A dark green background featuring a complex, glowing yellow network graph. The graph consists of numerous small, glowing yellow dots (nodes) connected by thin, glowing yellow lines (edges) in a non-uniform, organic pattern, resembling a molecular or neural network.

# COURSE NOTES: DESCRIPTIVE STATISTICS

# FIRST WE NEED TO KNOW THE VARIABLES ...



## ***TYPES OF DATA***

**CATEGORICAL**

**NUMERICAL**

## **MEASUREMENT LEVELS**



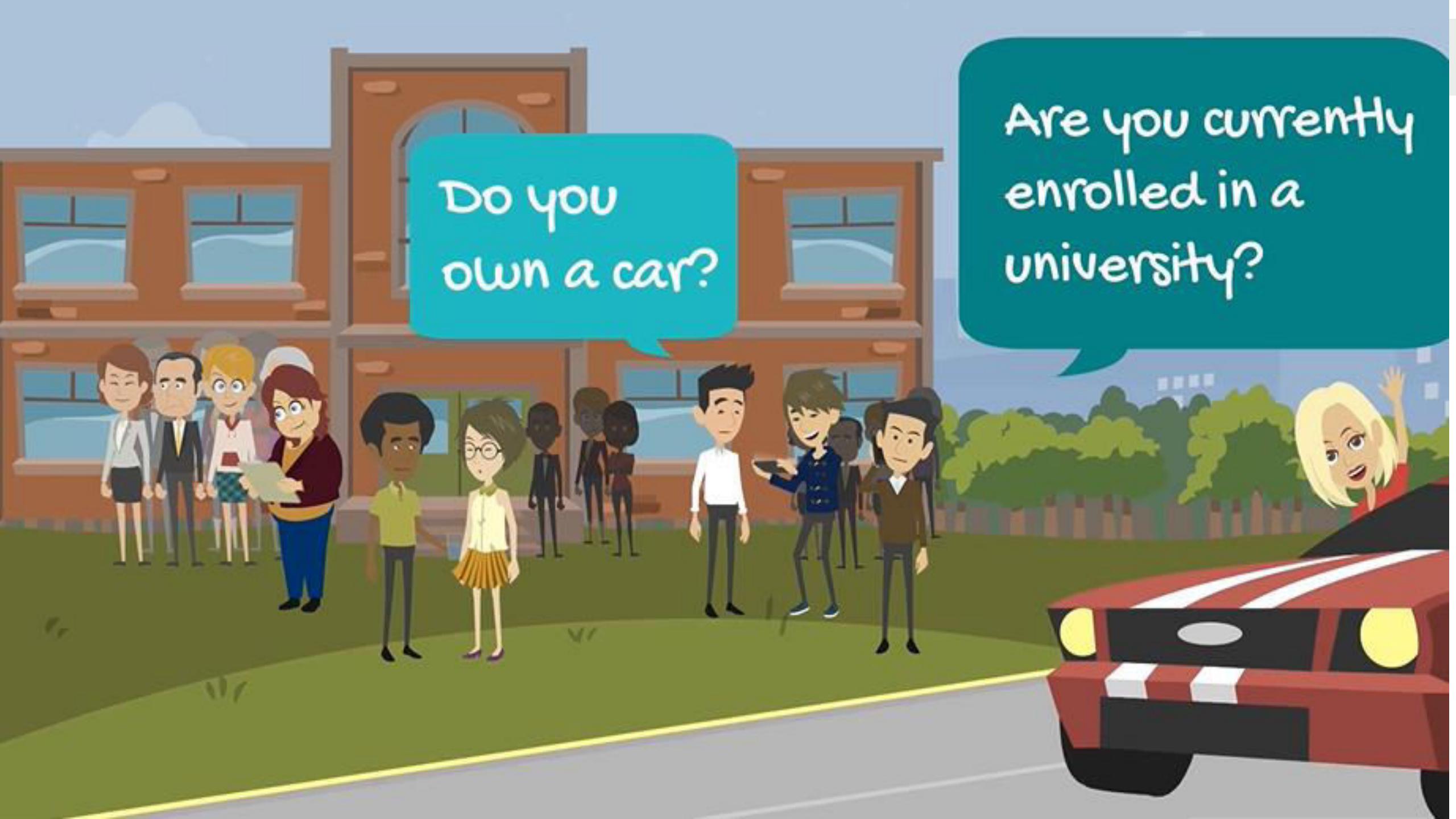
**CATEGORICAL**

Categories or groups

**CAR BRANDS**



**YES AND NO QUESTIONS**



Do you  
own a car?

Are you currently  
enrolled in a  
university?

**TYPES OF DATA**

**MEASUREMENT  
LEVELS**

**CATEGORICAL**

**NUMERICAL**

## **TYPES OF DATA**

## **MEASUREMENT LEVELS**

**CATEGORICAL**

**NUMERICAL**

**DISCRETE**

**CONTINUOUS**

**DISCRETE**

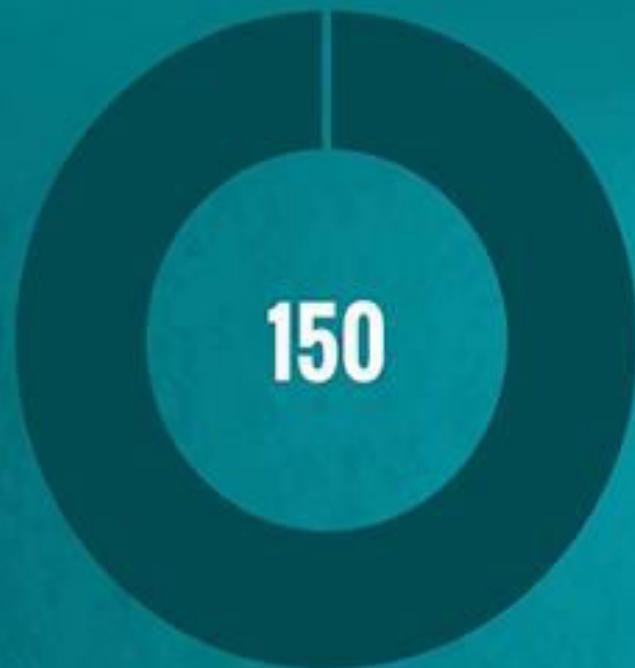
**#** of children

S.A.T score



1000; 1560; 1570; 2400

**CONTINUOUS**



68.0389 kg



**CONTINUOUS**

Weight



**DISCRETE**

# of children



## EXAMPLES OF DISCRETE

A, B, C,  
D, E, F  
or  
0 to 100%

Grades



#1,000

number of objects



Money

# EXAMPLES OF CONTINUOUS



Height



Area



Distance



Time



***TIME ON A CLOCK IS  
DISCRETE***



***TIME IN GENERAL IS  
CONTINUOUS***

**A variable represents the weight of a person. What type of data does it represent?**

1. **Categorical, discrete**
2. **Categorical, continuous**
3. **Numerical, discrete**
4. **Numerical, continuous**

Answer is

**numerical, continuous**

A variable represents the gender of a person.  
What type of data does it represent ?

- 1. Categorical**
- 2. numerical, discrete**
- 3. Numerical, continuous**

Answer is

**Categorical**

## MEASUREMENT LEVELS

**QUALITATIVE**

**QUANTITATIVE**

**NOMINAL**

**ORDINAL**

**INTERVAL**

**RATIO**

## MEASUREMENT LEVELS

**QUALITATIVE**

**QUANTITATIVE**

**NOMINAL**

**ORDINAL**

**NOMINAL**



1. Which of the following courses would you like to register for? (Choose any 3)

- Web development
- Product design
- Graphics design
- Content writing
- Animation

## II. What toppings would you like on your pizza?

- Pepperoni
- Sausage
- Spinach
- Sardines
- Extra Cheese

# ORDINAL



Disgusting

Unappetizing

Neutral

Tasty

Delicious



-----



# Ordinal Data Collection Tool

---

Which of the following best describes your current level of financial happiness?

- Very happy
- Happy
- Neutral
- Unhappy
- Very unhappy

# Ordinal Data Collection Tool

---

How can you rate your knowledge of Excel?

- Advanced
- Intermediate
- Basic
- Novice
- Zero

## MEASUREMENT LEVELS

**QUALITATIVE**

**QUANTITATIVE**

**NOMINAL**

**ORDINAL**

**INTERVAL**

**RATIO**

**RATIO**

**HAS A TRUE 0**

**INTERVAL**

**DOESN'T HAVE A TRUE 0**

# RATIO



You have 3 times  
as many as I do.

RATIO OF 6/2 IS

3

# RATIO

#



NUMBER OF  
OBJECTS



DISTANCE

TIME

# INTERVAL

DOESN'T HAVE A TRUE

0



# INTERVAL

**TODAY**

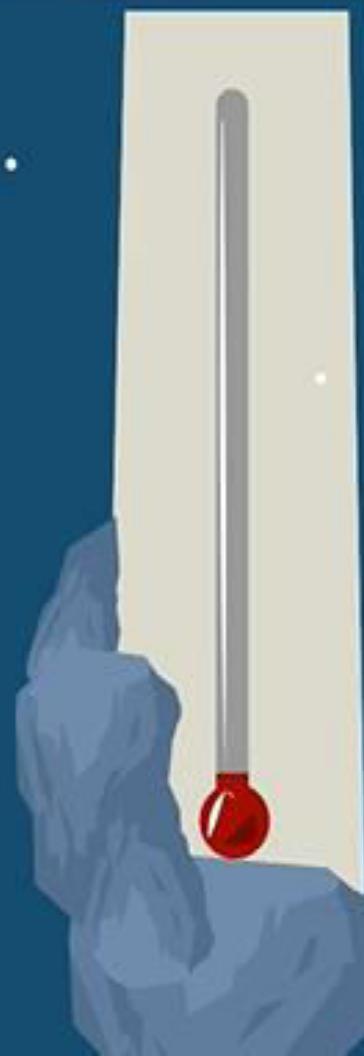
seems colder in  
C°

in terms of  
F°... not really

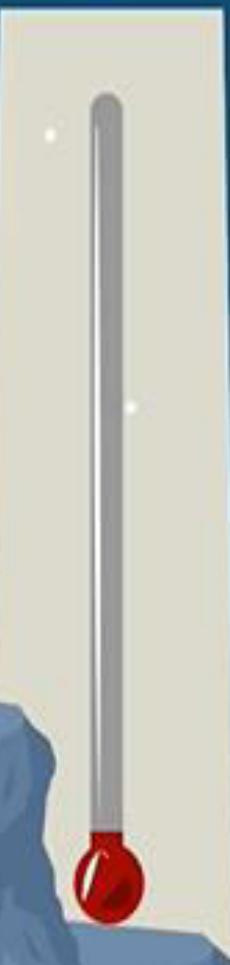


## INTERVAL

$0^{\circ}\text{C}$  AND  $0^{\circ}\text{F}$  ARE NOT TRUE ZEROS

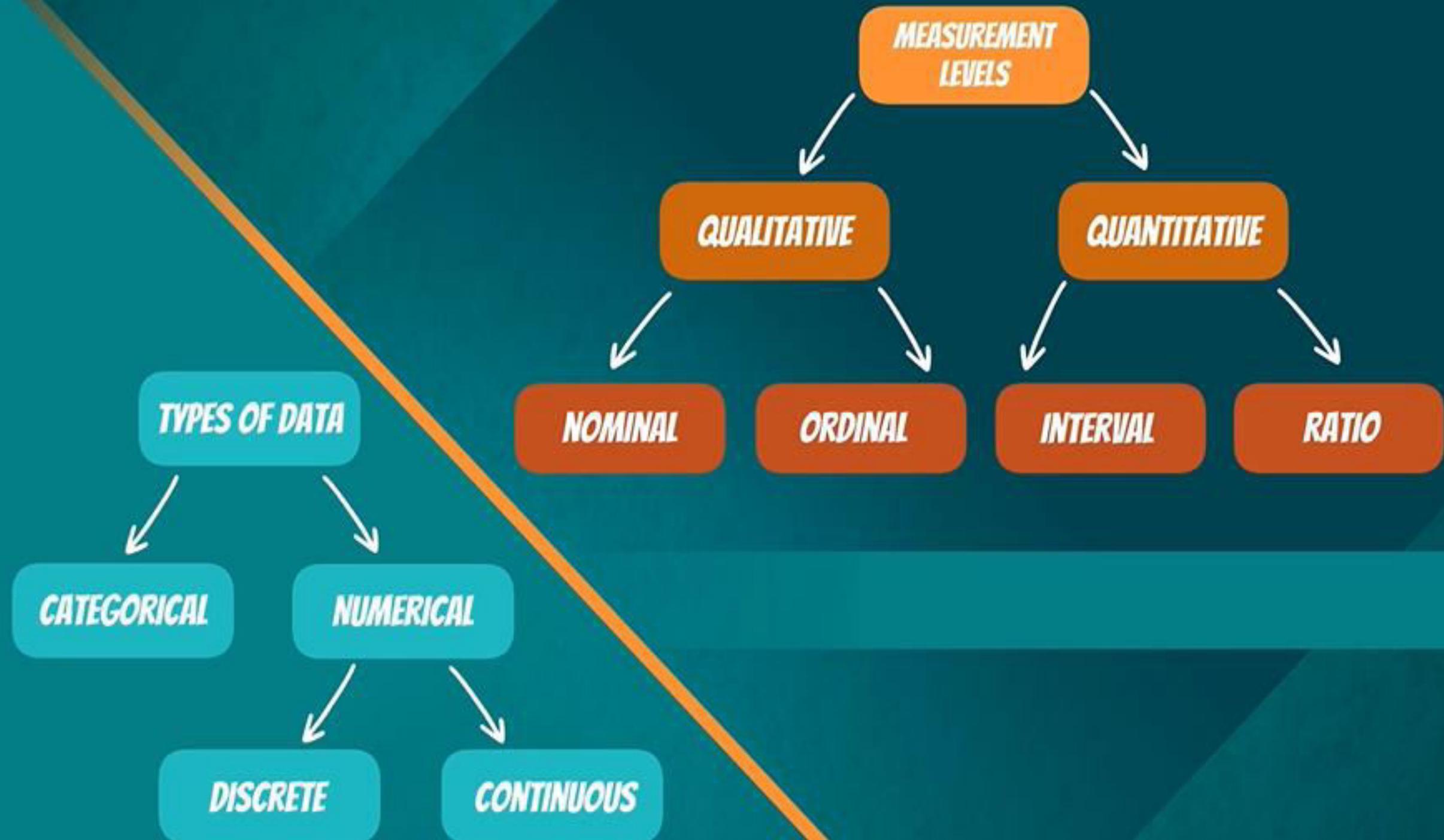


## INTERVAL



$0^\circ$ <sub>C</sub> AND  $0^\circ$ <sub>F</sub> ARE NOT TRUE ZEROS

$0^\circ$ <sub>K</sub> :-  $-273.15^\circ$ <sub>C</sub> :-  $-459.67^\circ$ <sub>F</sub>



Question 1:

**A variable represents the gender of a person. What is the level of measurement?**

nominal

ordinal

interval

ratio

Question 2:

**A variable represents the weight of a person. What is the level of measurement?**

nominal

ordinal

interval

ratio

***VISUALIZING DATA IS THE MOST INTUITIVE  
WAY TO INTERPRET IT***





## **TYPES OF DATA**

**CATEGORICAL**

**NUMERICAL**

## **MEASUREMENT LEVELS**



# REPRESENTATION OF CATEGORICAL VARIABLES

Frequency  
distribution  
tables



Bar charts

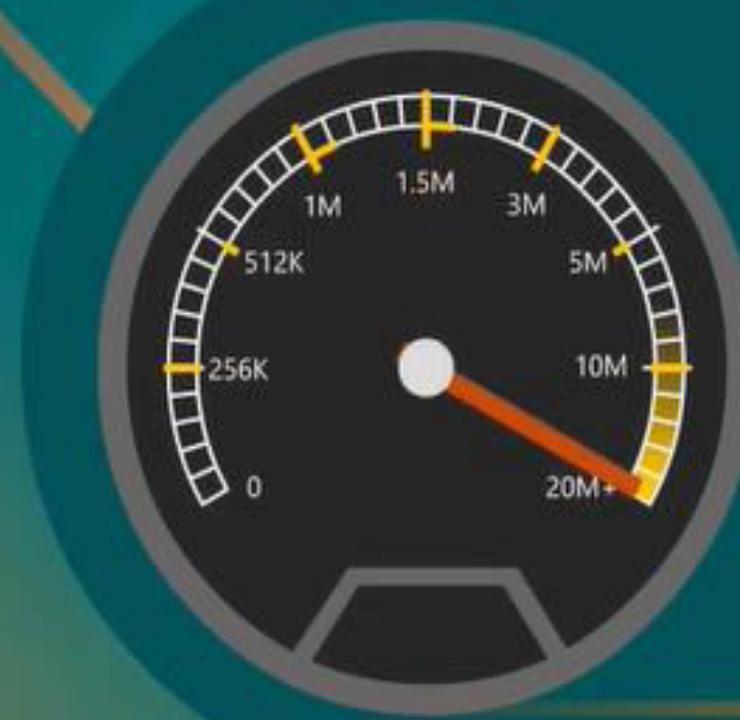


Pie charts



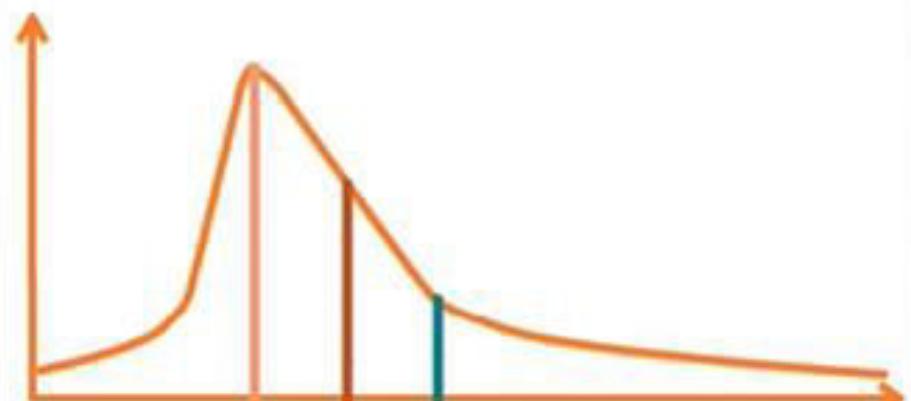
Pareto  
diagrams





## MEASURES OF CENTRAL TENDENCY

# MEASURES OF CENTRAL TENDENCY



Mean

Median

Mode

# MEASURES OF CENTRAL TENDENCY



a.k.a. simple average

MEAN

Population  $\mu$

Sample  $\bar{x}$

## HOW DO WE FIND THE MEAN

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

By adding up all the components and then dividing by the number of components

or

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

Cut  
Copy  
Format Painter

Paste

Font: Arial, Size: 9, Bold, Italic, Underline, Wrap Text, Alignment: Merge & Center, Number: General, Conditional Formatting, Format as Table, Cell Styles, Cells: Insert, Delete, Format, Editing: AutoSum, Fill, Clear, Sort & Find & Filter, Select

A1

A B C D E F G H I J K L M N C

## Mean, median, mode

Pizza prices example

### New York City Los Angeles

\$ 1.00	\$ 1.00
\$ 2.00	\$ 2.00
\$ 3.00	\$ 3.00
\$ 3.00	\$ 4.00
\$ 5.00	\$ 5.00
\$ 6.00	\$ 6.00
\$ 7.00	\$ 7.00
\$ 8.00	\$ 8.00
\$ 9.00	\$ 9.00
\$ 11.00	\$ 10.00
<hr/>	
\$ 66.00	

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste Clipboard Arial 9 A A Wrap Text General Conditional Format as Table Cell Styles Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select

Font Alignment Number Styles Cells Editing

A1 X ✓ fx

A B C D E F G H I J K L M N C

1 Mean, median, mode

2 Pizza prices example

3

4 New York City Los Angeles

5 \$ 1.00 \$ 1.00

6 \$ 2.00 \$ 2.00

7 \$ 3.00 \$ 3.00

8 \$ 3.00 \$ 4.00

9 \$ 5.00 \$ 5.00

10 \$ 6.00 \$ 6.00

11 \$ 7.00 \$ 7.00

12 \$ 8.00 \$ 8.00

13 \$ 9.00 \$ 9.00

14 \$ 11.00 \$ 10.00

15 \$ 66.00

Mean in NY = 
$$\frac{1+2+3+3+5+6+7+8+9+11+66}{11}$$

Mean in LA = 
$$\frac{1+2+3+4+5+6+7+8+9+10}{10}$$

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial, Size: 9, Bold, Italic, Underline, Font Color: Yellow, Font Color: Red

Wrap Text, General, Conditional Formatting, Format as Table, Cell Styles, Insert, Delete, Format Cells

AutoSum, Fill, Clear, Sort & Filter, Select

Clipboard

A1

A B C D E F G H I J K L M N C

1 Mean, median, mode

2 Pizza prices example

3

4 Position New York City Los Angeles

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	

5 New York City Los Angeles

6 Mean \$ 11.00

7 Median \$ 5.50

16

17

18

19

20

21

22

23

The median is the number at position  $(n+1)/2$  in the ordered list

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial Size: 9 Bold Italic Underline Alignment: Wrap Text Number: General Conditional Format as Table Cell Styles Insert Delete Format

AutoSum Fill Clear Sort & Filter Select

Clipboard

Font Alignment Number Styles Cells Editing

A1

A B C D E F G H I J K L M N C

1 Mean, median, mode

2 Pizza prices example

3

4 Position New York City Los Angeles

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	

5 New York City Los Angeles

6 Mean \$ 11.00 \$ 5.50

7 Median \$ 6.00

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

Median in NYC =  $(11+1)/2 = 6$ th position

Dataset mean Median Median2 Mode Mode2

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial, Size: 9, Bold, Italic, Underline, Font Color: Yellow, Font Style: A. Alignment: Wrap Text, General, Merge & Center, Conditional Formatting. Number: \$, %, 00.00, Number Format: Cell Styles, Insert, Delete, Format Cells, AutoSum, Fill, Sort & Find & Filter, Clear, Select.

Font: B, I, U, Font Color: Red, Font Style: A. Alignment: Wrap Text, General, Merge & Center, Conditional Formatting. Number: \$, %, 00.00, Number Format: Cell Styles, Insert, Delete, Format Cells, AutoSum, Fill, Sort & Find & Filter, Clear, Select.

Clipboard: Font, Alignment, Number, Styles, Cells, Editing.

A1

A B C D E F G H I J K L M N C

1 Mean, median, mode

2 Pizza prices example

3

4 Position New York City Los Angeles

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	

5 New York City Los Angeles

6 Mean \$ 11.00 \$ 5.50

7 Median \$ 6.00

8

9

10

11

12

13

14

15

16

17

18

19

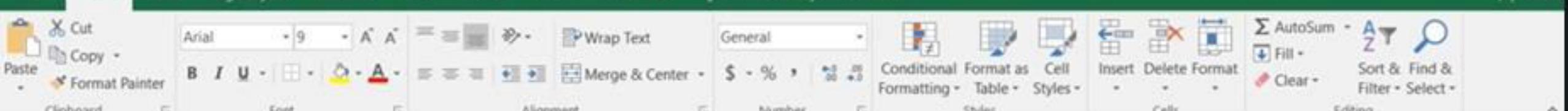
20

21

22

23

Median in LA =  $(10+1)/2 = 5.5$ th position



## Mean, median, mode

### Pizza prices example

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste Clipboard Arial 9 Wrap Text General Conditional Formatting Table Insert AutoSum Fill Clear Sort & Filter

Font B I U Merge & Center \$ % Number Cell Styles Delete Format

Font Alignment Number Styles Cells Editing

A1 X ✓ fx

Position	New York City	Los Angeles		New York City	Los Angeles
1	\$ 1.00	\$ 1.00		Mean	\$ 11.00
2	\$ 2.00	\$ 2.00		Median	\$ 6.00
3	\$ 3.00	\$ 3.00			\$ 5.50
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				

The mode is the value that occurs most often

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Paste Format Painter

Font: Arial Size: 9 Bold Italic Underline

Wrap Text Alignment: General

Conditional Formatting

Format as Table

Cell Styles

Insert Delete Format

AutoSum

Fill

Clear

Sort & Filter

Find & Select

Cells

Editing

A1

A B C D E F G H I J K L M N C

1 Mean, median, mode

2 Pizza prices example

3

4 Position New York City Los Angeles

5 1	\$ 1.00	\$ 1.00
6 2	\$ 2.00	\$ 2.00
7 3	\$ 3.00	\$ 3.00
8 4	\$ 3.00	\$ 4.00
9 5	\$ 5.00	\$ 5.00
10 6	\$ 6.00	\$ 6.00
11 7	\$ 7.00	\$ 7.00
12 8	\$ 8.00	\$ 8.00
13 9	\$ 9.00	\$ 9.00
14 10	\$ 11.00	\$ 10.00
15 11	\$ 66.00	

5 11.00

6 6.00

7 7.00

8 8.00

9 9.00

10 10.00

11 66.00

4 New York City Los Angeles

5 Mean	\$ 11.00	\$ 5.50
6 Median	\$ 6.00	\$ 5.50
7 Mode	\$ 3.00	

5 11.00

6 6.00

7 7.00

8 8.00

9 9.00

10 10.00

11 66.00

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste Clipboard Arial 9 A A Wrap Text General \$ % Conditional Formatting Table Cell Styles Insert Delete Format AutoSum Fill Clear Sort & Filter Select

A1 X ✓ fx

A B C D E F G H I J K L M N C

1 Mean, median, mode

2 Pizza prices example

3

4 Position New York City Los Angeles

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	

5 New York City Los Angeles

6 Mean \$ 11.00 \$ 5.50

7 Median \$ 6.00 \$ 5.50

8 Mode \$ 3.00 -

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

Which measure is best?

## Mean, median, mode

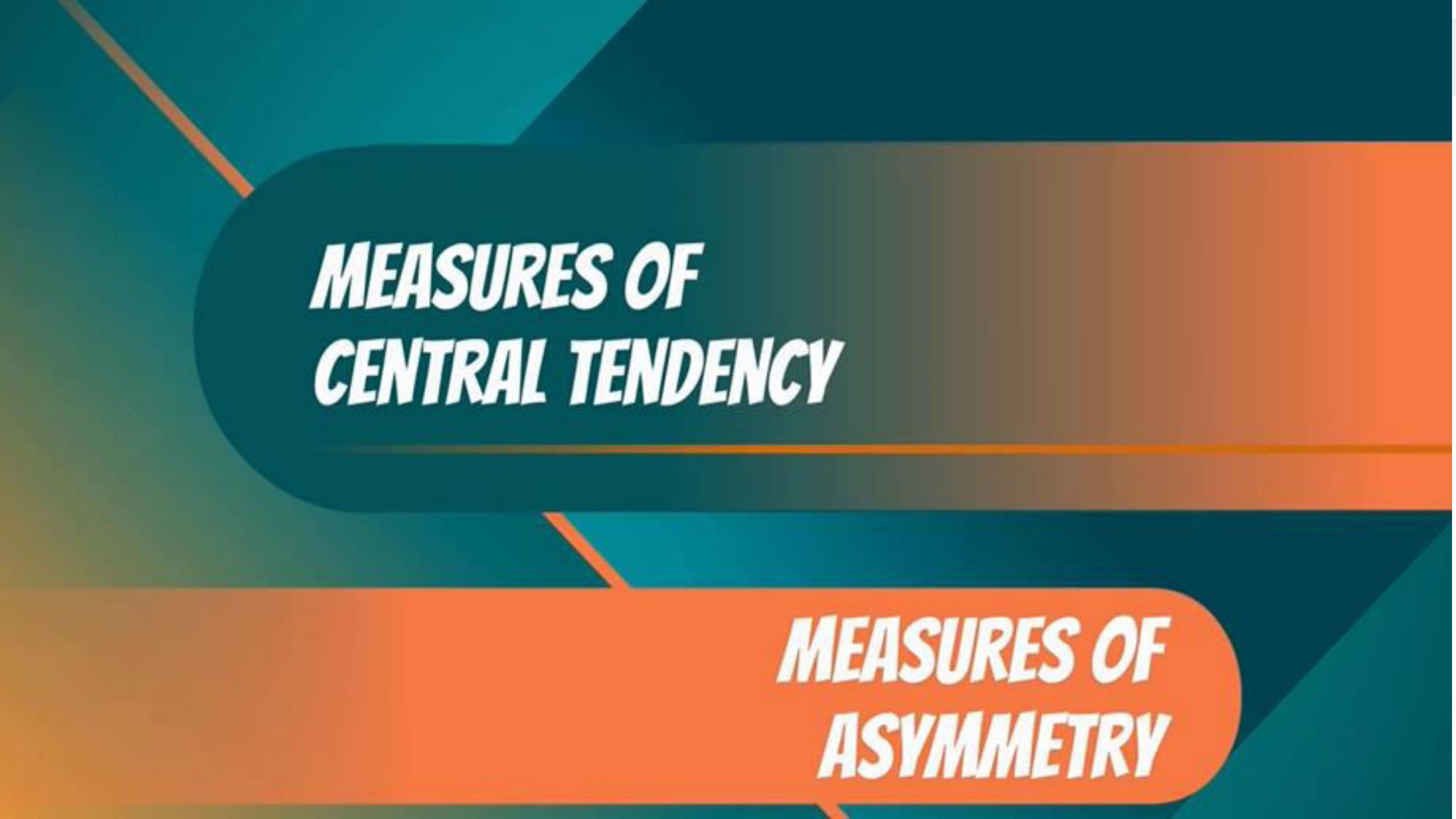
### Pizza prices example

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!

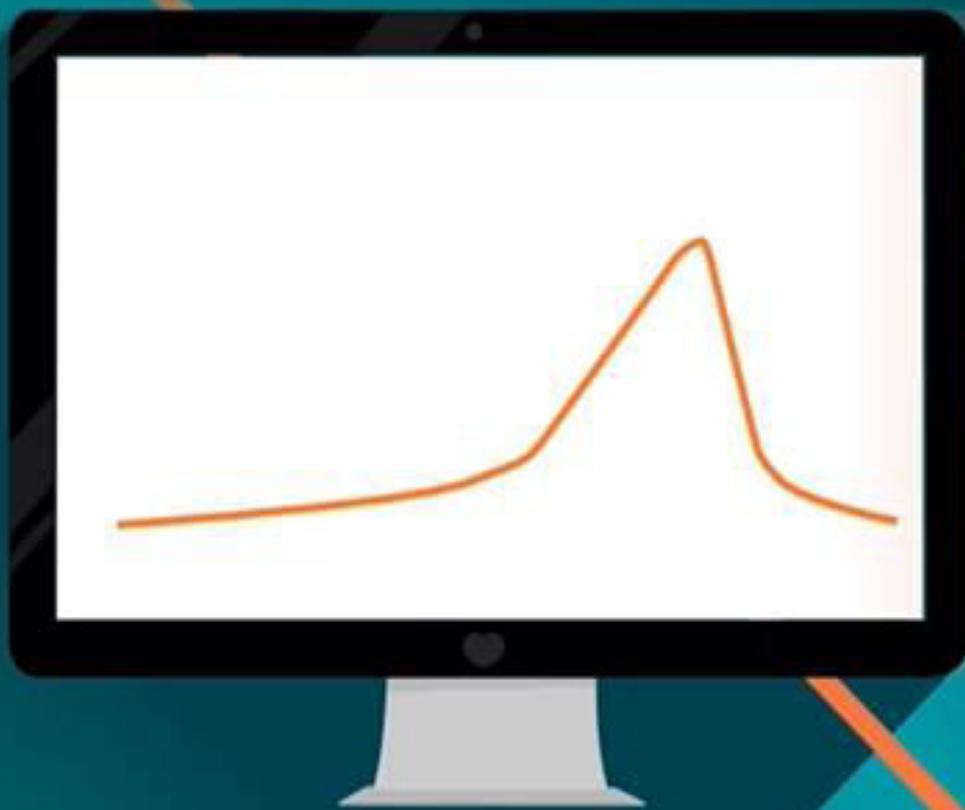


## **MEASURES OF CENTRAL TENDENCY**

## **MEASURES OF ASYMMETRY**

# SAMPLE SKEWNESS FORMULA

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



Almost always, you will  
use software that  
performs the calculation  
for you



Skewness indicates  
whether the data is  
concentrated on one side

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste Arial - 9 A A Wrap Text General Conditional Format as Cell Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select Clipboard Font Alignment Number Styles Cells Editing

A1 X ✓ fx

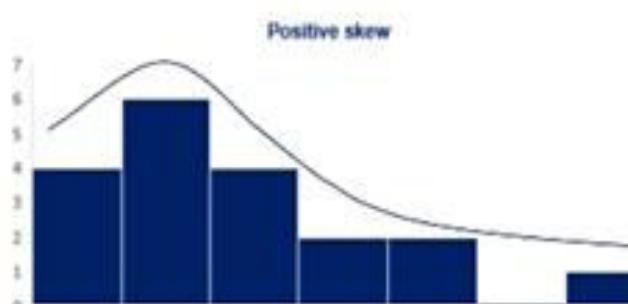
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA AB

Skewness

Positive (right)

Dataset 1			Dataset 2			Dataset 3		
	Interval	Frequency		Interval	Frequency		Interval	Frequency
1	0 to 1	4	1	0 to 1	2	1	0 to 1	1
1	1 to 2	6	1	1 to 2	4	2	1 to 2	1
1	2 to 3	4	2	2 to 3	3	3	2 to 3	2
1	3 to 4	2	2	3 to 4	5	3	3 to 4	3
2	4 to 5	2	3	4 to 5	3	4	4 to 5	4
2	5 to 6	0	3	5 to 6	2	4	5 to 6	6
2	6 to 7	1	3	6 to 7	2	4	6 to 7	3
2			4			5		
2			4			5		
3	Mean	Median	Mode	4	Mean	Median	Mode	5
3	2.75	2.00	2.00	4	4.00	4.00	4.00	5
3				5				6
3				5				6
3				6				6
4				6				6
4				7				7
5								
5								
7								

mean > median



41

1748

### Position (right)

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

mean > median

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

Mean Median Med

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

Mean Median Med



Skewness

**Positive (right)**

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

Mean Median Mode

### Zero (no skew)

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

Mean Median Mode

### Negative (left)

Dataset 3	Interval	Frequencies
1	0 to 1	
2	1 to 2	
3	2 to 3	
3	3 to 4	
4	4 to 5	
4	5 to 6	
4	6 to 7	

Mean Median Mod

mean = median = mode



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste Clipboard

Font: Arial Size: 9 Bold Italic Underline Color: Yellow Red

Wrap Text Alignment: Merge & Center Number: \$ % , .

Conditional Formatting Table Styles Cell Styles

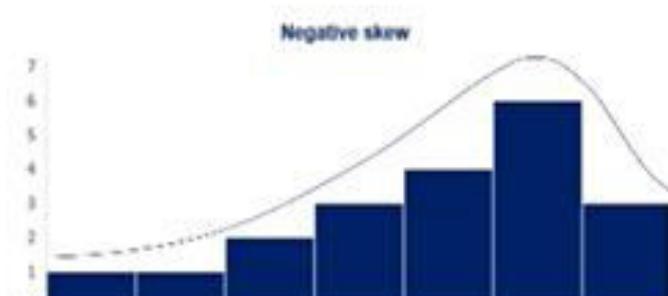
AutoSum Fill Clear Sort & Filter

A1

Skewness

Positive (right)			Zero (no skew)			Negative (left)		
Dataset 1	Interval	Frequency	Dataset 2	Interval	Frequency	Dataset 3	Interval	Frequency
1	0 to 1	4	1	0 to 1	2	1	0 to 1	1
1	1 to 2	6	1	1 to 2	2	2	1 to 2	1
1	2 to 3	4	2	2 to 3	3	2	2 to 3	2
1	3 to 4	2	2	3 to 4	5	3	3 to 4	3
2	4 to 5	2	3	4 to 5	3	4	4 to 5	4
2	5 to 6	0	3	5 to 6	2	4	5 to 6	6
2	6 to 7	1	3	6 to 7	2	4	6 to 7	3
	Mean	Median	Mean	Median	Mode		Mean	Median
	2.79	2.00	4.00	4.00	4.00		4.90	5.00

mean < median





**Measuring how data is spread out :  
calculating variance**



MEASURES OF

Central tendency

Asymmetry

Variability

WE WILL COVER

- VARIANCE
- STANDARD DEVIATION

➤ COEFFICIENT OF VARIATION

MEASURES OF

Central tendency

Asymmetry

Variability

# DIFFERENT FORMULAS FOR

**SAMPLE DATA**



**POPULATION DATA**



# WHEN YOU TAKE SAMPLE DATA

## SAMPLE DATA



A sample statistic is an approximation of the population parameter

# WHEN YOU TAKE SAMPLE DATA



## SAMPLE DATA

**10** different samples give  
**10** different measures

# WHEN YOU TAKE SAMPLE DATA

FORMULA



population formula

sample formula

# Mean

$$\frac{\sum_{i=1}^n x_i}{n}$$

sample formula

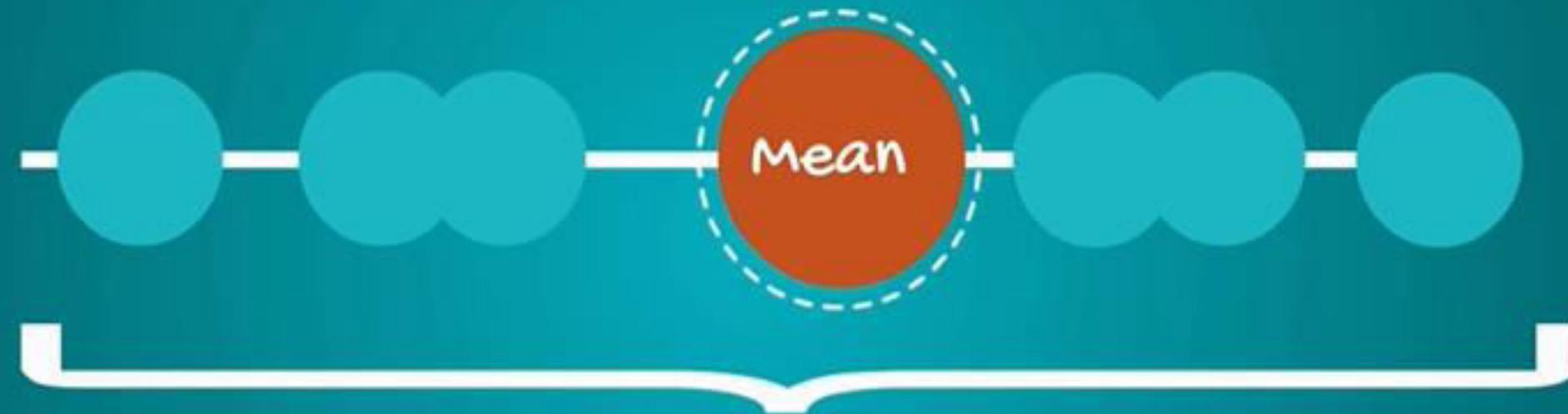
n is the size of  
the sample

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

population formula

N is the size of  
the population

# VARIANCE



Variance measures the dispersion of a set of data points around their mean

# VARIANCE

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

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



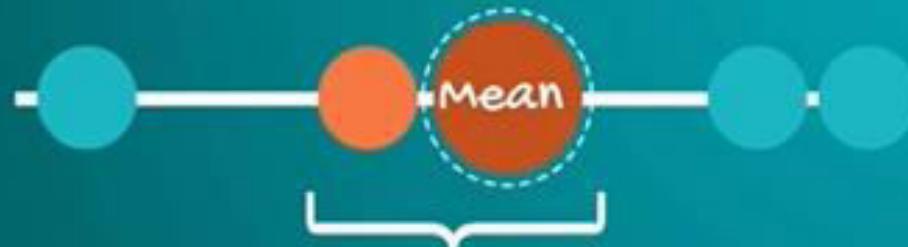
population  
variance



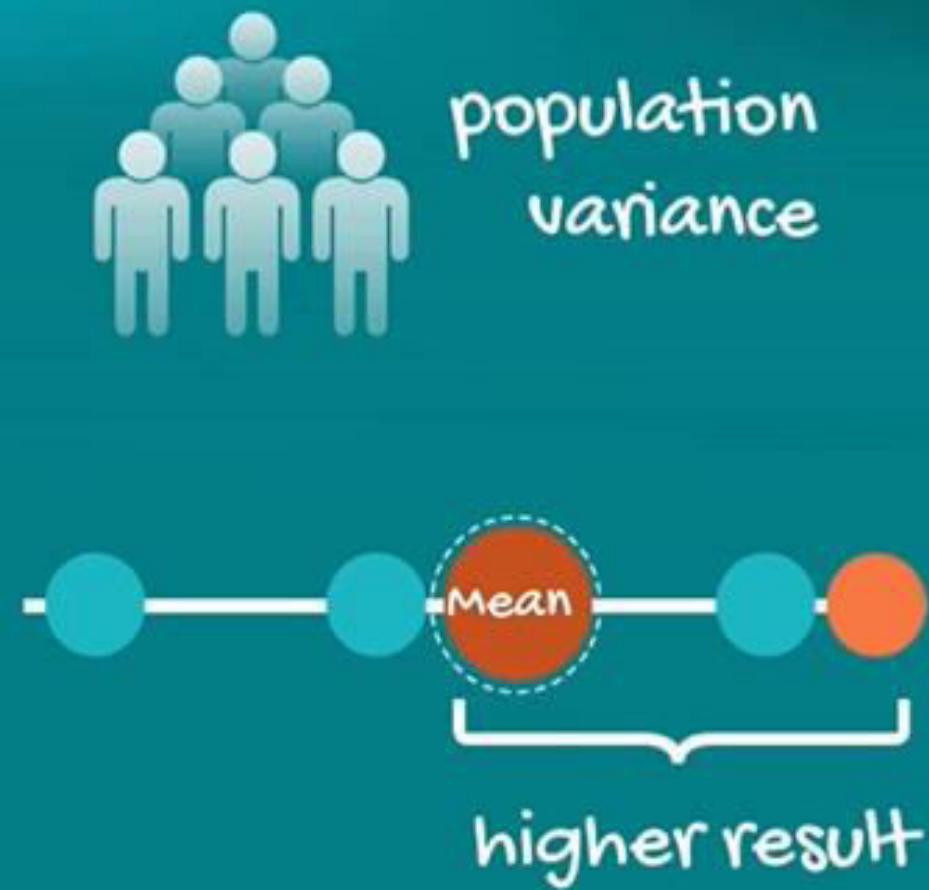
sample  
variance

# VARIANCE

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

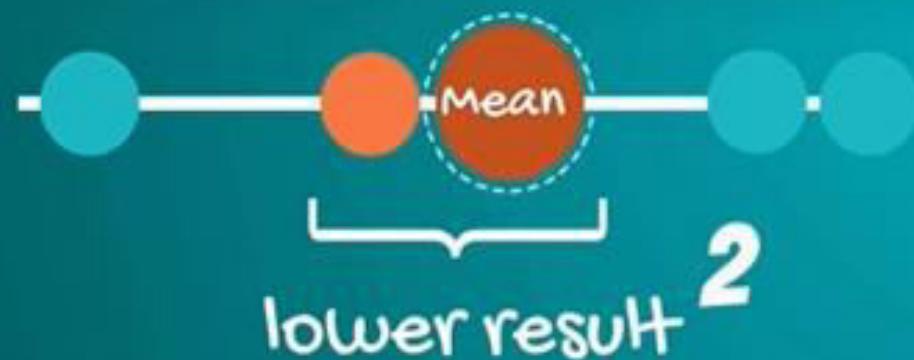


lower result

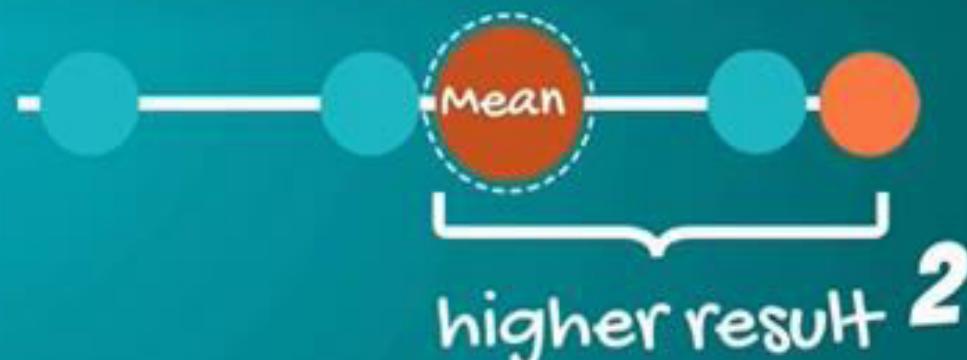


population  
variance

higher result



- Dispersion is non-negative.  
Non-negative values don't cancel out
- Amplifies the effect of large differences



1 Variance

2  
3 Population  
4 1  
5 2  
6 3  
7 4  
8 5  
9

Mean 3.00



We start by calculating the mean. Mean =  $\frac{1+2+3+4+5}{5}$

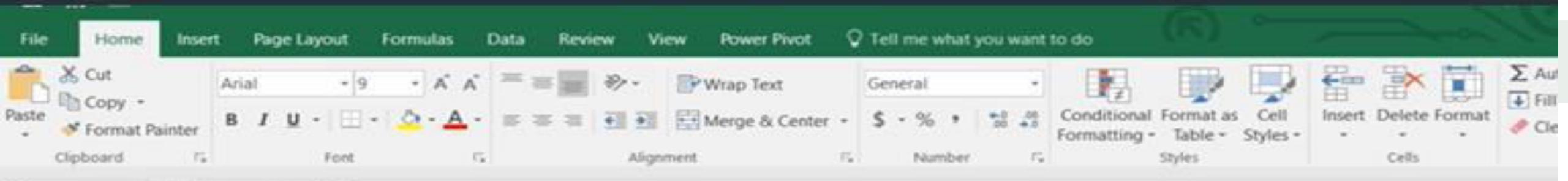
1 Variance

### 3 Population

1	Mean	3.00
2	Population variance	2.00
3		
4		
5		

$$\frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

## Population variance formula



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Paste Format Painter

Font Alignment Number Styles

Insert Delete Format

Σ Au

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Variance													
2														
3	<u>Population</u>													
4	1	Mean		3.00										
5	2	Population variance		2.00										
6	3													
7	4													
8	5													
9														
10														
11														
12														
13														
14														
15														
16														
17														
18														
19														
20														
21														
22														
23														
24														
25														
26														
27														
28														
29														
30														
31														

Sample variance formula is used when our set of observations is a sample drawn from a bigger population



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Variance																

Population		
1	Mean	3.00
2	Population variance	2.00
3		
4		
5		

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{4}$$

Cut  
Copy  
Format Painter

Clipboard

A1

A

Variance

Population

1	Mean	3.00
2	Population variance	2.00
3	Sample variance	2.50
4		
5		

We had all the data and we calculated the variance.  
We had a sample, but did not know the population.  
Therefore, there is more uncertainty.

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial Size: 9 Bold Italic Underline Alignment: Merge & Center Number: General \$ % , . , . Conditional Formatting

Format as Table Cell Styles Insert Delete Cells

AutoSum Fill Clear Sort & Filter Select

A1 X ✓ fx

A B C D E F G H I J K L M N O P Q

1 Variance

2

3 Population

4 1 Mean 3.00

5 2 Population variance 2.00

6 3 Sample variance 2.50

7 4

8 5

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28 Our sample variance has rightfully corrected upwards,

29 in order to reflect the higher potential variability

30

31



WE WILL COVER

- VARIANCE
- STANDARD DEVIATION

➤ COEFFICIENT OF VARIATION

VARIANCE

large units 2



WE WILL COVER

➤ VARIANCE

➤ STANDARD  
DEVIATION

➤ COEFFICIENT OF  
VARIATION

VARIANCE

large  
units 2



## ***WHEN YOU TAKE SAMPLE DATA***

**FORMULA**



**population formula**



**sample formula**

# STANDARD DEVIATION FORMULAS

population standard  
deviation

sample standard  
deviation

# STANDARD DEVIATION FORMULAS

$$\sigma = \sqrt{\sigma^2}$$

population standard deviation

sample standard deviation

$$S = \sqrt{S^2}$$

# COEFFICIENT OF VARIATION (CV)

relative standard deviation

standard deviation

mean

# COEFFICIENT OF VARIATION (CV)

$$C_v = \frac{\sigma}{\mu}$$

Population formula

Sample formula

$$\hat{C}_v = \frac{s}{\bar{x}}$$

 $\sigma$ 

Standard deviation is the  
most common measure of  
variability for a **SINGLE**  
**DATASET**

 $C_v$



$\sigma$

Standard deviation is the  
most common measure of  
variability for a **SINGLE**  
**DATASET**



Comparing **TWO OR**  
**MORE** datasets

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste Arial 9 A A Wrap Text General Conditional Format as Cell Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select

B I U Merge & Center \$ % , . , . Conditional Formatting Table Styles Cell Styles Cells Editing

Clipboard Font Alignment Number Styles

A1 X ✓ fx

A B C D E F G H I J K L M N O P Q R S

1 Standard deviation and coefficient of variation

2 Pizza price example

3

	NY Dollars	Pesos
5	\$ 1.00	MXN 18.81
6	\$ 2.00	MXN 37.62
7	\$ 3.00	MXN 56.43
8	\$ 3.00	MXN 56.43
9	\$ 5.00	MXN 94.05
10	\$ 6.00	MXN 112.86
11	\$ 7.00	MXN 131.67
12	\$ 8.00	MXN 150.48
13	\$ 9.00	MXN 169.29
14	\$ 11.00	MXN 208.91

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste Arial 9 A A Wrap Text General Conditional Formatting Merge & Center \$ % , Number Conditional Formatting Cell Styles Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select Clipboard Font Alignment Number Styles Cells Editing

A1 Standard deviation and coefficient of variation Pizza price example

	NY Dollars	Pesos
5	\$ 1.00	MXN 18.81
6	\$ 2.00	MXN 37.62
7	\$ 3.00	MXN 56.43
8	\$ 3.00	MXN 56.43
9	\$ 5.00	MXN 94.05
10	\$ 6.00	MXN 112.86
11	\$ 7.00	MXN 131.67
12	\$ 8.00	MXN 150.48
13	\$ 9.00	MXN 169.29
14	\$ 11.00	MXN 206.91

Step 1: Sample or population? It is a sample => we have to use the sample formulas

## 1 Standard deviation and coefficient of variation

2 Pizza price example

NY Dollars	Pesos		Dollars	Pesos
\$ 1.00	MXN 18.81	Mean	\$ 5.50	MXN 103.46
\$ 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			

## Step 1: Sample or population?

It is a **sample** => we have to use the **sample** formulas

## Step 2: Find the mean

Cut  
Copy  
Format Painter

Paste

Clipboard

Font

Alignment

Number

Styles

Cells

Editing

Font

&lt;

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste Arial 9 A A Wrap Text General Conditional Format as Table Cell Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select

B I U Merge & Center \$ % , Conditional Format as Table Cell Insert Delete Format AutoSum Fill Clear Sort & Find & Filter Select

Clipboard Font Alignment Number Styles Cells Editing

A1 X ✓ fx

A B C D E F G H I J K L M N O P Q R

1 Standard deviation and coefficient of variation

2 Pizza price example

3

	NY Dollars	Pesos		Dollars	Pesos
5	\$ 1.00	MXN 18.81	Mean	\$ 5.50	MXN 103.46
6	\$ 2.00	MXN 37.62	Sample variance	\$ <sup>2</sup> 10.72	MXN <sup>2</sup> 3793.69
7	\$ 3.00	MXN 56.43	Sample standard deviation	\$ 3.27	MXN 61.59
8	\$ 3.00	MXN 56.43			
9	\$ 5.00	MXN 94.05			
10	\$ 6.00	MXN 112.86			
11	\$ 7.00	MXN 131.67			
12	\$ 8.00	MXN 150.48			
13	\$ 9.00	MXN 169.29			
14	\$ 11.00	MXN 206.91			

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18 Step 1: Sample or population?

19

20

21

22 Step 2: Find the mean

23

24

25

26

27

28 Step 3: Find the sample variance

29

30

31

## Sample standard deviation

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

Step 4: Find the sample standard deviation

### Standard deviation and coefficient of variation

### Pizza price example

NY Dollars	Pesos	Dollars	Pesos
\$ 1.00	MXN 18.81	Mean	\$ 5.50 MXN 103.46
\$ 2.00	MXN 37.62	Sample variance	\$ <sup>2</sup> 10.72 MXN <sup>2</sup> 3793.60
\$ 3.00	MXN 56.43	Sample standard deviation	\$ 3.27 MXN 61.59
\$ 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		

Standard deviation is the preferred measure of variability, as it is directly interpretable

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font Arial Size 9 Bold Italic Underline Merge & Center Wrap Text General Conditional Formatting

Font Alignment Number Styles

Cells Insert Delete Format

AutoSum Fill Clear Sort & Filter

Format Painter

Clipboard

Font Alignment Number Styles

Cells Insert Delete Format

AutoSum Fill Clear Sort & Filter

Format Painter

Clipboard

A1

A B C D E F G H I J K L M N O P Q R

1 Standard deviation and coefficient of variation

2 Pizza price example

3

	NY Dollars	Pesos		Dollars	Pesos
5	\$ 1.00	MXN 18.81	Mean	\$ 5.50	MXN 103.46
6	\$ 2.00	MXN 37.62	Sample variance	\$ <sup>2</sup> 10.72	MXN <sup>2</sup> 3793.69
7	\$ 3.00	MXN 56.43	Sample standard deviation	\$ 3.27	MXN 61.59
8	\$ 3.00	MXN 56.43	Sample coefficient of variation	0.60	0.60
9	\$ 5.00	MXN 94.05			
10	\$ 6.00	MXN 112.86			
11	\$ 7.00	MXN 131.67			
12	\$ 8.00	MXN 150.48			
13	\$ 9.00	MXN 169.29			
14	\$ 11.00	MXN 208.91			
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					

NY Dollars Pesos

Mean

Sample variance

Sample standard deviation

Sample coefficient of variation

Dollars Pesos

103.46

3793.69

61.59

0.60

0.60

18.81

37.62

56.43

56.43

94.05

112.86

131.67

150.48

169.29

208.91

The screenshot shows a Microsoft Excel spreadsheet with the following data and formulas:

	NY Dollars	Pesos
5	\$ 1.00	MXN 18.81
6	\$ 2.00	MXN 37.62
7	\$ 3.00	MXN 56.43
8	\$ 3.00	MXN 56.43
9	\$ 5.00	MXN 94.05
10	\$ 6.00	MXN 112.86
11	\$ 7.00	MXN 131.67
12	\$ 8.00	MXN 150.48
13	\$ 9.00	MXN 169.29
14	\$ 11.00	MXN 206.91

Below the table, the following formulas are listed:

	Dollars	Pesos
Mean	\$ 5.50	MXN 103.48
Sample variance	\$ <sup>2</sup> 10.72	MXN <sup>2</sup> 3793.69
Sample standard deviation	\$ 3.27	MXN 61.59
Sample coefficient of variation	0.60	0.60

**Standard deviation and coefficient of variation**  
Pizza price example

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

$\sigma^2$ 

Variance

 $\sigma$ 

Standard deviation

 $c_v$ 

Coefficient of  
variation

RECAP



Question 1:

Johnny wanted to know how many days per week people in his class exercise, so he asked 3 of his friends. The answers he got were 1, 3 and 5. Based on this, which of the following is true about the results Johnny found?

*Hint: Do the numbers [1,3,5] represent a population or a sample?*

$\mu = 3, \sigma = 2$

$\bar{x} = 3, \sigma = -2$

$\bar{x} = 3, s = 2$

$\bar{x} = 3, s = 4$

Univariate measures

Central tendency ✓

Asymmetry ✓

Variability ✓



in the next two  
lessons...

## MEASURES OF RELATIONSHIP BETWEEN VARIABLES



bivariate measures

Central tendency ✓

Asymmetry ✓

Variability ✓



in the next two  
lessons...

## MEASURES OF RELATIONSHIP BETWEEN VARIABLES



OUR FOCUS:

- Covariance
- Linear correlation coefficient

# REAL ESTATE

what determines house prices?



# REAL ESTATE

Their size!



22. Calculating and understanding covariance

Press Esc to exit full screen

	Size (ft.)	Price (\$)
6	650	772,000
7	785	998,000
8	1200	1,200,000
9	720	800,000
10	975	895,000

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial Size: 9 Bold Italic Underline Font Color: Yellow, Text Color: Red Alignment: Wrap Text, Merge & Center, Horizontal: Center, Vertical: Middle Number: General, \$, %, 00.00 Conditional Formatting, Table, Cell Styles, Insert, Delete, Format Cells

AutoSum, Fill, Sort & Filter, Find & Select, Clear

Clipboard, Font, Alignment, Number, Styles, Cells, Editing

A1

A B C D E F G H I J K L M N O P Q R S

1 Covariance

2 Housing data

3

4

5 Size (ft.) Price (\$)

6 650 772,000

7 785 998,000

8 1200 1,200,000

9 720 800,000

10 975 895,000

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

Dataset Scatter Mean Covariance

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial, Size: 9, Bold, Italic, Underline, Wrap Text, General, Conditional Formatting, Insert, AutoSum, Sort & Filter.

Font: B I U, Alignment: Merge & Center, Number: \$ % , Styles: Cell Styles, Conditional Format as Table, Cell Styles, Insert, Delete Format, Cells, Sort & Filter, Select.

Clipboard, Alignment, Number, Styles, Cells, Editing.

A1 X ✓ fx

A B C D E F G H I J K L M N O P Q R S

1 Covariance

2 Housing data

3

4

5 Size (ft.) Price (\$)

6 650 772,000

7 785 998,000

8 1200 1,200,000

9 720 800,000

10 975 895,000

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

Price (y)

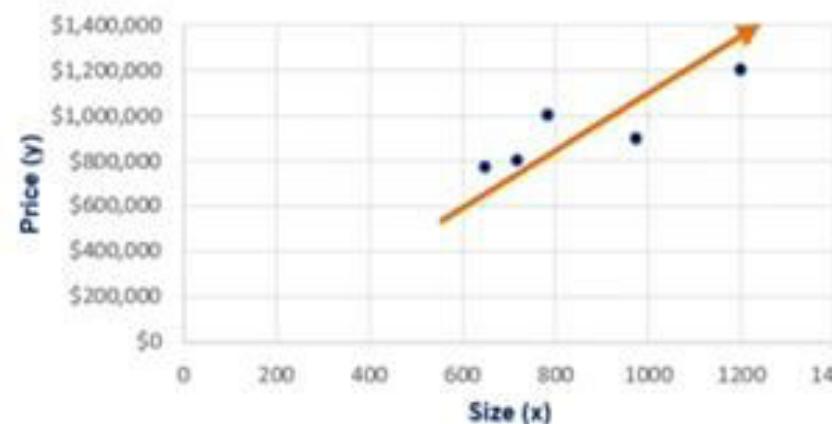
Size (x)

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

## 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



Cut X

Copy Copy

Format Painter Format Painter

Font: Arial, Size: 9, Bold: B, Italic: I, Underline: U, Color: Yellow, Red.

Font: Arial, Size: 9, Bold: B, Italic: I, Underline: U, Color: Yellow, Red.

Wrap Text Wrap Text

General General

Conditional Formatting Conditional Formatting

Format as Table Format as Table

Cell Styles Cell Styles

Insert Insert

Delete Format Delete Format

AutoSum AutoSum

Fill Fill

Sort & Filter Sort & Filter

Clear Clear

Select All Select All

A1 B C D E F G H I J K L M N O P Q R S

1 Covariance Covariance

2 Housing data

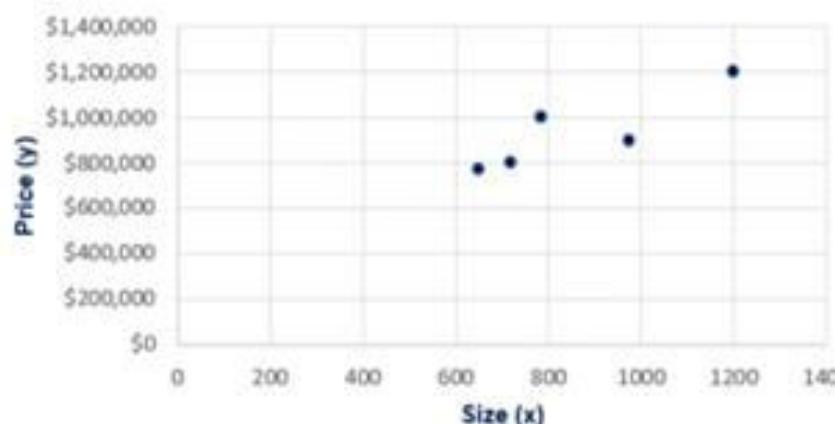
Size (ft.)	Price (\$)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000

## Sample formula

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

## Population formula

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



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial, Size: 9, Bold, Italic, Underline, Merge & Center, Alignment: General, Number: \$ % , Conditional Formatting, Format as Table, Cell Styles, Insert, Delete, Format Cells, AutoSum, Fill, Sort & Find & Select, Clear, Sort & Filter, Select

Clipboard

A1

**Covariance**  
Housing data

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

**Sample formula**

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

**Population formula**

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

Price (Y) vs. Size (X)

Price (Y)

Size (X)

## Covariance Housing data

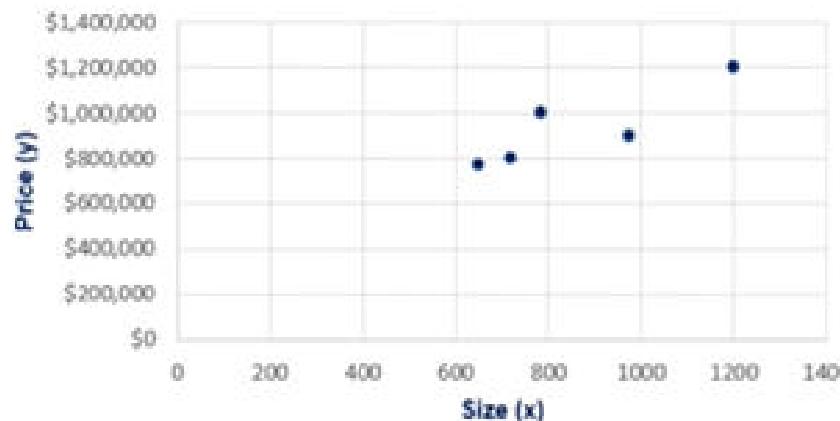
<b>X</b>	<b>y</b>
Size (ft.)	Price (\$)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000
886	933,000

Sum  
Sample size  
Cov. Sample

$$(x - \bar{x})^2 + (y - \bar{y})^2$$

## Sample formula

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

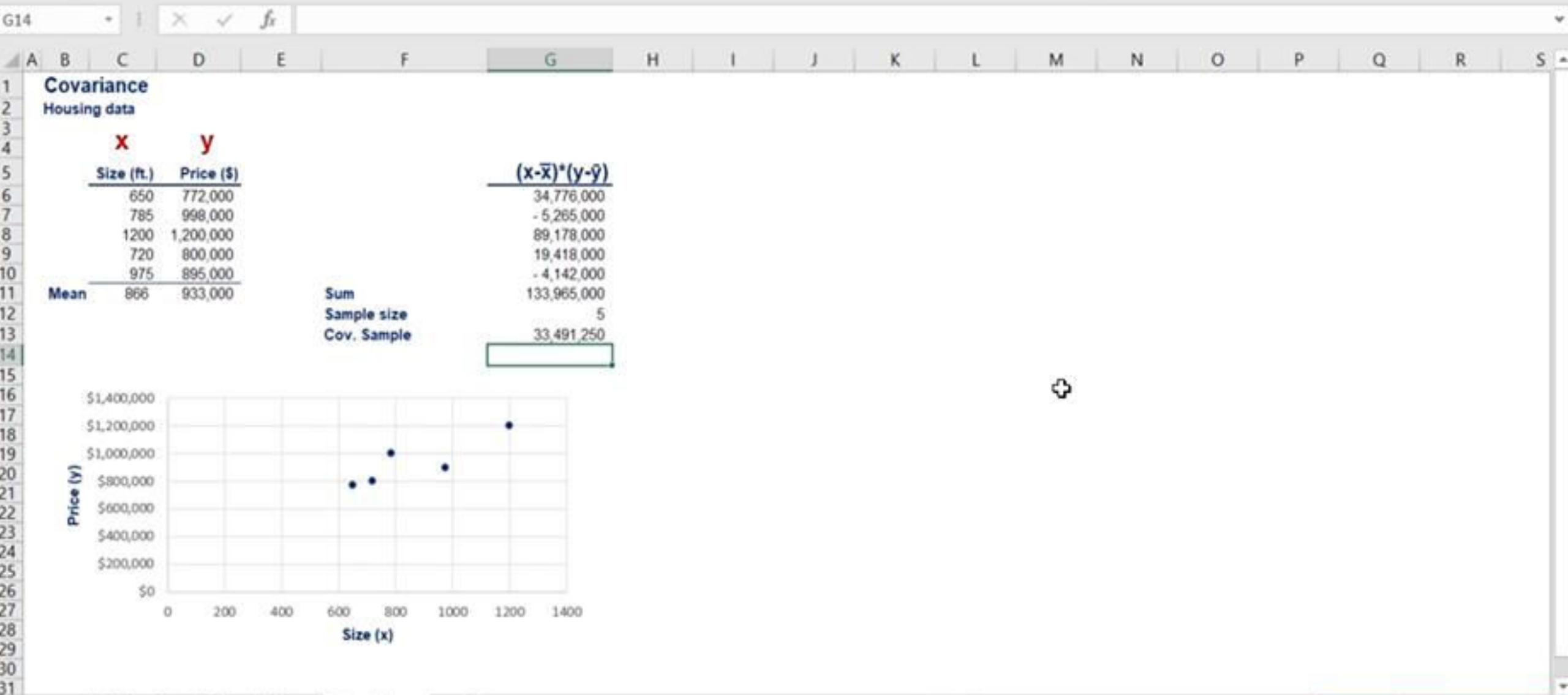


File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

fx **Sum** Trace Precedents Watch Window

Insert Function AutoSum Recently Used Date & Time Text Reference Logical Functions Math & Trig Functions More Functions Name Manager Defined Names Trace Precedents Show Formulas Trace Dependents Error Checking Remove Arrows Evaluate Formula Watch Window Calculation Options Calculate Sheet Calculation

Function Library



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

fx  $\sum$  Insert Function

AutoSum Recently Used Financial Logical Text Date & Time Lookup & Reference Math & Trig More Functions

Function Library

Define Name Use in Formula Trace Precedents Show Formulas Trace Dependents Error Checking Remove Arrows Evaluate Formula Name Manager Create from Selection Defined Names

Watch Window Calculation Options Calculate Now Calculate Sheet Formula Auditing

Calculation

G14

A B C D E F G H I J K L M N O P Q R S

1 Covariance

2 Housing data

3

4

5 **x y**

	<b>x</b>	<b>y</b>
Size (ft.)	Price (\$)	
650	772,000	$(x - \bar{x})(y - \bar{y})$
785	998,000	34,776,000
1200	1,200,000	- 5,265,000
720	800,000	89,178,000
975	895,000	19,418,000
Mean	866	- 4,142,000
		133,965,000
		5
		33,491,250

6

7

8

9

10

11 Sum

12 Sample size

13 Cov. Sample

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

**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

Price (y)

Size (x)

**NEXT LESSON: CORRELATION COEFFICIENT**

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial Size: 9 Bold Italic Underline Color: Yellow Red

Wrap Text Alignment: Merge & Center Number: \$ % , . , .

Conditional Formatting as Table Cell Styles Insert Delete Format

AutoSum Fill Clear Sort & Find & Filter Select

Clipboard Font Alignment Number Styles Cells Editing

A1 X ✓ fx

A B C D E F G H I J K L M N O P Q R

1 Correlation coefficient

2 Housing data

3

4

5 Size (ft.) Price (\$)

6 650 772,000

7 785 998,000

8 1200 1,200,000

9 720 800,000

10 975 895,000

11 Mean 866 933,000

12 Standard dev. 222 173,615

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

$$\frac{(x-\bar{x})(y-\bar{y})}{\text{Stdev}(x) * \text{Stdev}(y)}$$

$$\frac{\text{Cov}(x, y)}{\text{Stdev}(x) * \text{Stdev}(y)}$$

	Sum	5
Mean	133,965,000	
Sample size		5
Cov. Sample	33,491,250	

Price (Y)

Size (X)

Scatter plot showing Price (Y) vs Size (X). The x-axis is labeled 'Size (X)' and ranges from 0 to 1400. The y-axis is labeled 'Price (Y)' and ranges from \$0 to \$1,400,000. Five data points are plotted, showing a positive correlation between size and price.

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial Size: 9 **B** *I* U **A** *A* Wrap Text Alignment: General Merge & Center Number: \$ % , . , . Conditional Formatting: Cell Styles Insert Delete Format

Cells: AutoSum Fill Clear Sort & Find & Filter Select Editing

A1 X ✓ fx

A B C D E F G H I J K L M N O P Q R

1 Correlation coefficient

2 Housing data

3

4

5 Size (ft.) Price (\$)

6 650 772,000

7 785 998,000

8 1200 1,200,000

9 720 800,000

10 975 895,000

11 Mean 900 913,000

12 Standard dev. 222 173,615

13 Sum 133,965,000

Sample size 5

Cov. Sample 33,491,250

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

Price (y)

Size (x)

$$\frac{(x-\bar{x})(y-\bar{y})}{\text{Stdev}(x) * \text{Stdev}(y)}$$

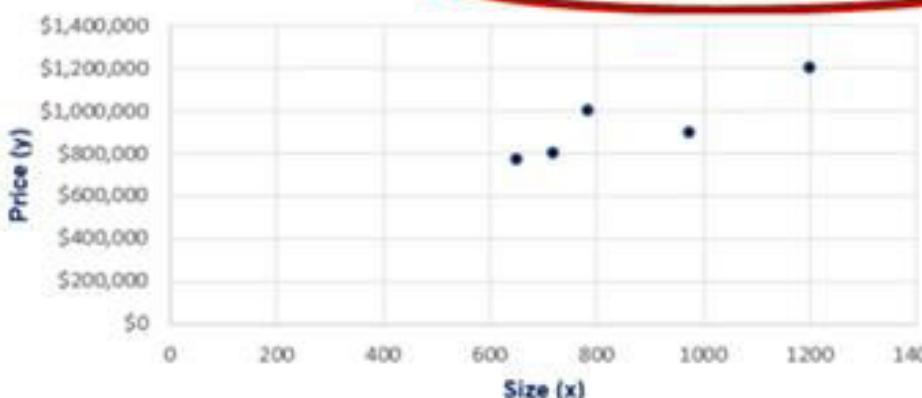
$$\frac{s_{xy}}{s_x s_y}$$

$$\frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

## Correlation coefficient

### Housing data

Size (ft.)	Price (\$)	$(x - \bar{x})^*(y - \bar{y})$
650	772,000	34,776,000
785	998,000	- 5,265,000
1200	1,200,000	89,178,000
720	800,000	19,418,000
975	895,000	- 4,142,000
Mean	866	933,000
Standard dev.	222	173,615
		Sum
		Sample size
		Corr. sample
		Correlation coeff.



-1 ≤ correlation coefficient ≤ 1

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste

Font: Arial Size: 9 Bold Italic Underline Text Color: Yellow Red Alignment: Wrap Text Merge & Center Number: General \$ % , . . . Conditional Formatting Table Styles Insert Delete Format

Cells Editing

Sort & Filter: Select Clear Fill

A1 B C D E F G H I J K L M N O P Q R

Correlation coefficient  
Housing data

	Size (ft.)	Price (\$)	$(x - \bar{x})(y - \bar{y})$
650	772,000	34,776,000	
785	998,000	- 5,265,000	
1200	1,200,000	89,178,000	
720	800,000	19,418,000	
975	895,000	- 4,142,000	
Mean	866	933,000	133,965,000
Standard dev.	222	173,615	33,491,250

Sum  
Sample size  
Cov. Sample  
Correlation coeff. 0.87

Price (y)

Size (x)

There is a STRONG relationship between the two variables

# PERFECT POSITIVE CORRELATION

X      Y

Correlation coeff. = 1

the entire variability of  
one variable is explained  
by the other

# RELATIONSHIP DIRECTION

Size determines price



# ***CORRELATION OF 0***

**Absolutely independent variables**

# CORRELATION OF 0

They have nothing in common!



Coffee in Brazil



Houses in London

# NEGATIVE CORRELATION

Perfect negative correlation of - 1

Imperfect negative correlation: (- 1,0)

# NEGATIVE CORRELATION



# CORRELATION

$$\begin{matrix} x & y \\ \text{---} \\ u & u \end{matrix} = \begin{matrix} y & x \\ \text{---} \\ u & u \end{matrix}$$

# CORRELATION

$$\frac{\text{Cov}(x, y)}{\text{Stdev}(x) * \text{Stdev}(y)} = \frac{\text{Cov}(y, x)}{\text{Stdev}(y) * \text{Stdev}(x)}$$

Symmetrical with respect to both variables

# CAUSALITY

Important to understand the direction of causal relationships

## **CAUSALITY**

Important to understand the direction of causal relationships

**CORRELATION DOES NOT IMPLY CAUSATION**

Question 1:

**Which of these is true about correlation and causality?**

- Causality is a symmetric relation
- Correlation is an asymmetric relation ( $x$  affects  $y$  is different from  $y$  affects  $x$ )
- Causality is an asymmetric relation. ( $x$  causes  $y$  is different from  $y$  causes  $x$ )
- None of the above