	Upper Limit			
32	51	41.5	1	2.50%
51	70	60.5	2	5.00%
70	89	79.5	3	7.50%
89	108	98.5	10	25.00%
108	127	117.5	12	30.00%
127	146	136.5	9	22.50%
146	165	155.5	3	7.50%
TOTAL			40	100.00%

GROUPED FREQUENCY DISTRIBUTIONS

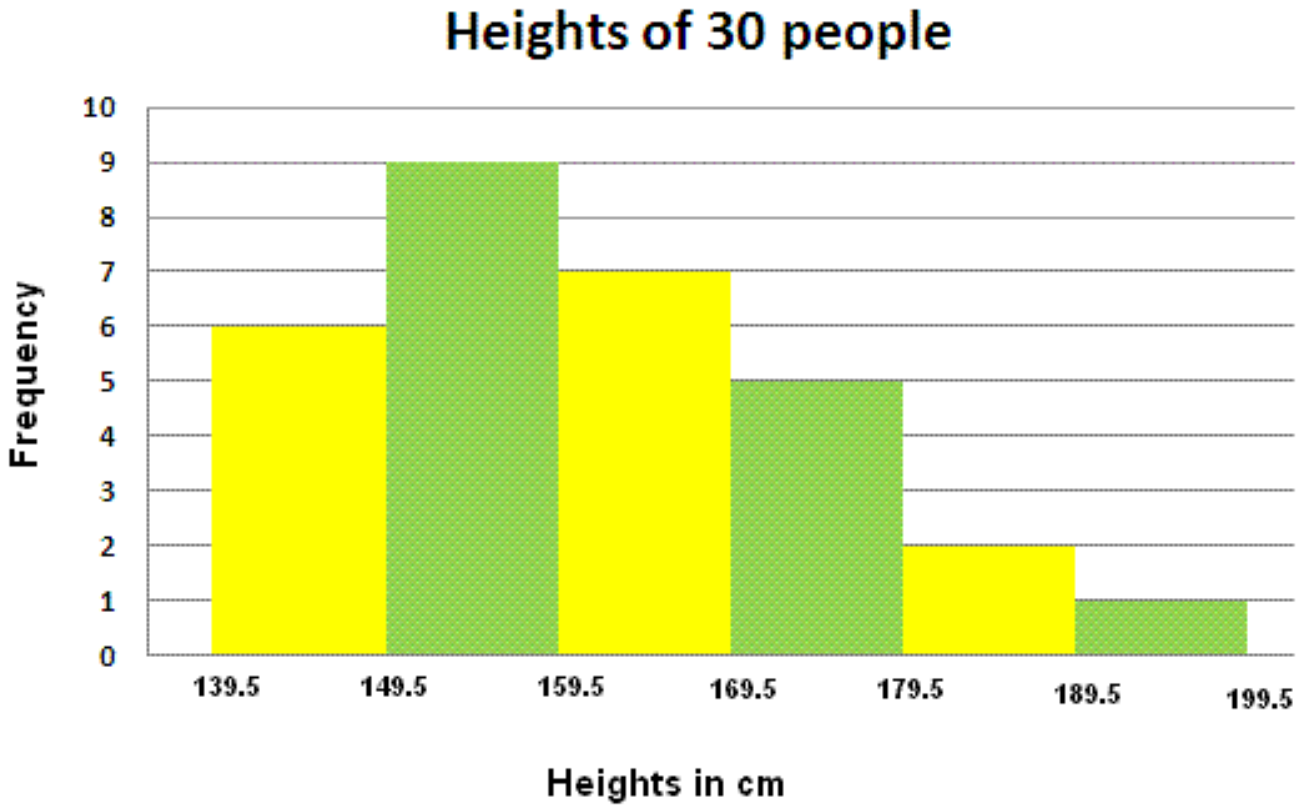


GROUPED FREQUENCY DISTRIBUTIONS



GROUPED FREQUENCY DISTRIBUTIONS

Example (3)



GROUPED FREQUENCY DISTRIBUTIONS

Example (3)

This histogram shows the heights (in cm) distribution of 30 people.

a) How many people have heights between 159.5 and 169.5 cm?

b) How many people have heights less than 159.5 cm?

c) How many people have heights more than 169.5 cm?

d) What percentage of people have heights between 149.5 and 179.5 cm?

GROUPED FREQUENCY DISTRIBUTIONS

Example (3) Solution

a) How many people have heights between 159.5 and 169.5 cm?

7 people

b) How many people have heights less than 159.5 cm?

9 + 6 = 15 people

c) How many people have heights more than 169.5 cm?

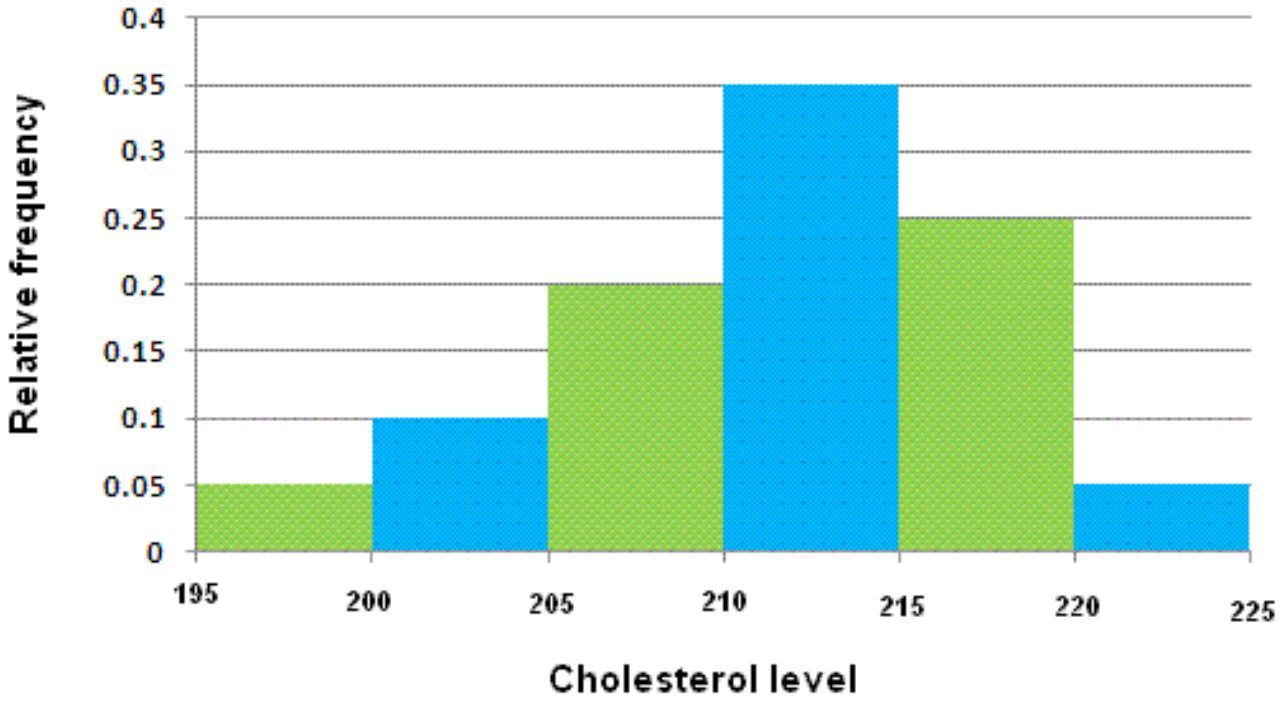
5 + 2 + 1 = 8 people

d) What percentage of people have heights between 149.5 and 179.5 cm?

(9 + 7 + 5)/30 = 0.7 = 70%

GROUPED FREQUENCY DISTRIBUTIONS

Example (4)



GROUPED FREQUENCY DISTRIBUTIONS

Example (4)

This histogram shows the level of cholesterol (in mg per dl) of 200 people.

a) How many people have a level of cholesterol between 205 and 210 mg per dl?

b) How many people have a level of cholesterol less than 205 mg per dl?

c) What percentage of people have a level of cholesterol more than 215 mg per dl?

d) How many people have a level of cholesterol between 205 and 220 mg per dl?

GROUPED FREQUENCY DISTRIBUTIONS

Example (4) Solution

a) How many people have a level of cholesterol between 205 and 210 mg per dl?

0.2*200 = 40 people

b) How many people have a level of cholesterol less than 205 mg per dl?

(0.05 + 0.1)*200 = 30 people

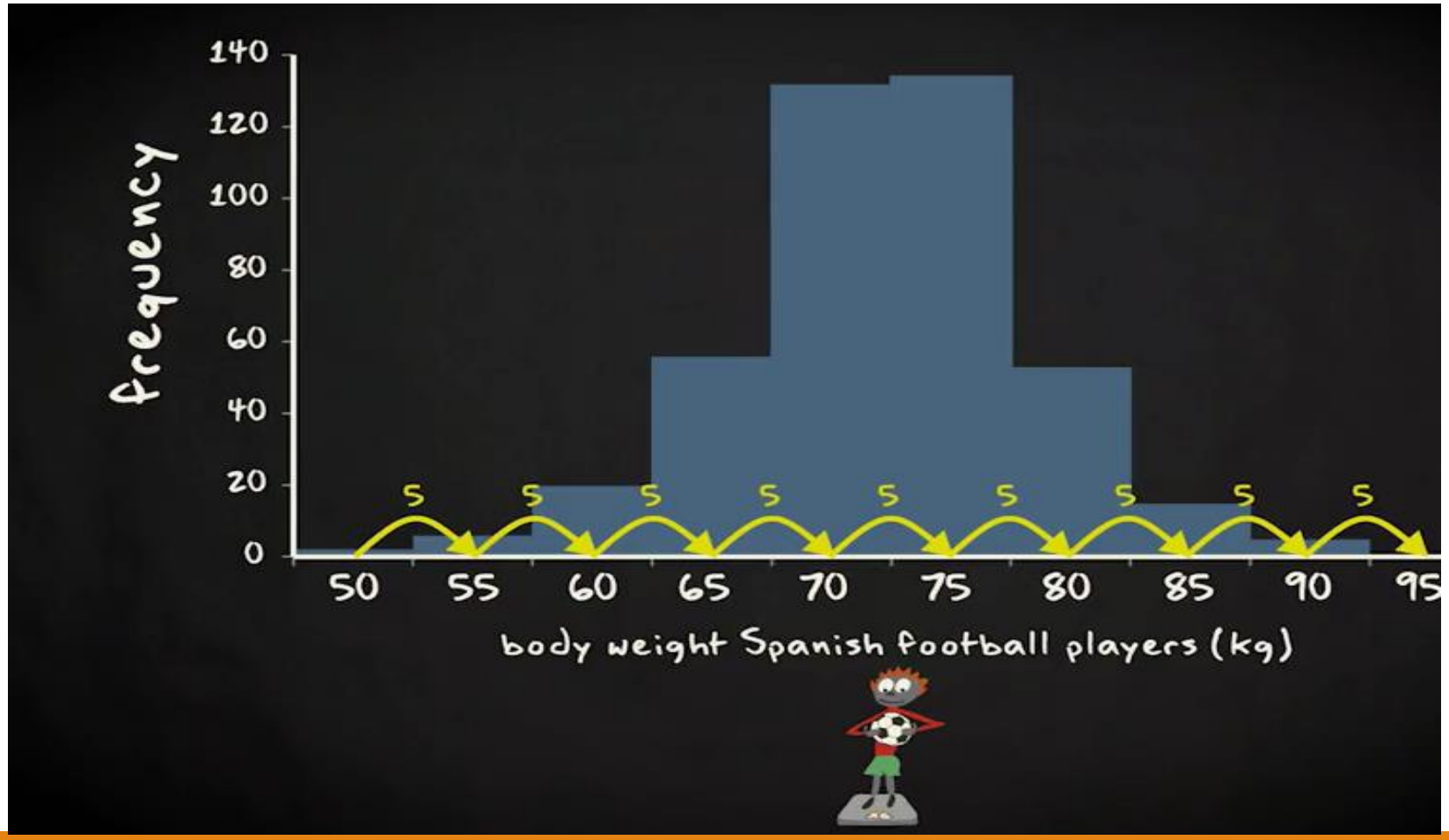
c) What percentage of people have a level of cholesterol more than 215 mg per dl?

(0.25 + 0.05) = 0.3 = 30%

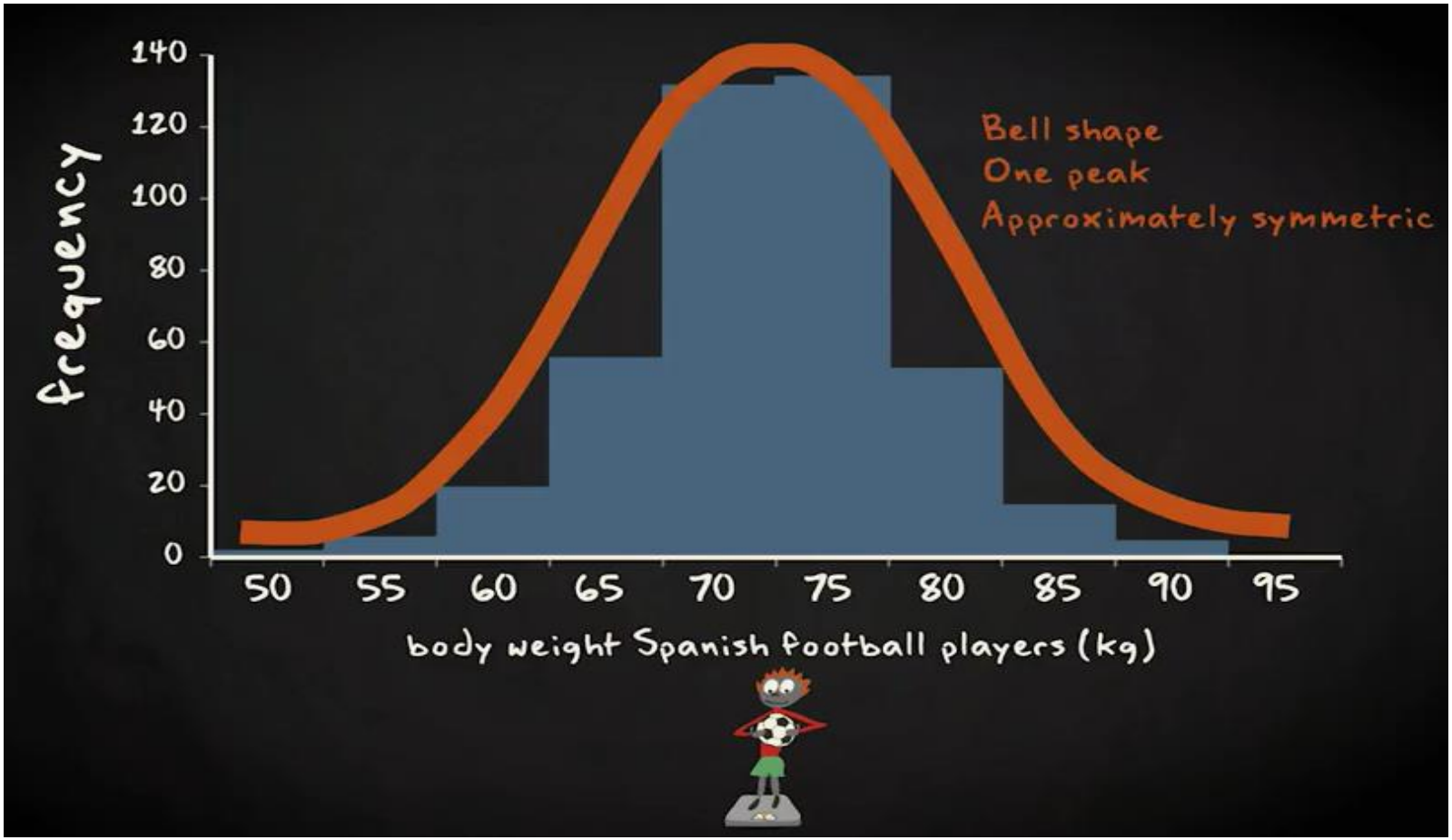
d) How many people have a level of cholesterol between 205 and 220 mg per dl?

(0.2 + 0.35 + 0.25)*200 = 160 people

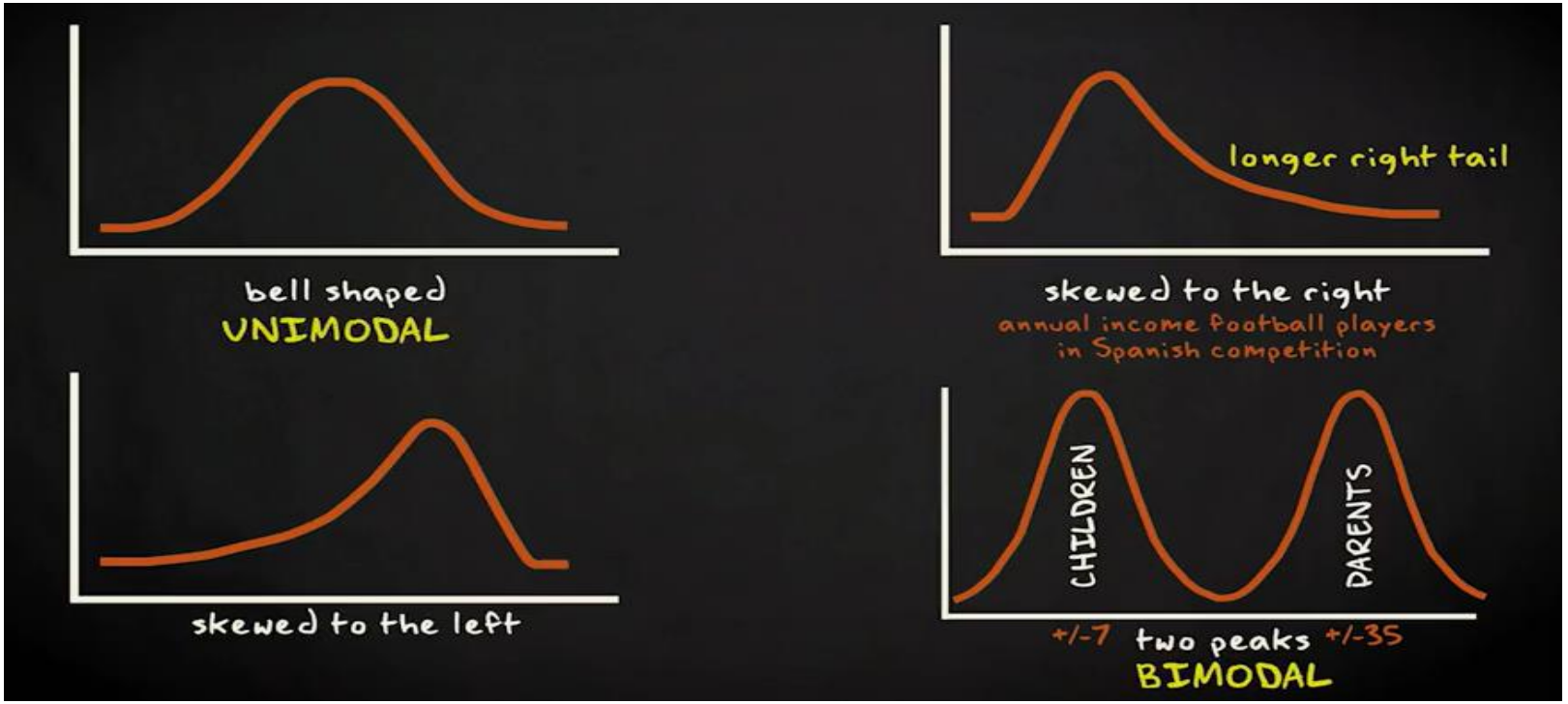
GRAPHS AND SHAPES OF DISTRIBUTIONS



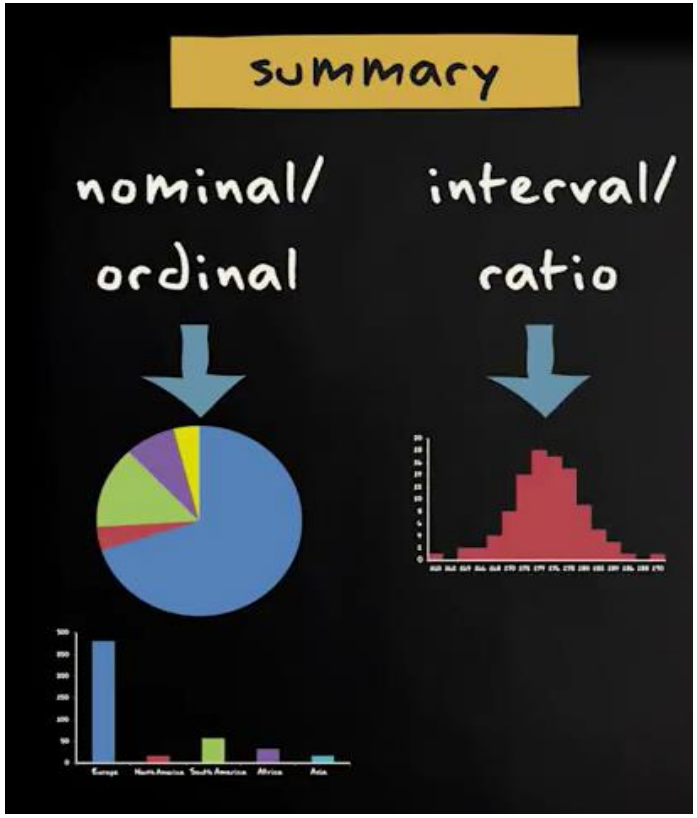
GRAPHS AND SHAPES OF DISTRIBUTIONS

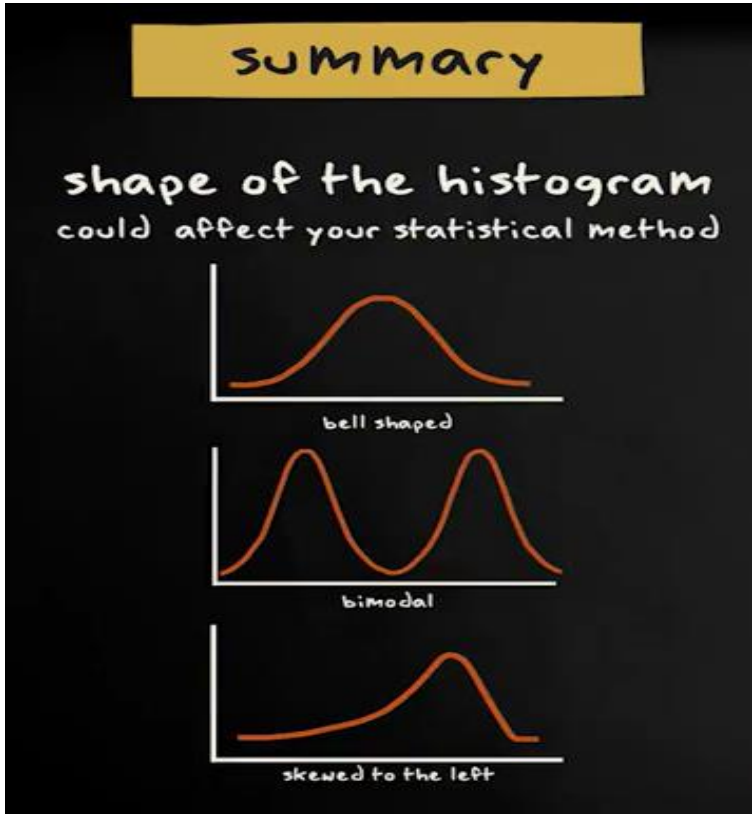


GRAPHS AND SHAPES OF DISTRIBUTIONS



GRAPHS AND SHAPES OF DISTRIBUTIONS





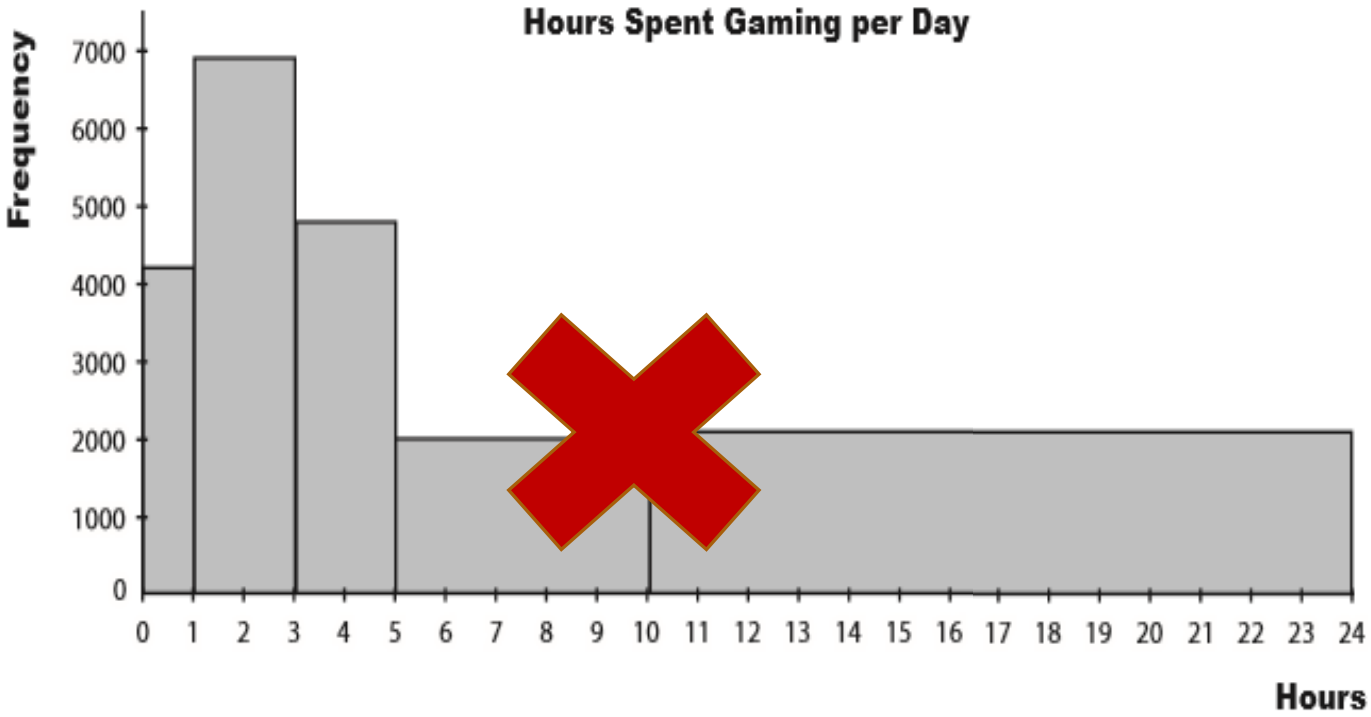
Exercise

This is the number of hours people play for

Hours	Frequency
0–1	4,300
1–3	6,900
3–5	4,900
5–10	2,000
10–24	2,100

This is frequency with which people play for this lengh of time

Exercise


Hours Spent Gaming per Day
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Hours

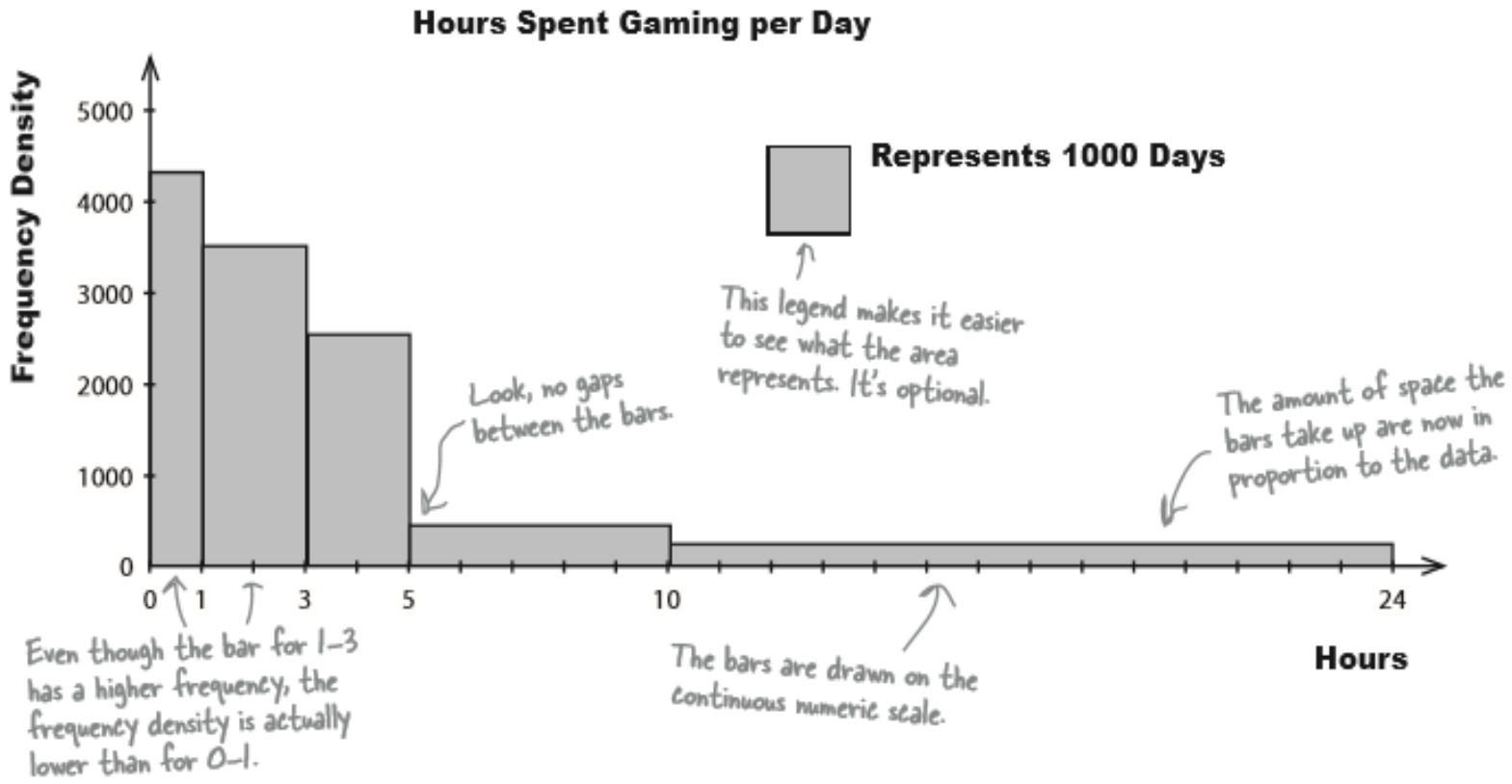
Exercise

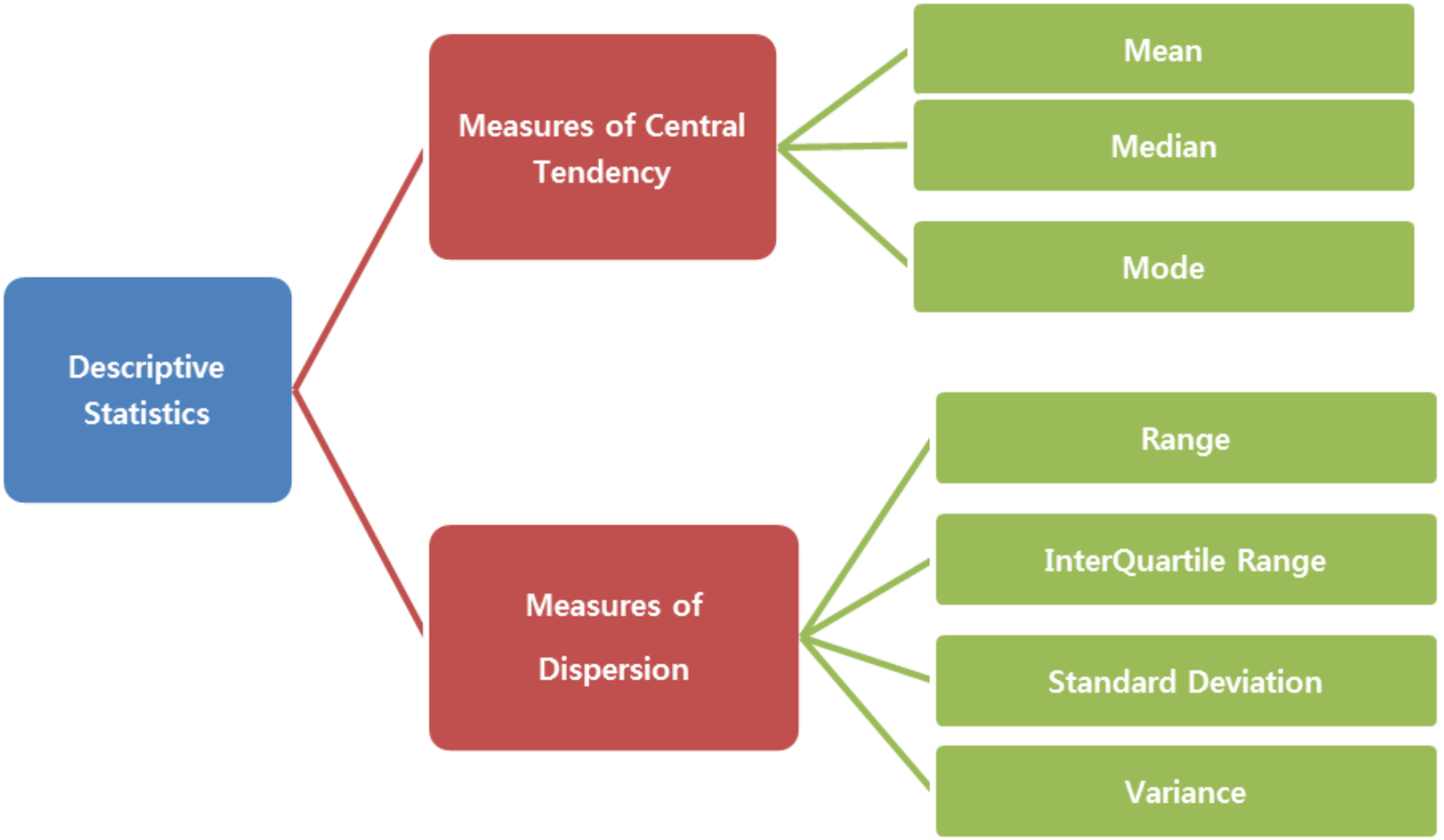
This is the number of hours people play for

Hours	Frequency
0–1	4,300
1–3	6,900
3–5	4,900
5–10	2,000
10–24	2,100

This is frequency with which people play for this lengh of time

Exercise


Hours Spent Gaming per Day
Frequency Density
5000
4000
3000
2000
1000
0
Represents 1000 Days
This legend makes it easier to see what the area represents. It's optional.
Look, no gaps between the bars.
The amount of space the bars take up are now in proportion to the data.
0
1
3
5
10
24
Hours
Even though the bar for 1–3 has a higher frequency, the frequency density is actually lower than for 0–1.
The bars are drawn on the continuous numeric scale.


Descriptive Statistics
Measures of Central Tendency
Mean
Median
Mode
Measures of Dispersion
Range
InterQuartile Range
Standard Deviation
Variance

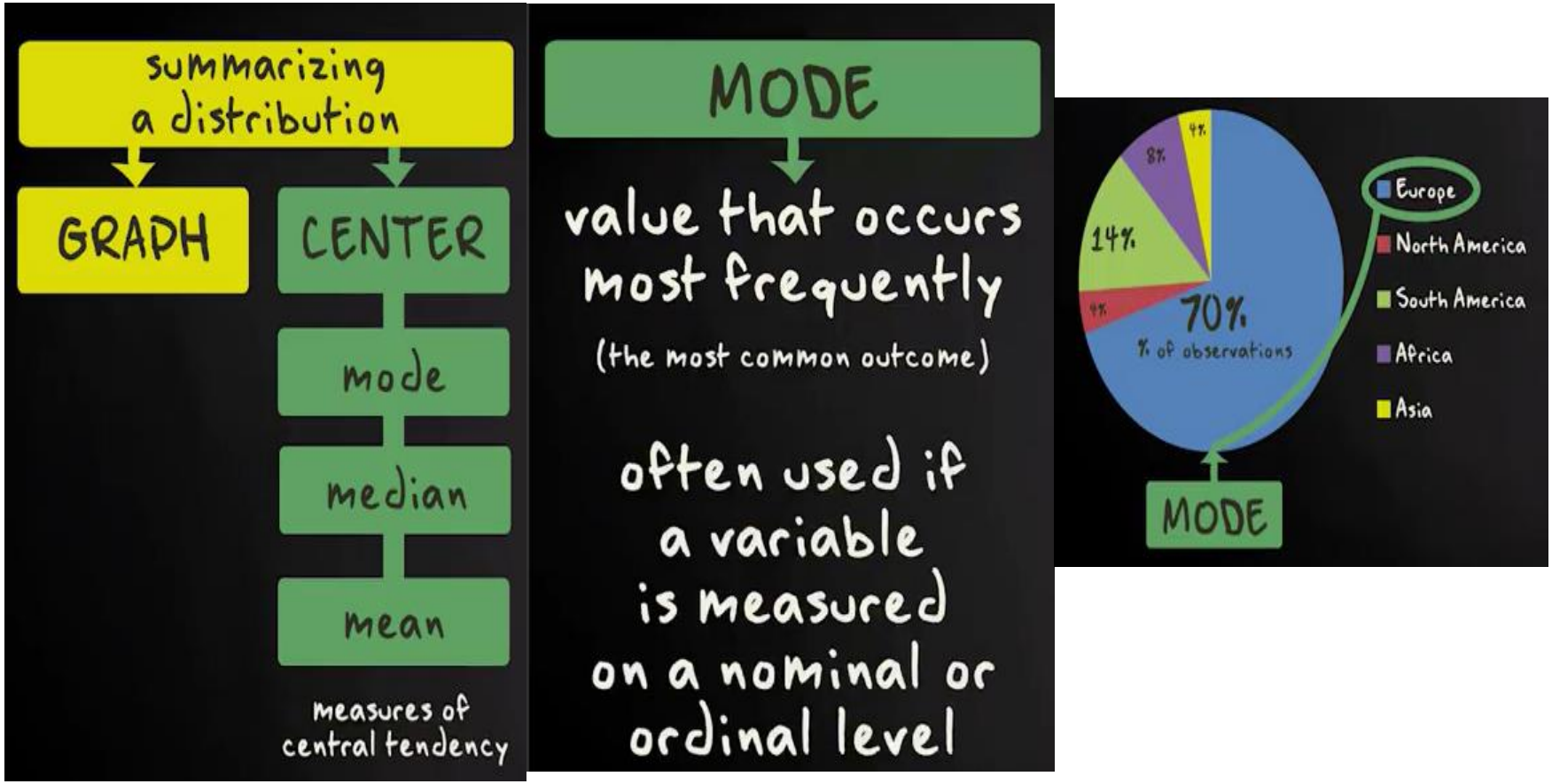
MEASURES OF CENTRAL TENDENCY AND DISPERSION

- 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**.

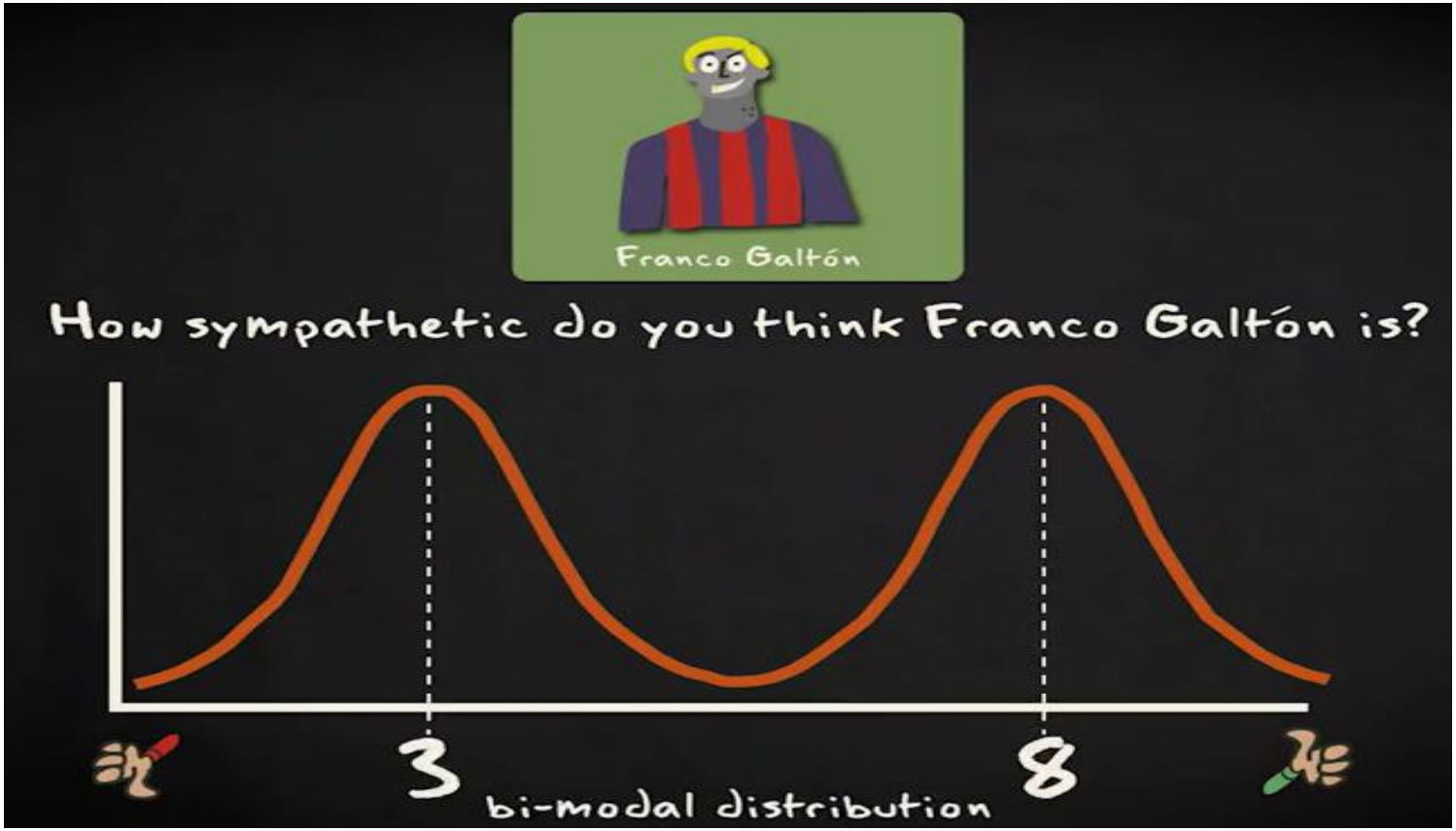
MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY

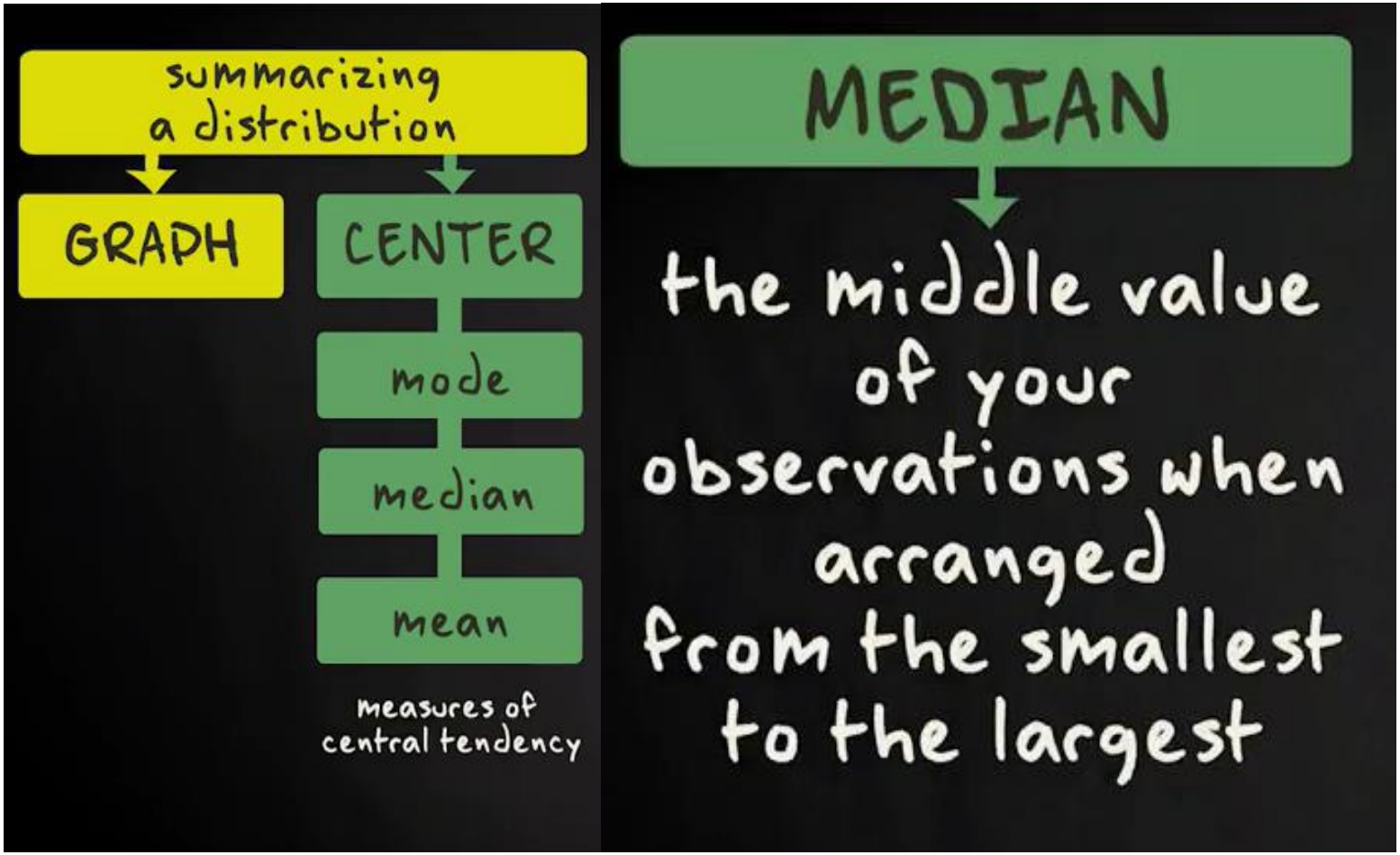


MEASURES OF CENTRAL TENDENCY

- **MODE**

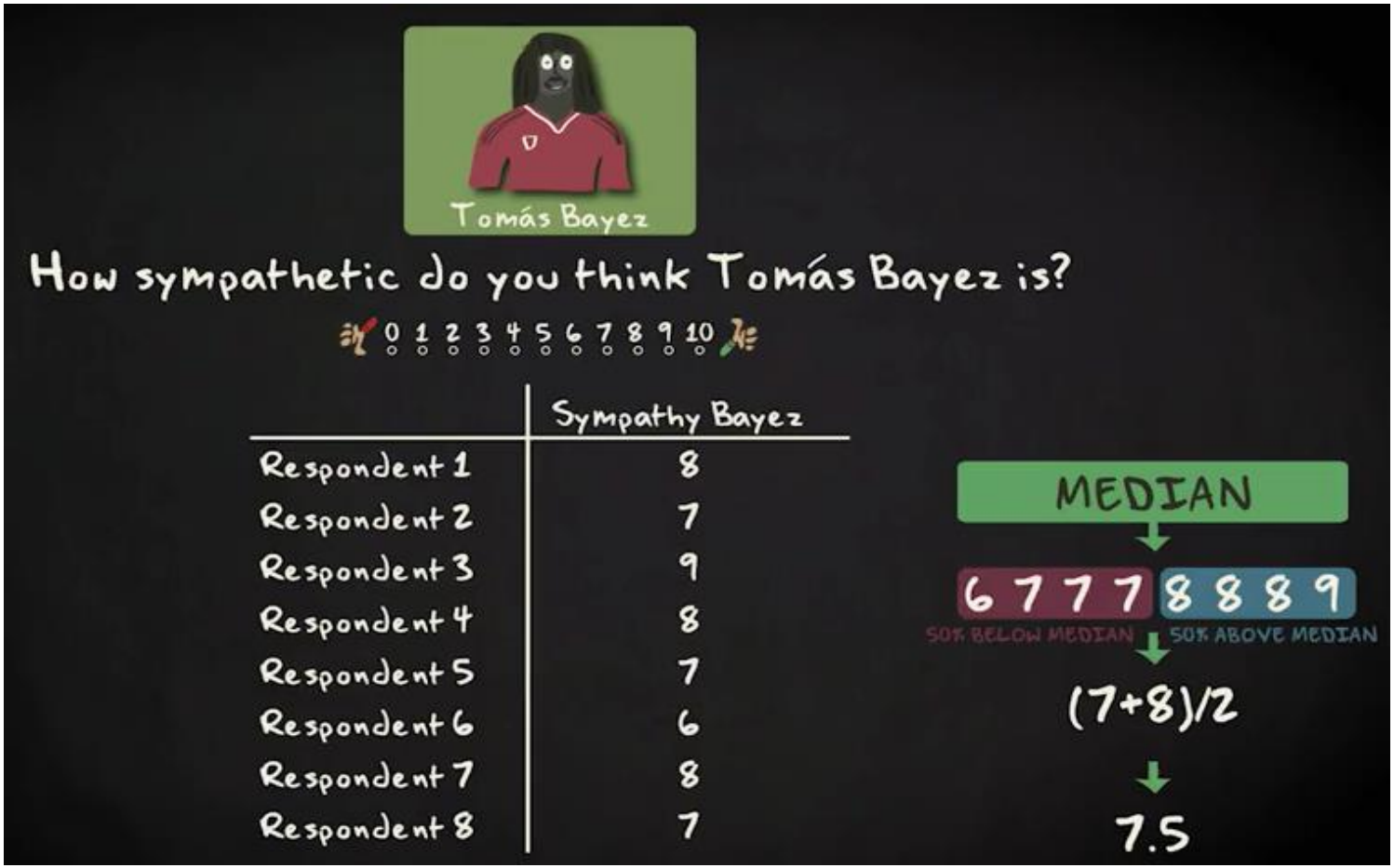


MEASURES OF CENTRAL TENDENCY

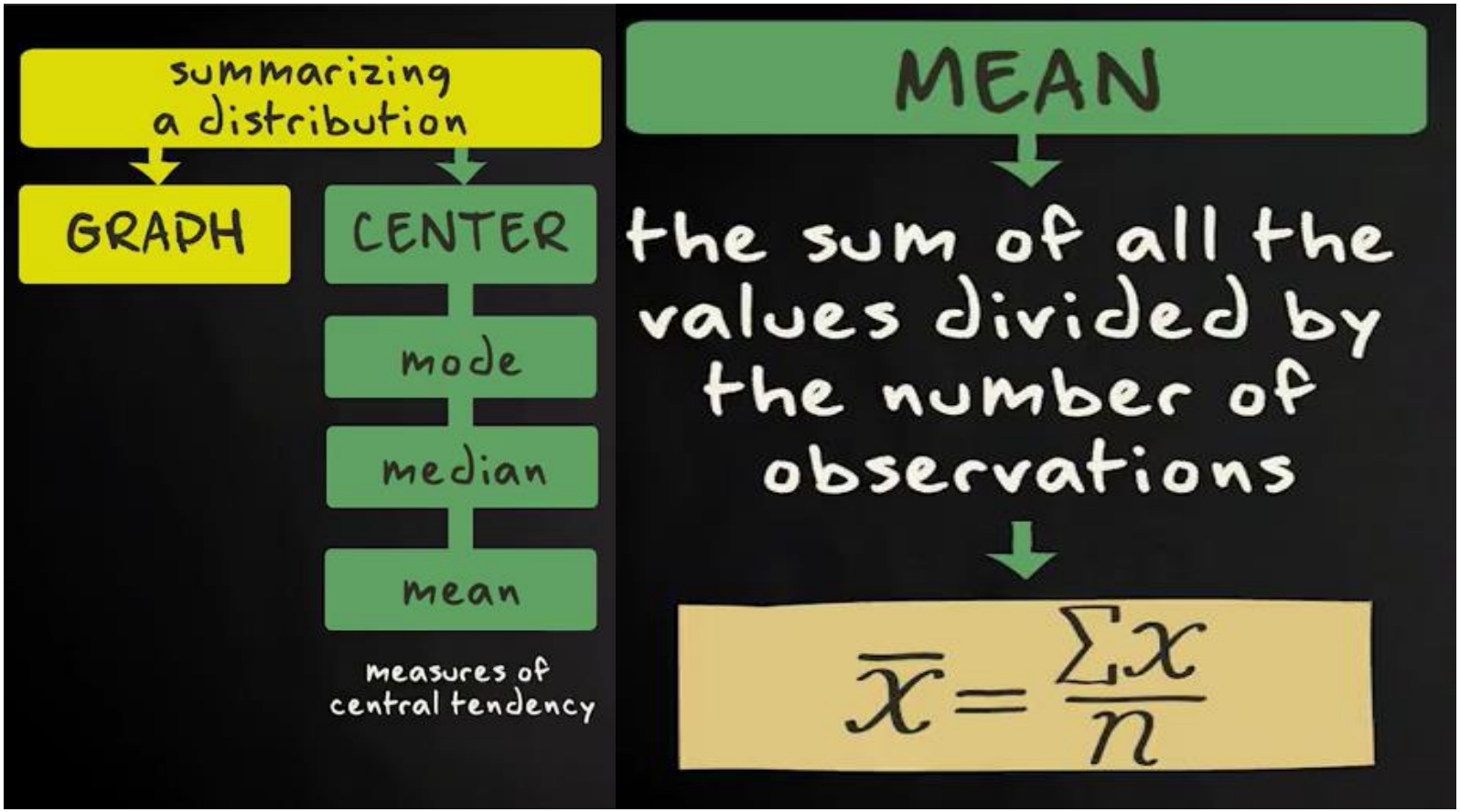


MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY

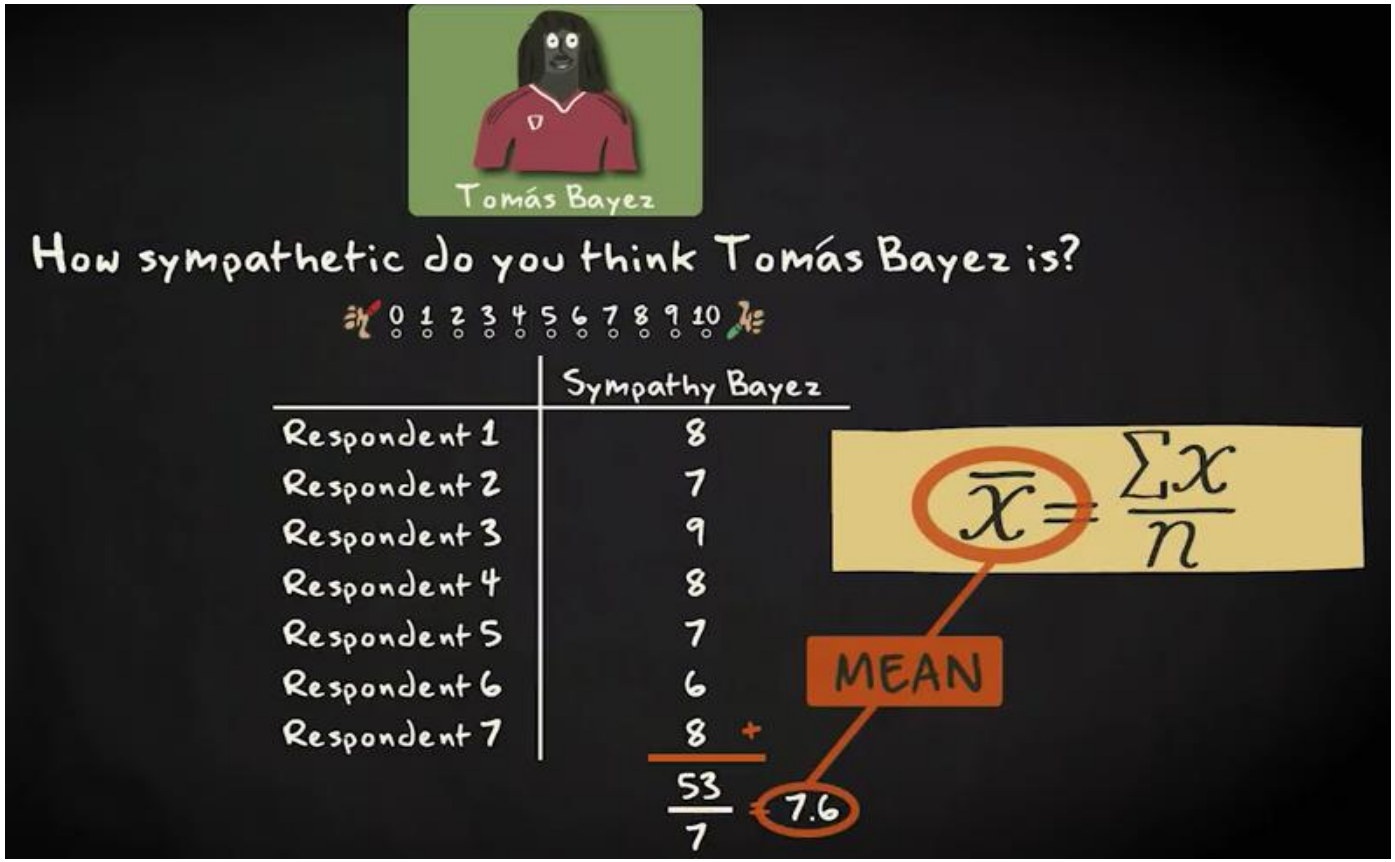


MEASURES OF CENTRAL TENDENCY



MEASURES OF CENTRAL TENDENCY

❑ MEAN (UNGROUPED DATA)



MEASURES OF CENTRAL TENDENCY

- **MEAN (GROUPED DATA)**

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

x = class midpoint

MEASURES OF CENTRAL TENDENCY

❑ MEAN (GROUPED DATA)

Age	*Frrquency (f)*	*Midpoint (x)*	*fx*
30-34	*4*	*32*	*128*
35-39	*5*	*37*	*185*
40-44	*2*	*42*	*84*
45-49	*9*	*47*	*423*
Total	*20*		*820*

$\sum f = n = 20$

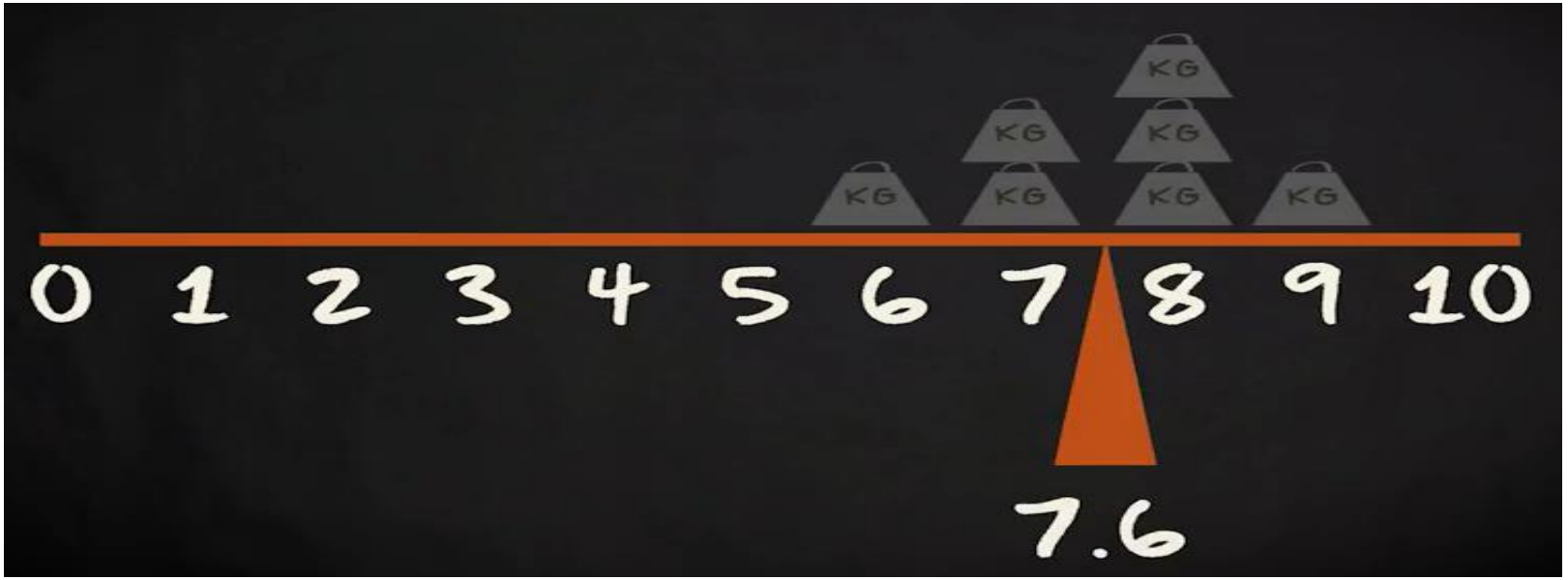
$\sum fx = 820$

Mean = 820/20 = 41

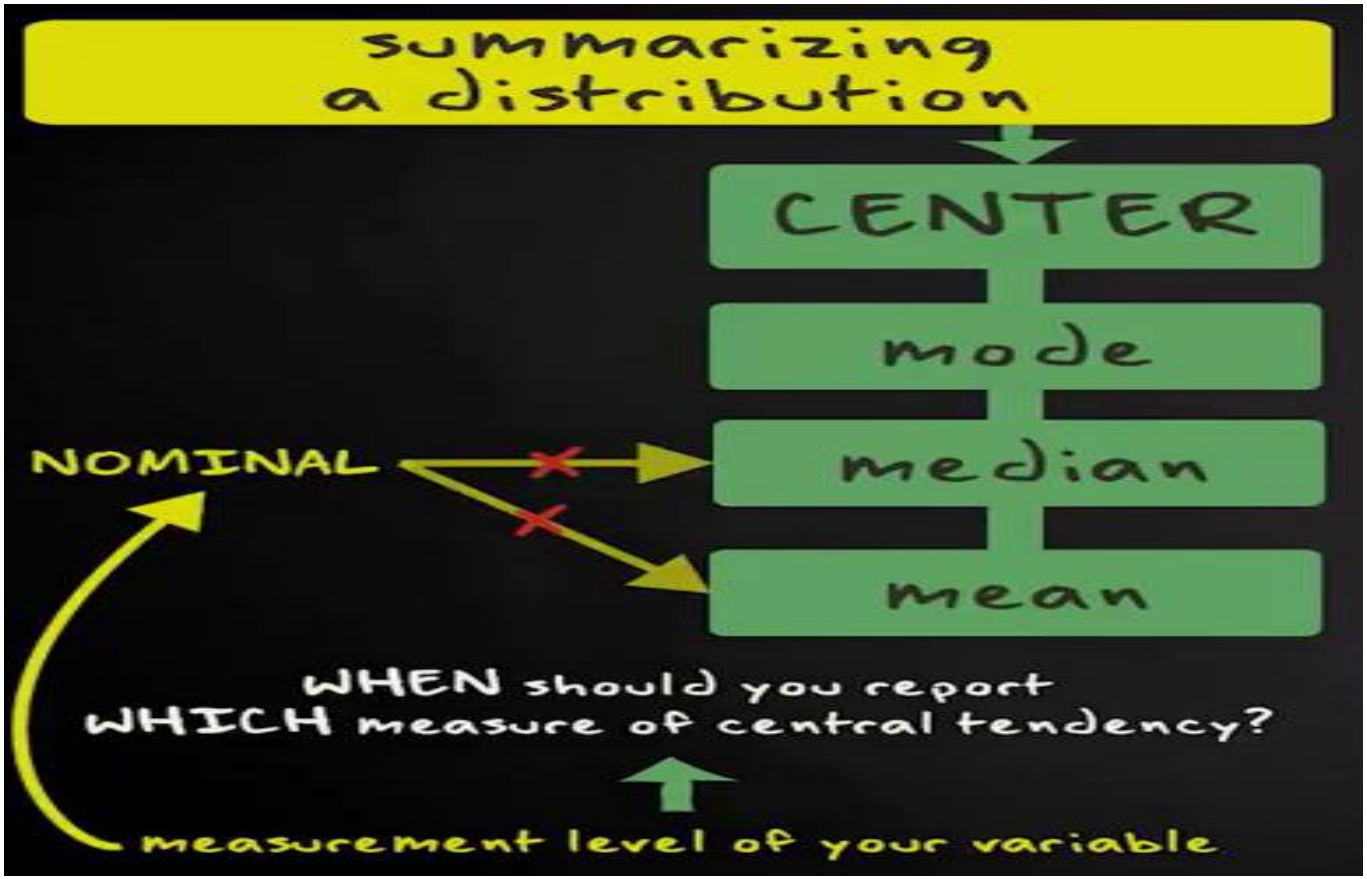
MEASURES OF CENTRAL TENDENCY

- **MEAN:**

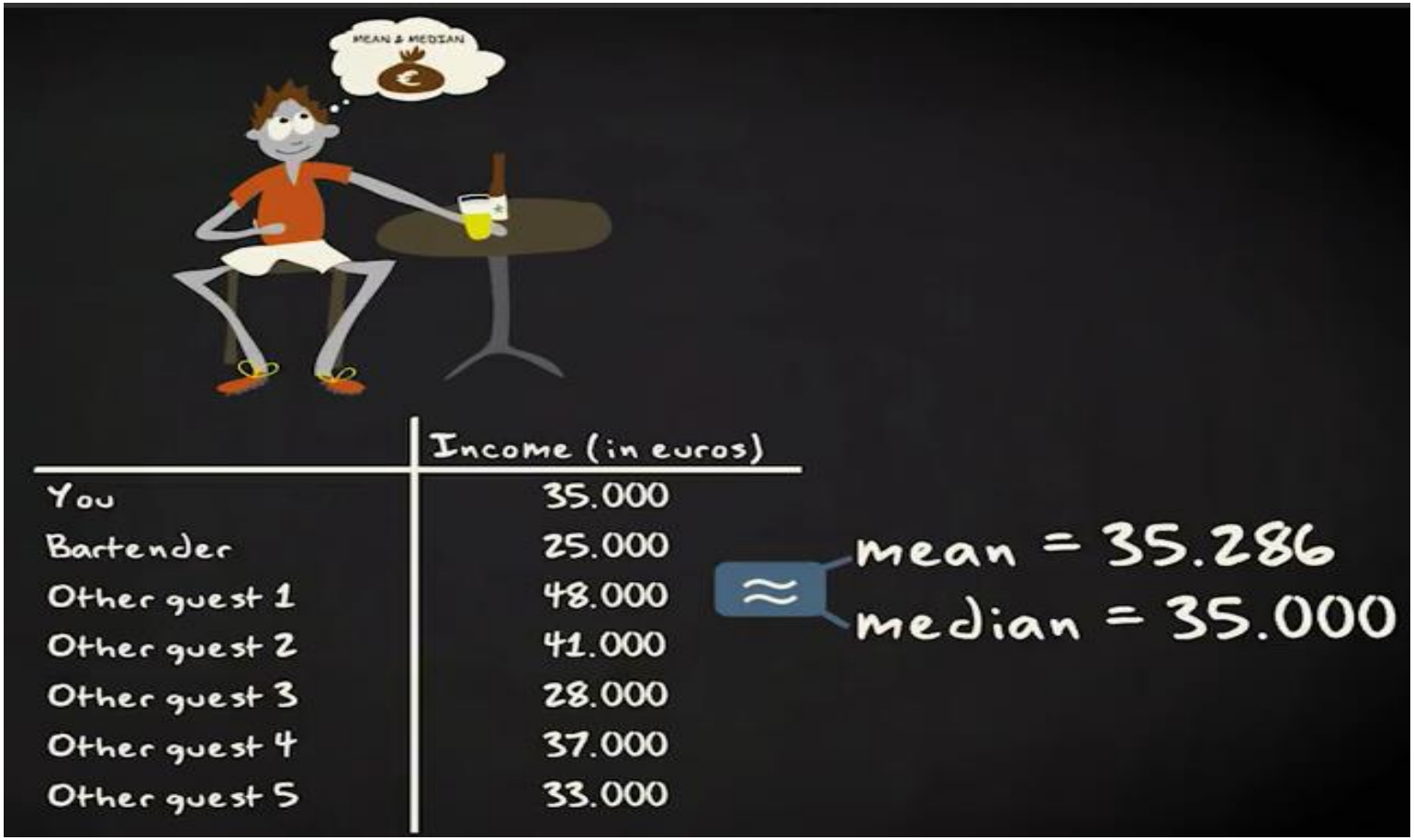
  Balance point of the data



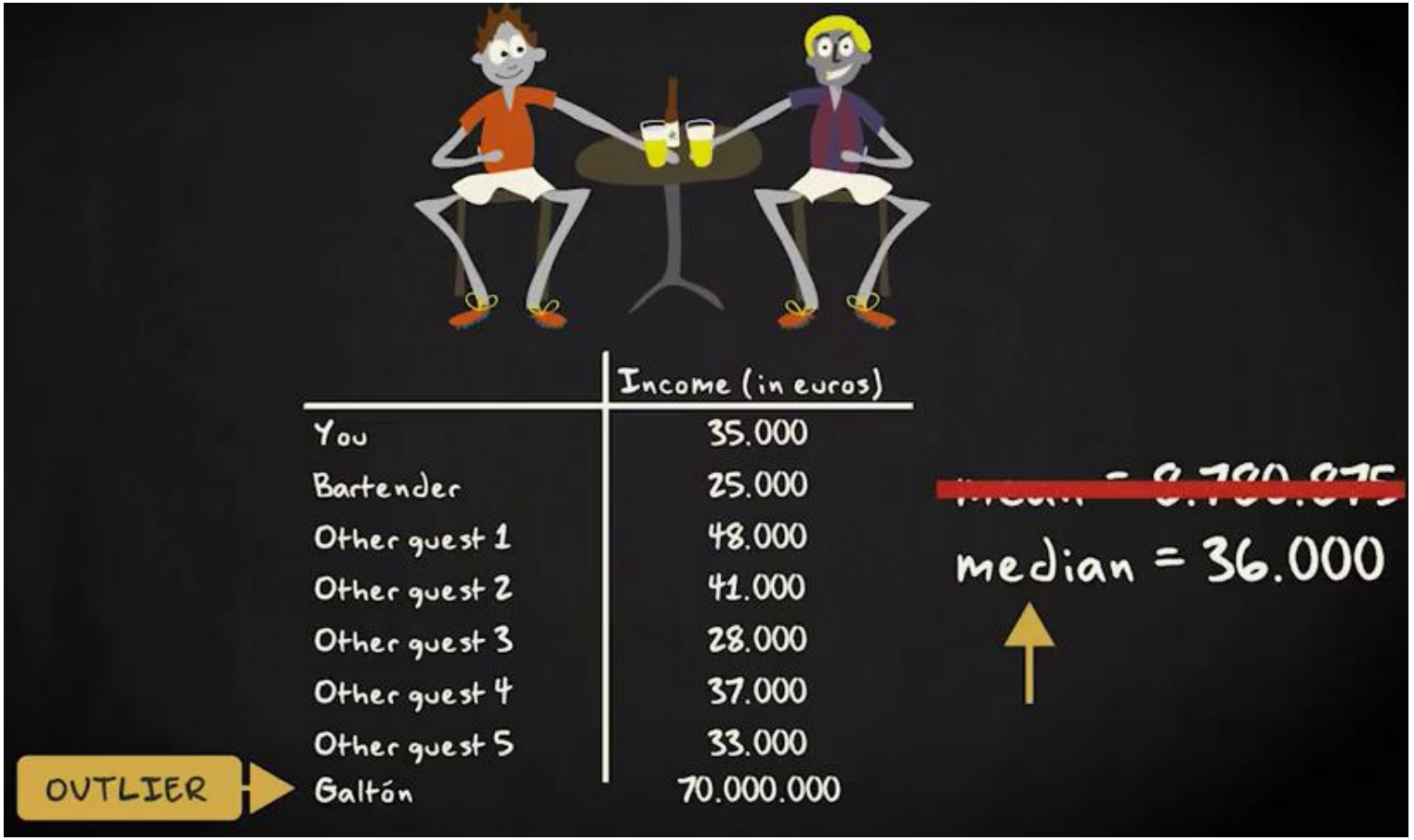
MEASURES OF CENTRAL TENDENCY



MEASURES OF CENTRAL TENDENCY

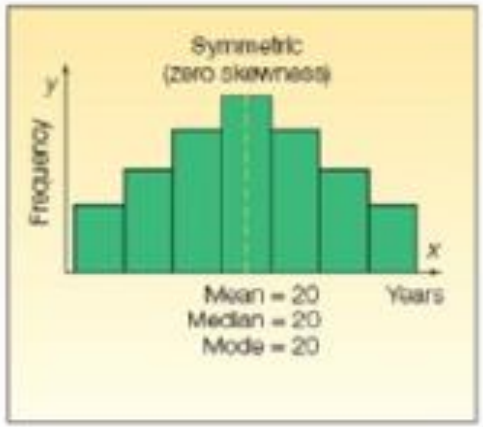


MEASURES OF CENTRAL TENDENCY



MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY

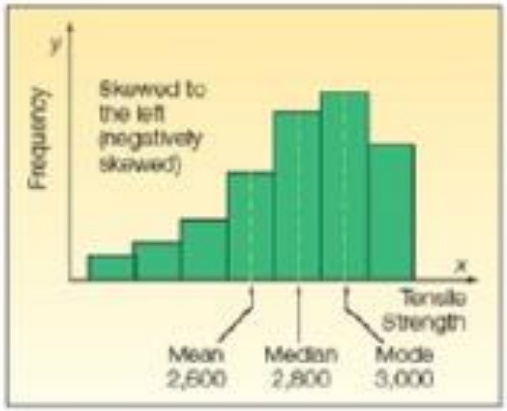


zero skewness

mode = median = mean

positive skewness

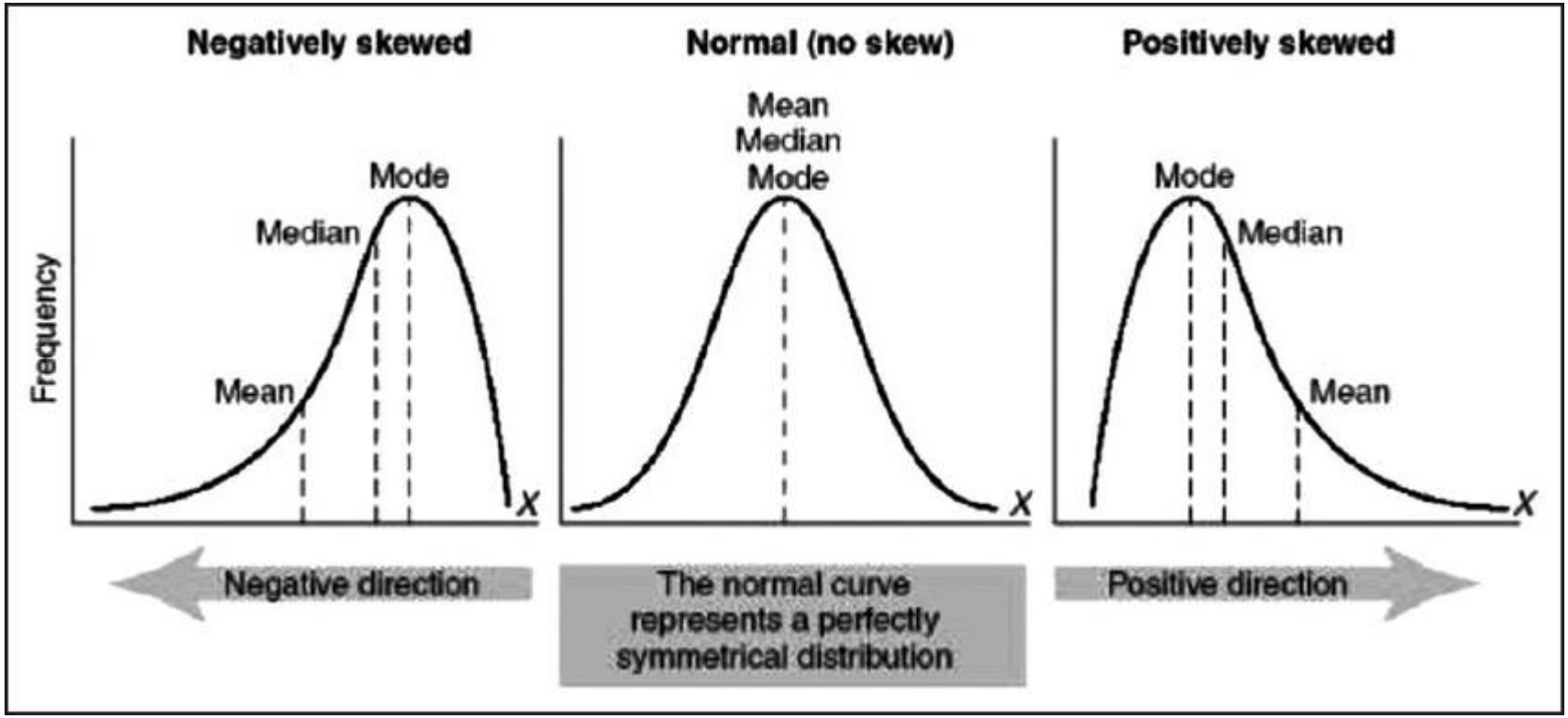
mode < median < mean



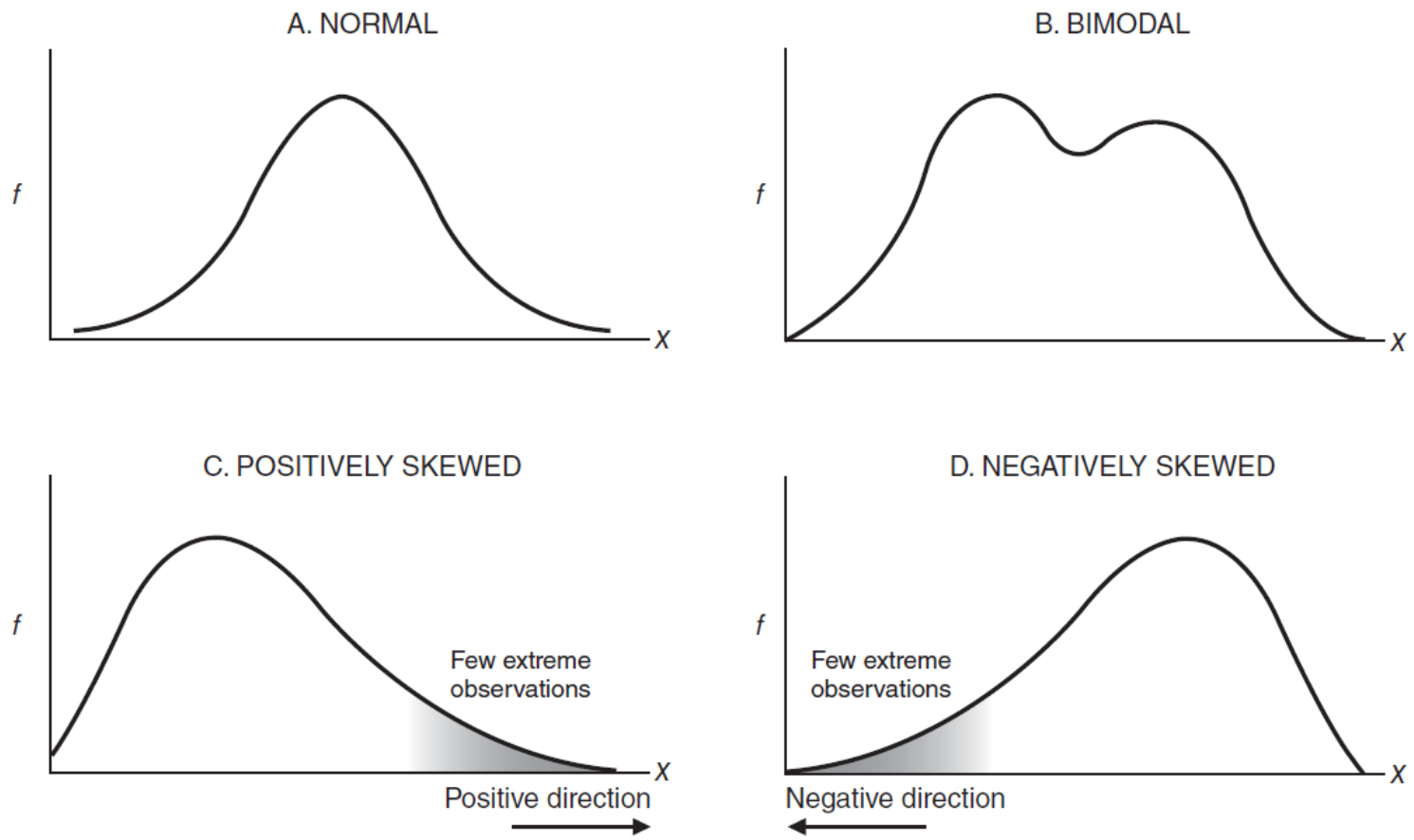
negative skewness

mode > median > mean

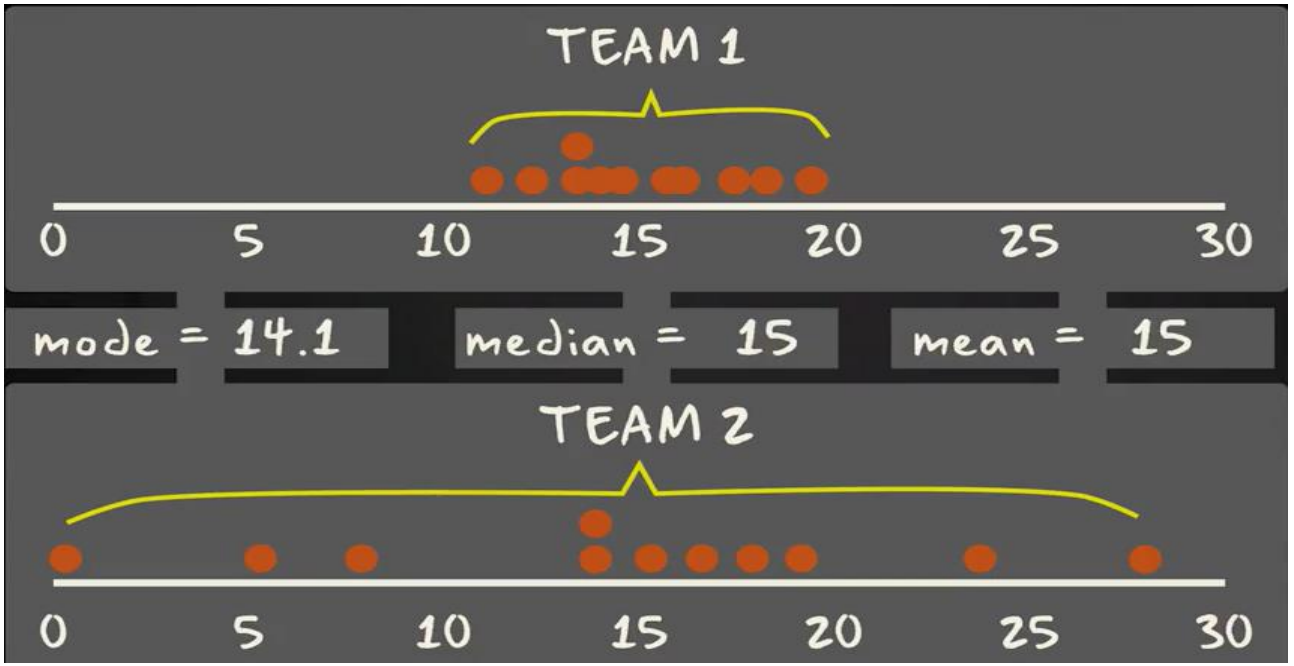
MEASURES OF CENTRAL TENDENCY



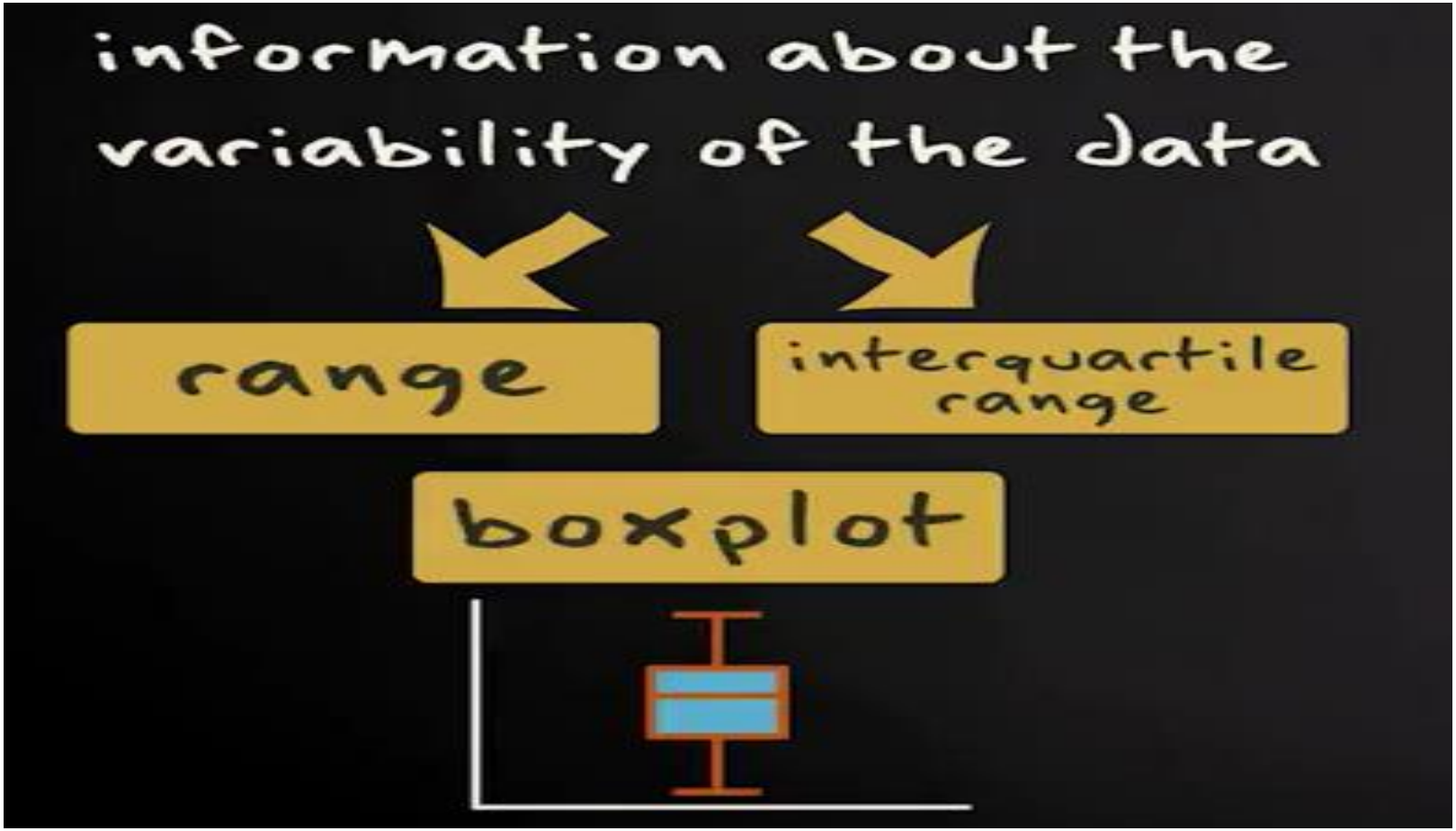
MEASURES OF CENTRAL TENDENCY



MEASURES OF DISPERSION OR VARIABILITY

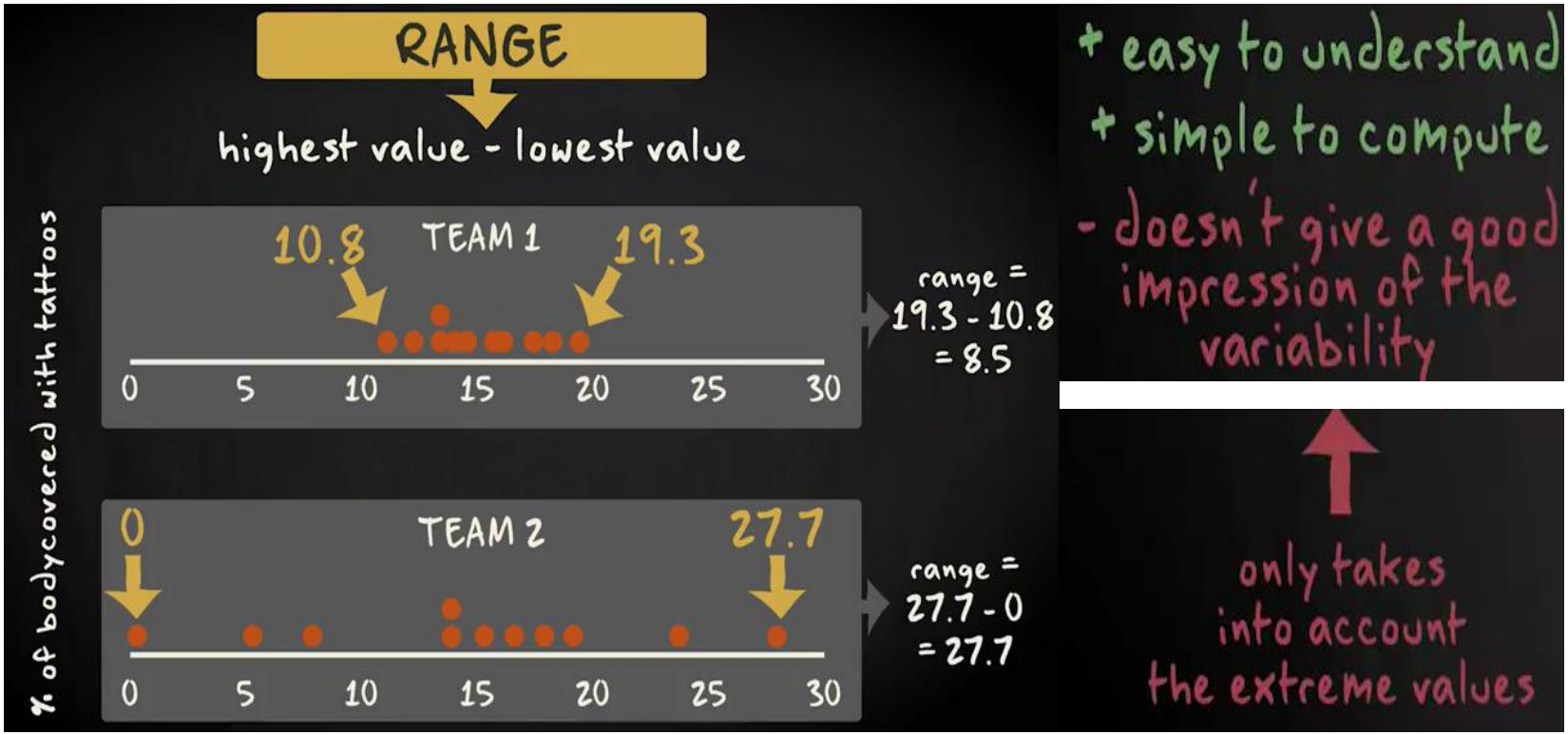


RANGE, INTERQUARTILE RANGE AND BOX PLOT



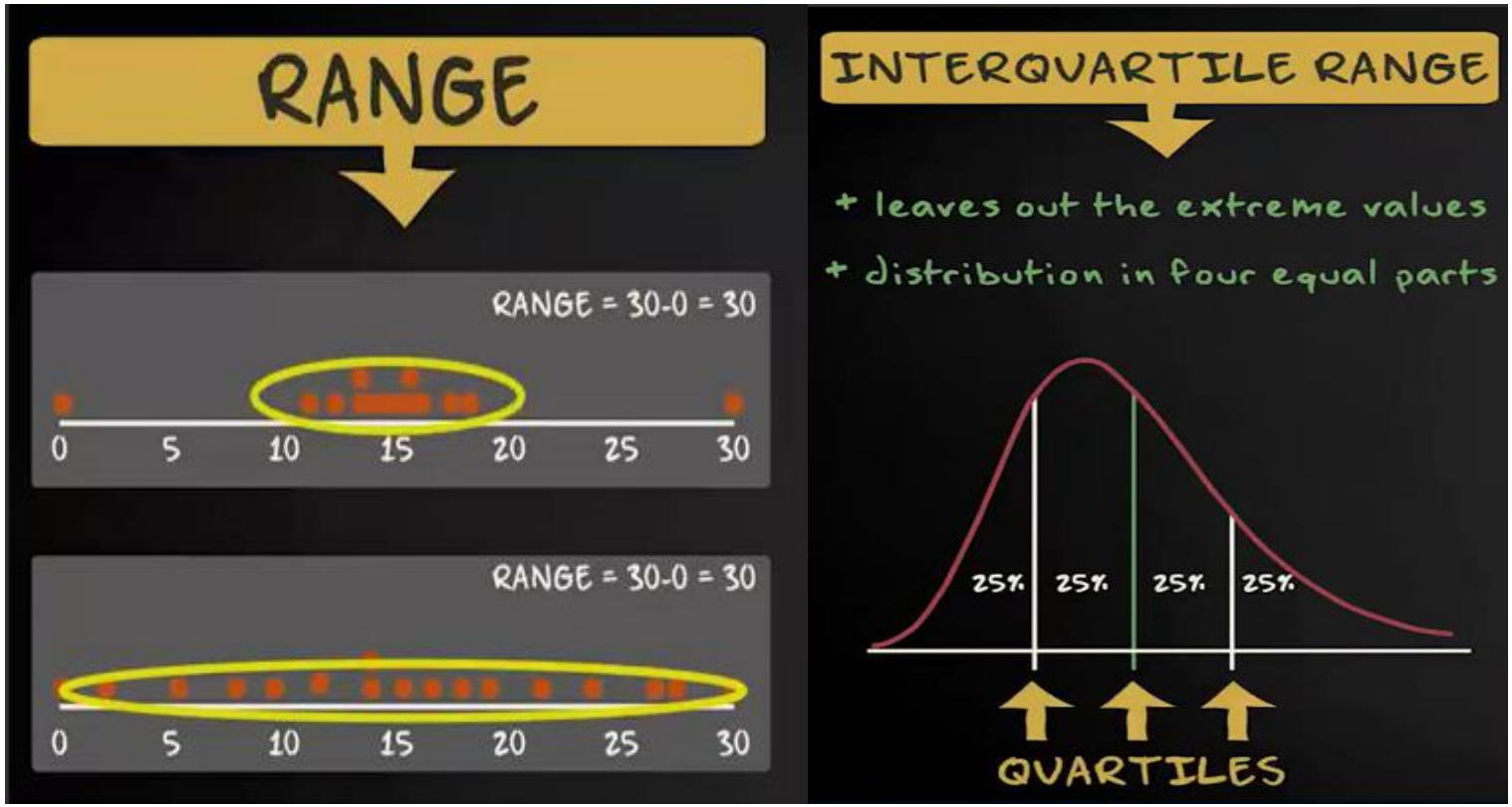
RANGE, INTERQUARTILE RANGE AND BOX PLOT

❑ RANGE (R):



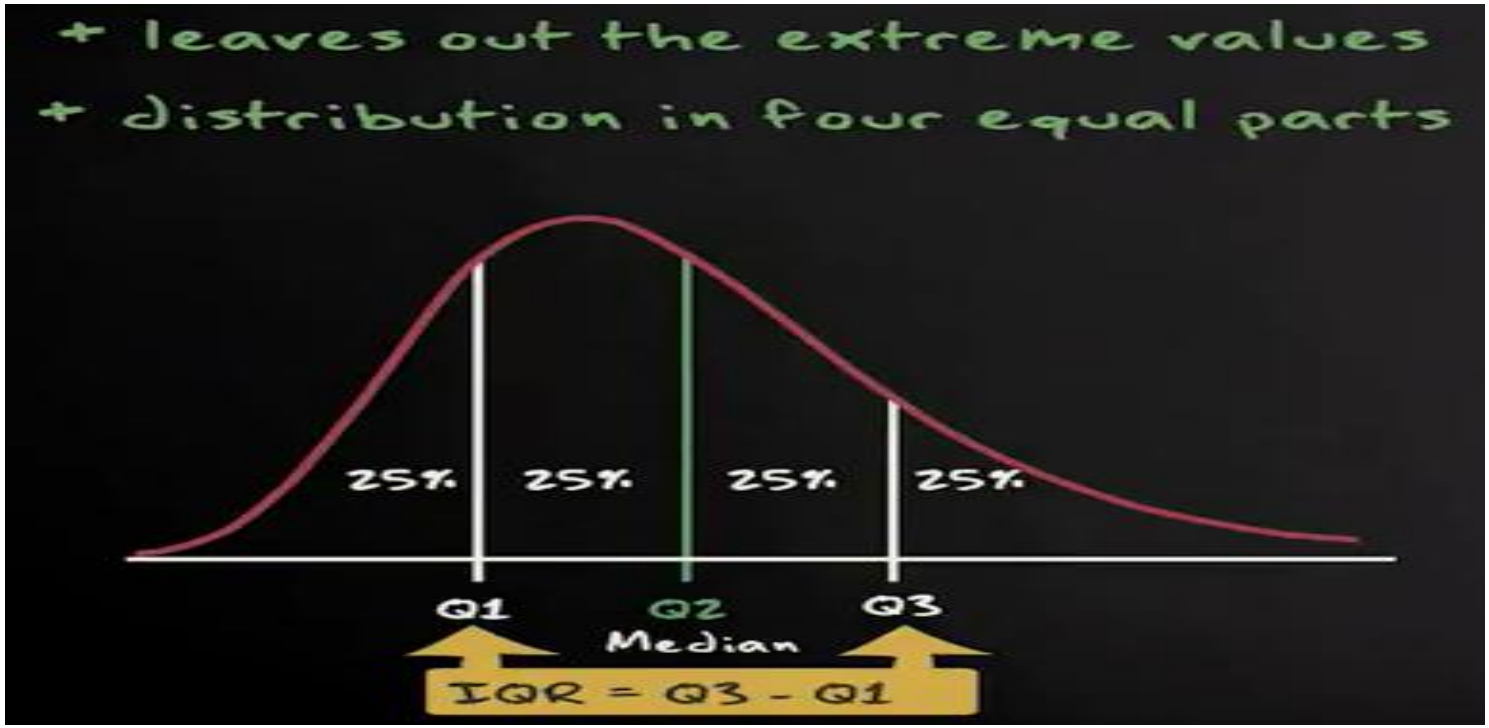
RANGE, INTERQUARTILE RANGE AND BOX PLOT

❑ INTERQUARTILE RANGE (IQR):



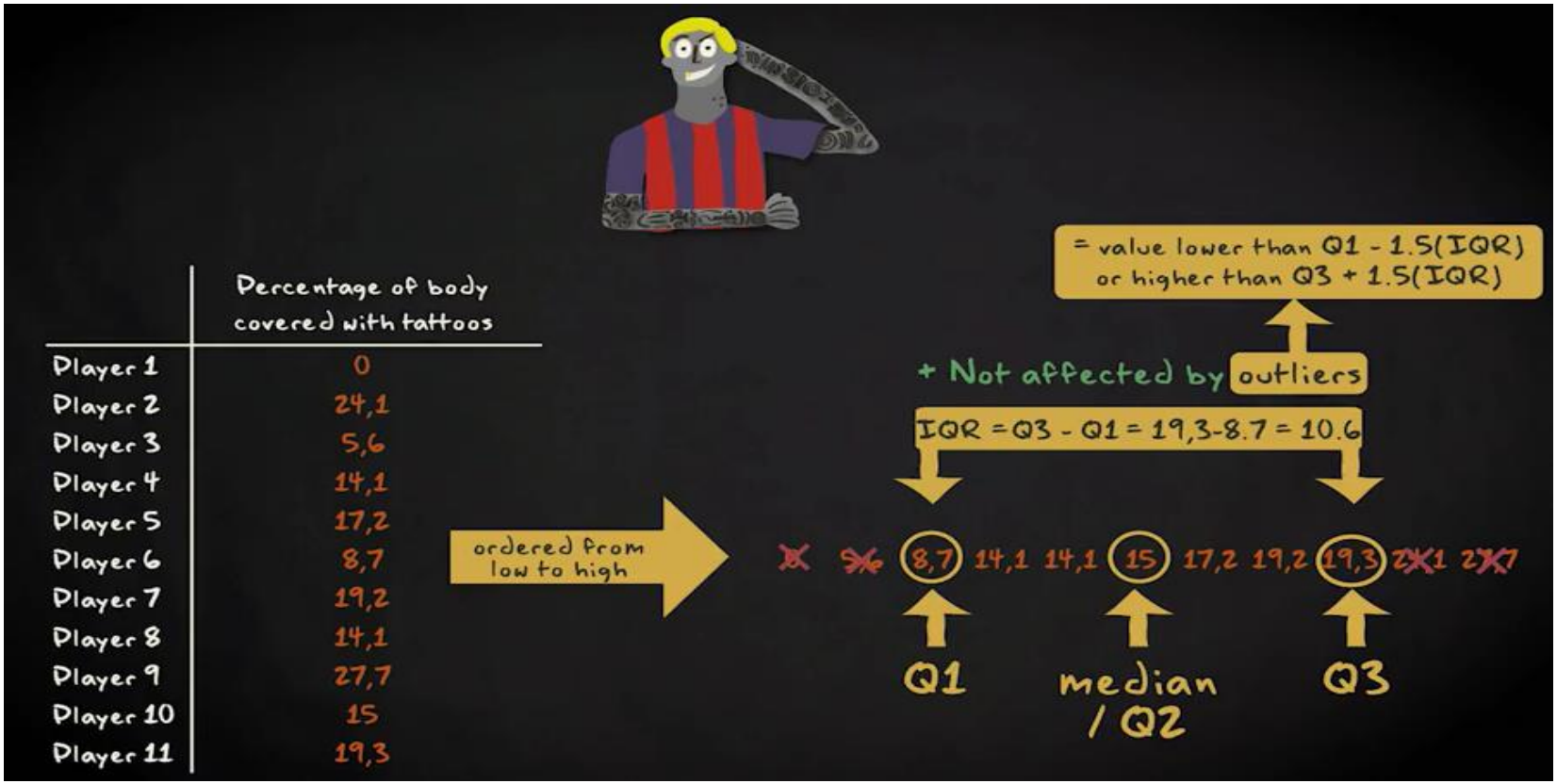
RANGE, INTERQUARTILE RANGE AND BOX PLOT

- **INTERQUARTILE RANGE (IQR):**



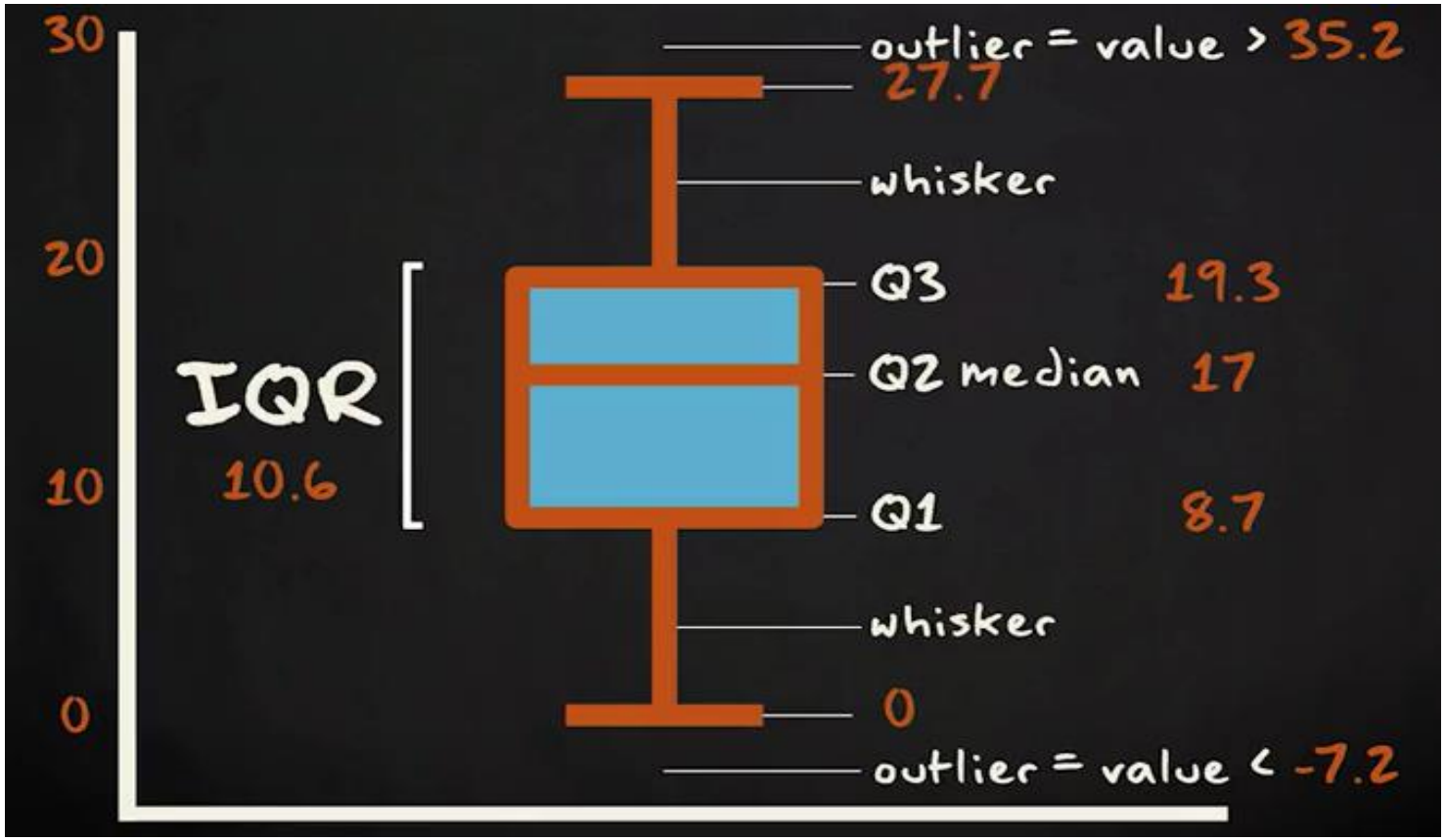
RANGE, INTERQUARTILE RANGE AND BOX PLOT

❑ INTERQUARTILE RANGE (IQR):



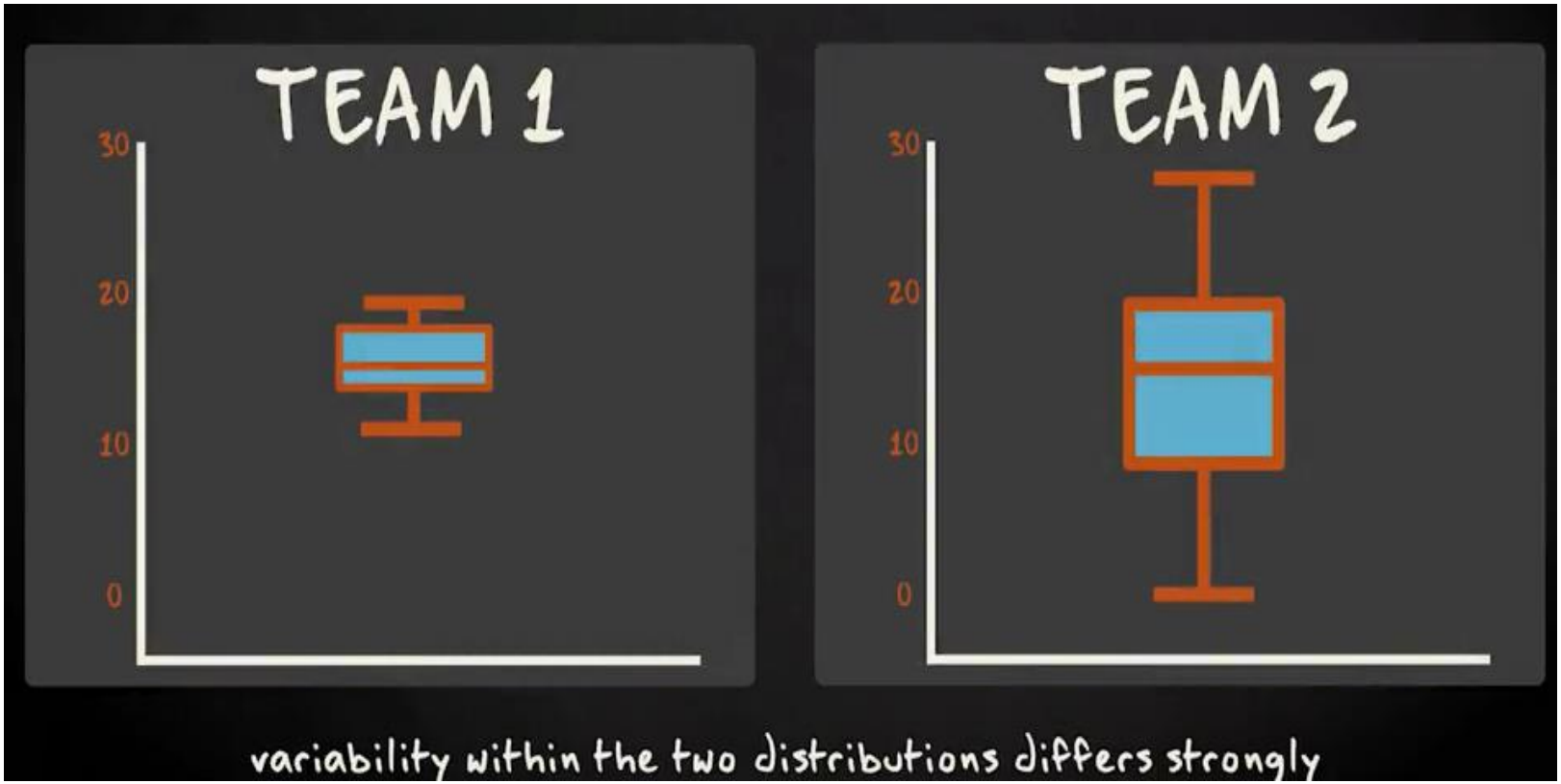
RANGE, INTERQUARTILE RANGE AND BOX PLOT

❑ BOX PLOT

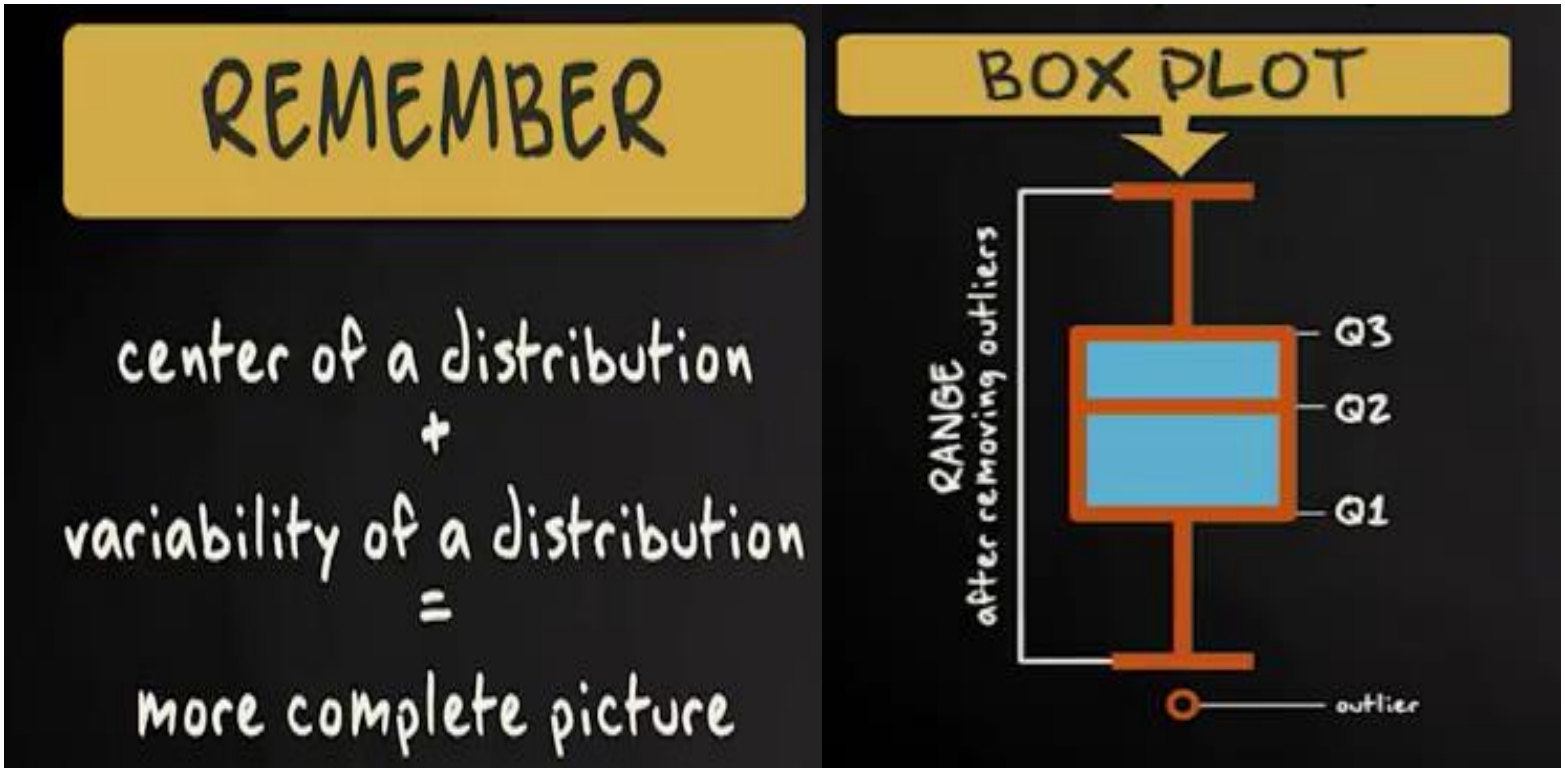


RANGE, INTERQUARTILE RANGE AND BOX PLOT

❑ BOX PLOT



RANGE, INTERQUARTILE RANGE AND BOX PLOT



RANGE, INTERQUARTILE RANGE AND BOX PLOT

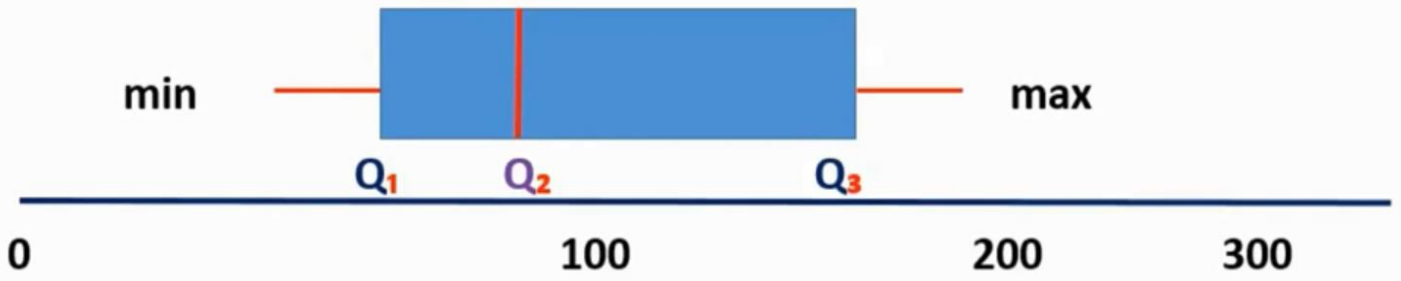
construct a boxplot for the data:

89, 47, 164, 296, 30, 215, 138, 78, 48, 39

Arrange: 30, 39, 47, 48, 78, 89, 138, 164, 215, 296

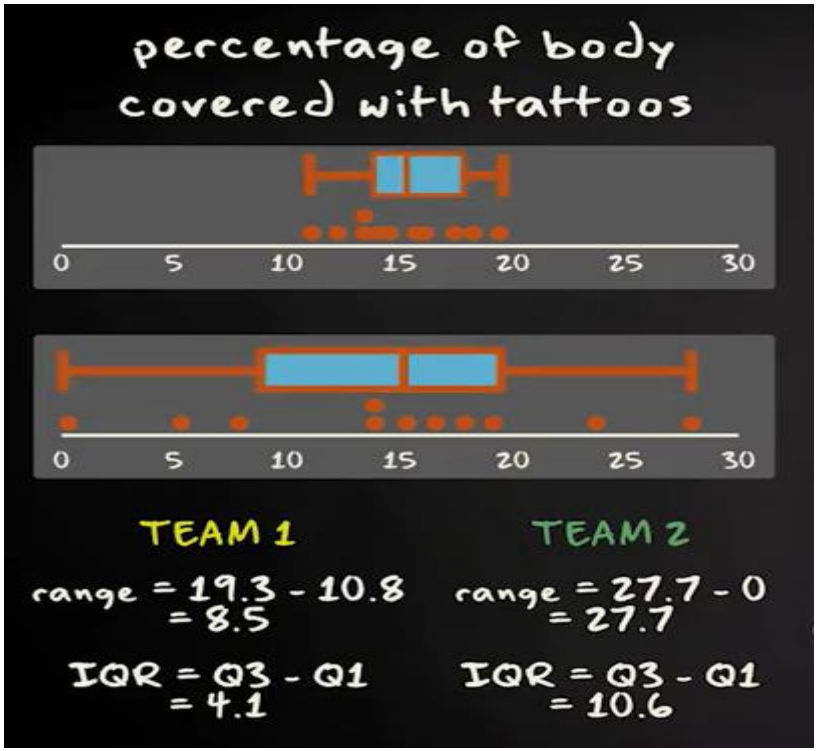
The median: 78 + 89 / 2 = 83.5

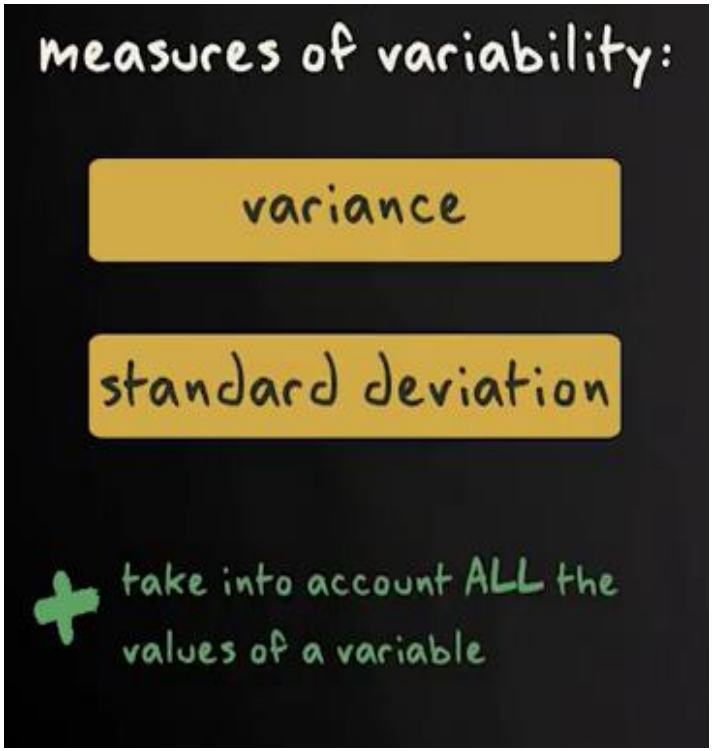
Q_1 = 30, 39, 47, 48, 78 = 47 Q_3 = 89, 138, 164, 215, 296 = 164



The distribution is positively skewed

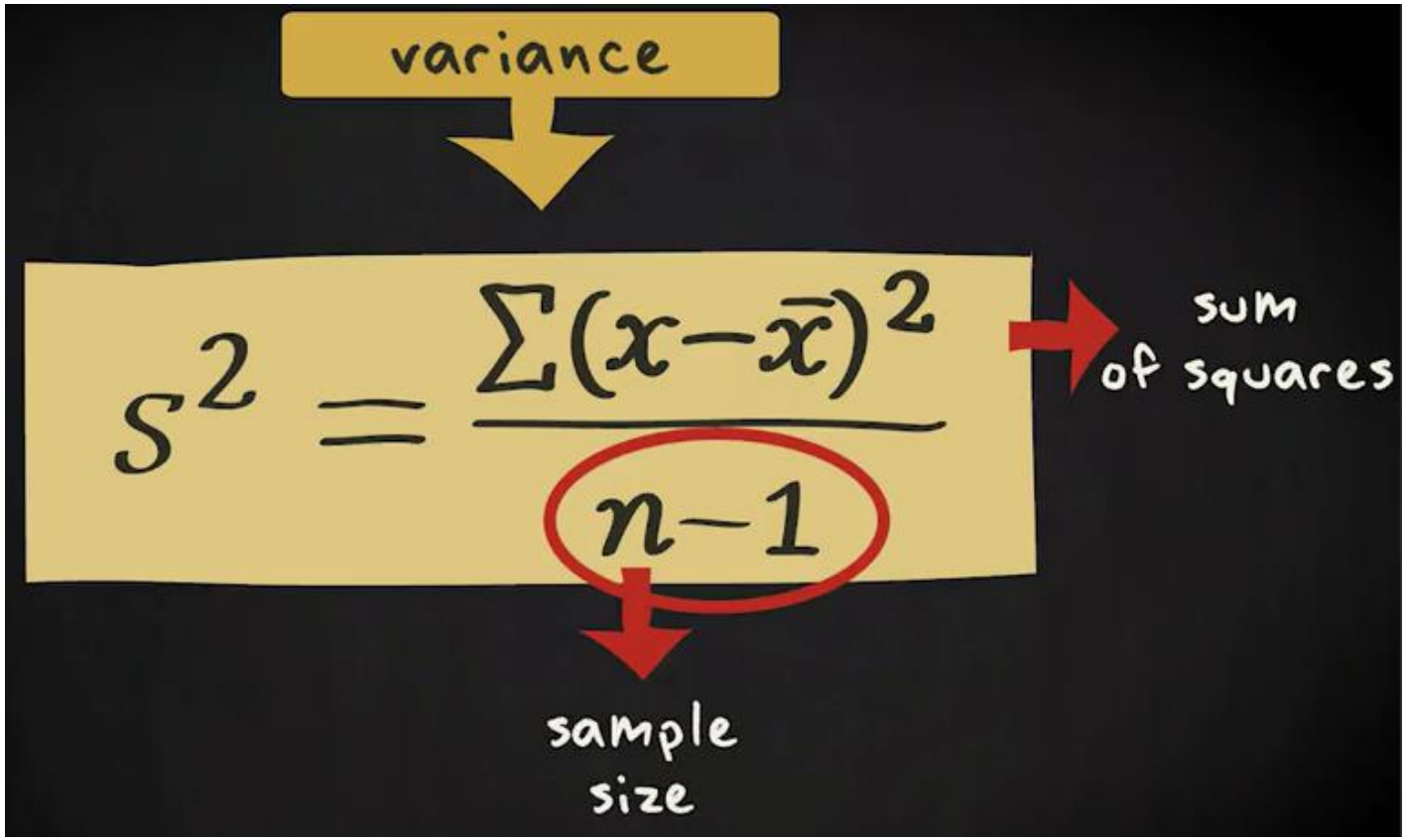
VARIANCE AND STANDARD DEVIATION





VARIANCE AND STANDARD DEVIATION

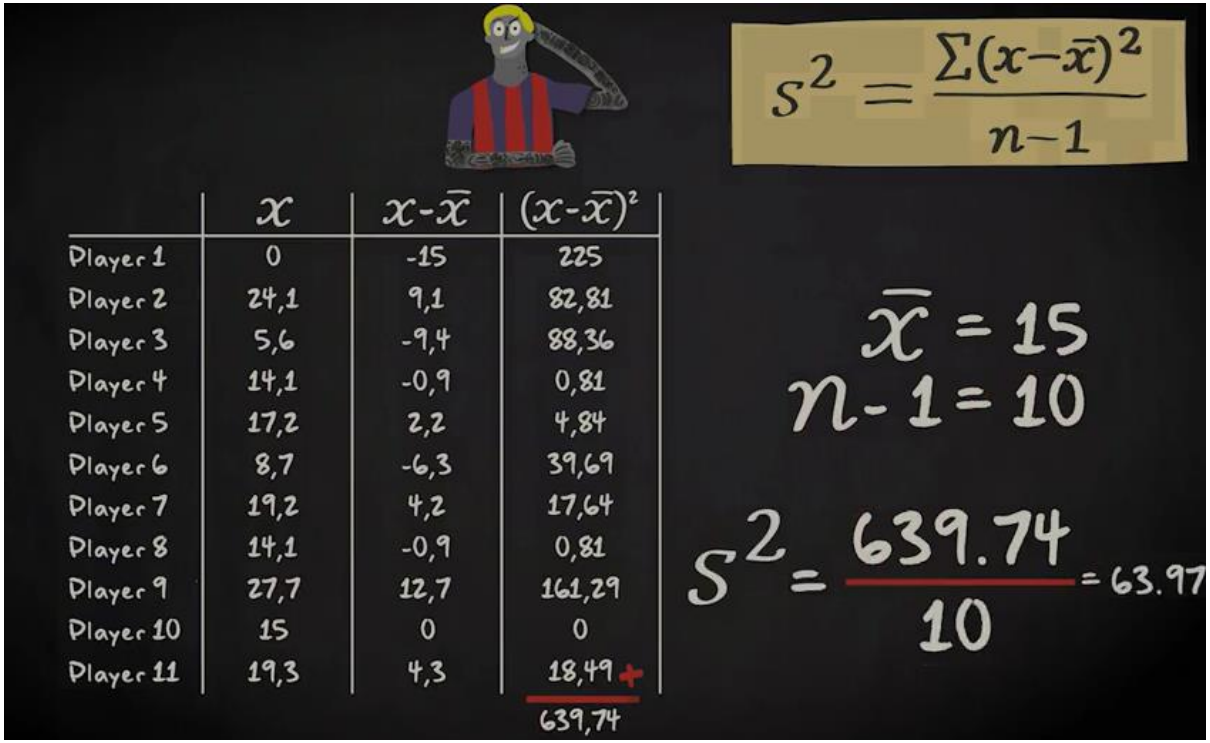
- **VARIANCE (UNGROUPED DATA)**



VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (UNGROUPED DATA)

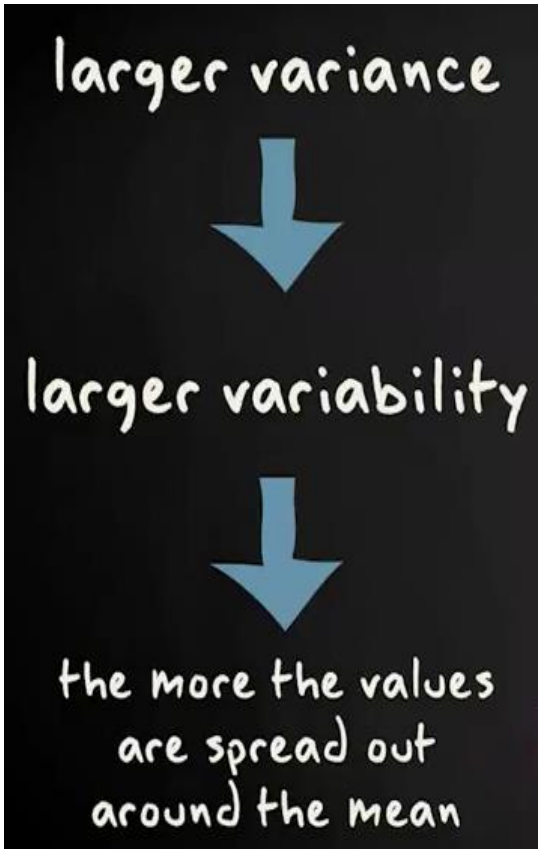
- Mean is the point of balance, so we have positive and negative deviations from the mean.
- The sum of deviation sum to zero. That's why we don't use the original deviations, but the squared deviations.

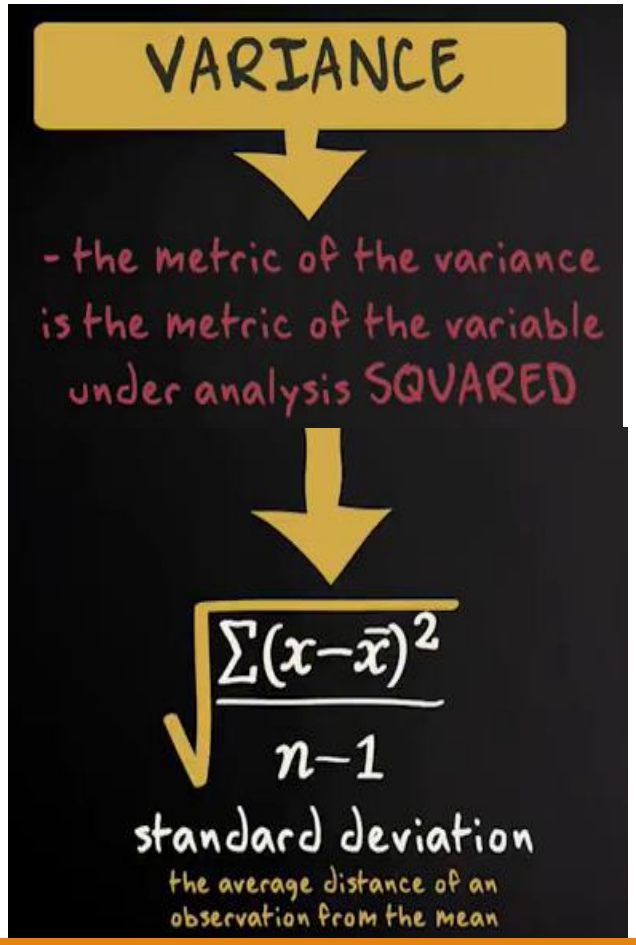


	x	$x-\bar{x}$	$(x-\bar{x})^2$
Player 1	0	-15	225
Player 2	24,1	9,1	82,81
Player 3	5,6	-9,4	88,36
Player 4	14,1	-0,9	0,81
Player 5	17,2	2,2	4,84
Player 6	8,7	-6,3	39,69
Player 7	19,2	4,2	17,64
Player 8	14,1	-0,9	0,81
Player 9	27,7	12,7	161,29
Player 10	15	0	0
Player 11	19,3	4,3	18,49
			639,74

VARIANCE AND STANDARD DEVIATION

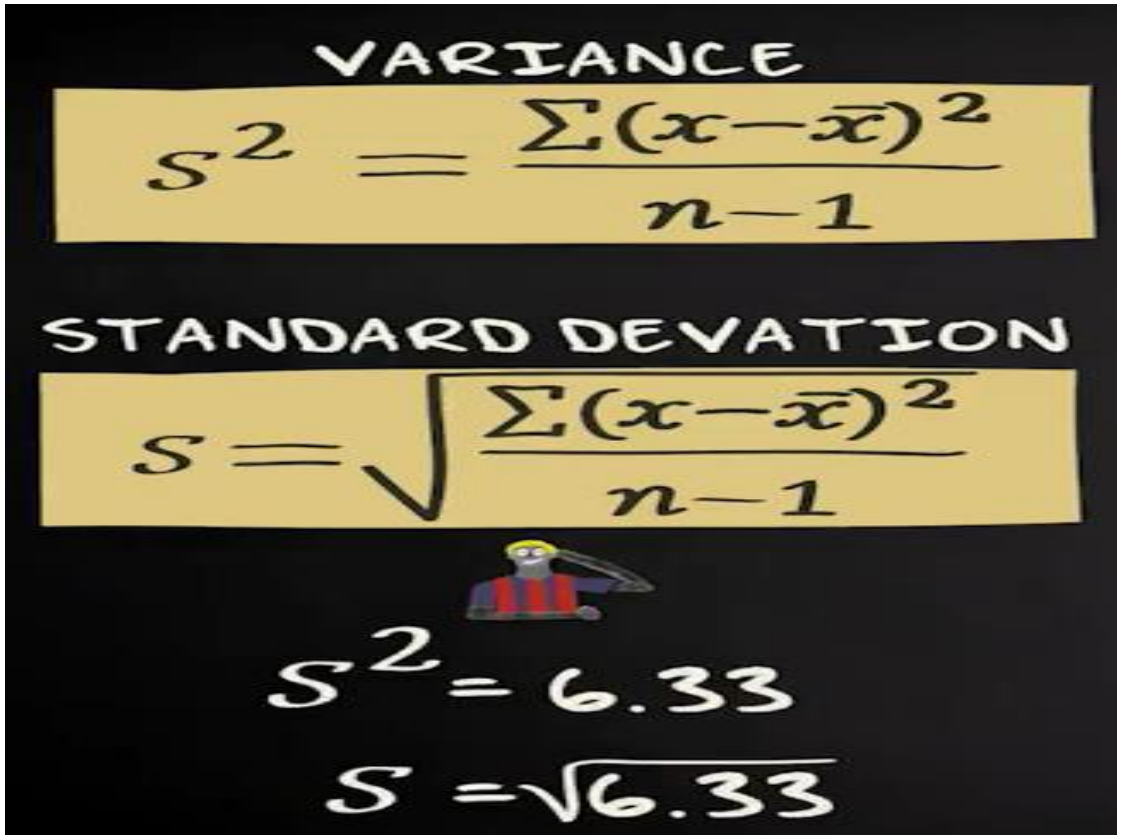
❑ VARIANCE (UNGROUPED DATA)





VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (UNGROUPED DATA)



$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (GROUPED DATA)

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

x = class midpoint

VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (GROUPED DATA)

$$s = \sqrt{\frac{\sum x^2 f - \frac{(\sum xf)^2}{n}}{n-1}}$$

x = class midpoint

VARIANCE AND STANDARD DEVIATION

❑ VARIANCE (GROUPED DATA)

Age	Frrquency (f)	Midpoint (x)	X-Mean	$(X\text{-Mean})^2$	$(X\text{-Mean})^2 f$
30-34	4	32	-9	81	324
35-39	5	37	-4	16	80
40-44	2	42	1	1	2
45-49	9	47	6	36	324
Total	20				730

$\sum f = n = 20$

Mean = 820/20 = 41

$\sum (X\text{-Mean})^2 f = 730$

$$S = \sqrt{\frac{730}{20 - 1}}$$

$$= \sqrt{38.42} \approx 6.20$$