Basic Statistics

- Saurav Poudel

Statistics

- Science of gathering, analyzing, interpreting, and presenting data.

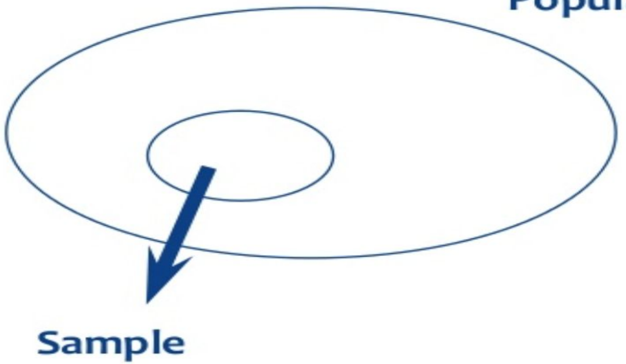
- Branch of Mathematics.

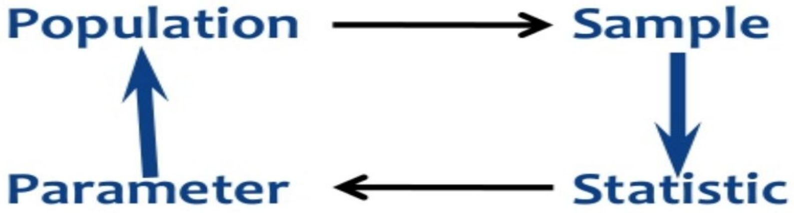
- Foundation of Machine Learning

Example: Height

- Sample vs Population
- Descriptive vs Inferential
- Statistics vs Statistic vs Parameter
- Best Guess of Parameter
- Math or No Math
- Sample Size & Estimation

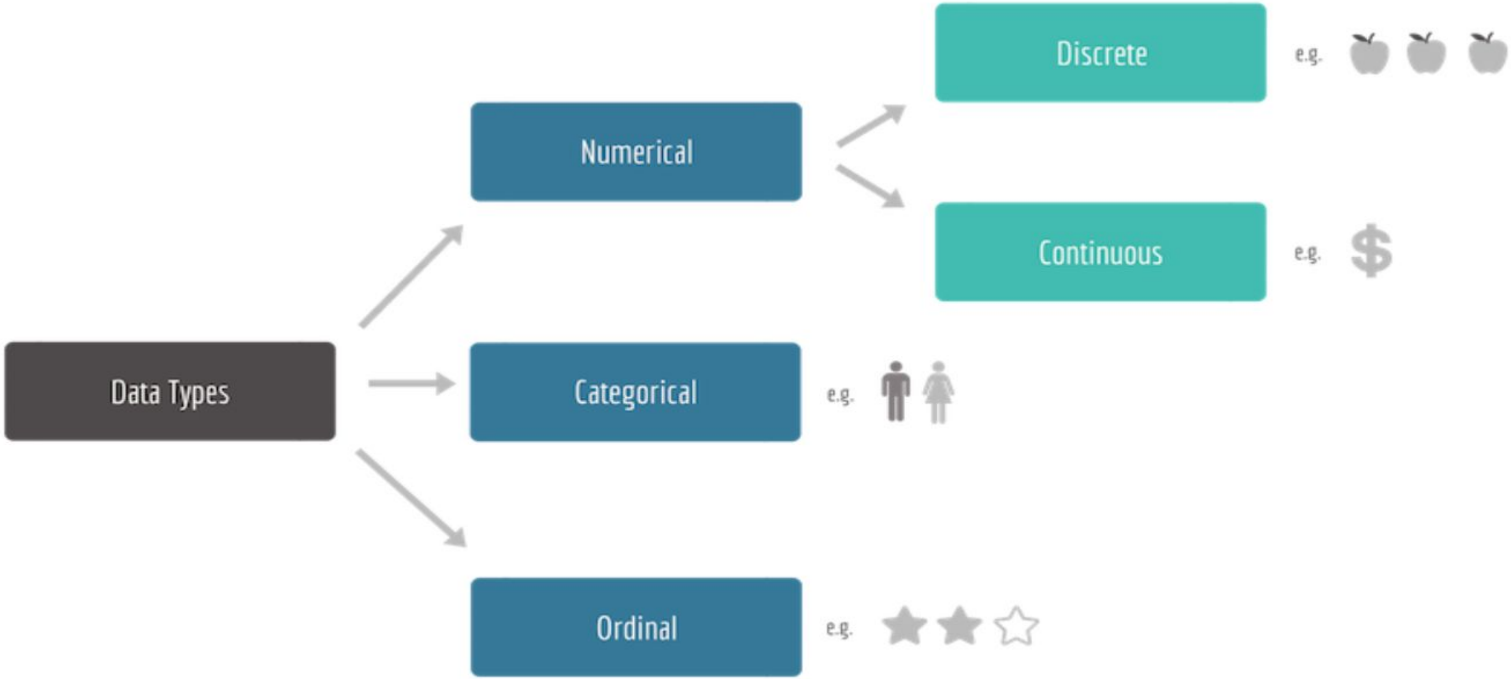
Population Vs. Sample





We measure the sample using statistics in order to draw inferences about the parameters of the population.

Data Types


Data Types
Numerical
Discrete
e.g.
Continuous
e.g.
Categorical
e.g.
Ordinal
e.g.

Can we summarize the data by one point?

Central Tendency

- Mean
- Median
- Mode

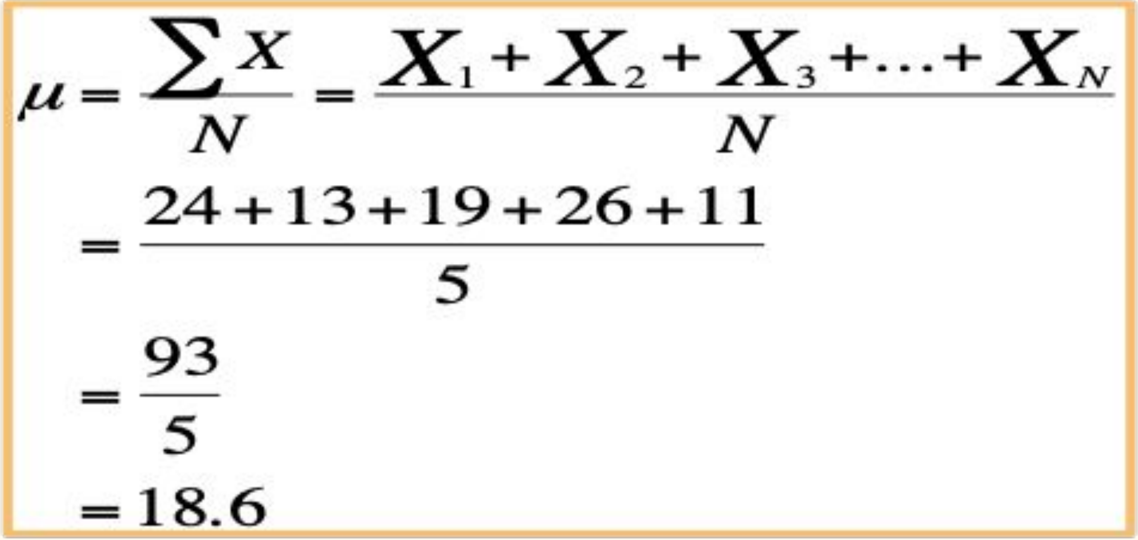
Arithmetic Mean

- Commonly called ‘the mean’
- Is the average of a group of numbers
- Affected by each value in the data set, including extreme values

Arithmetic Mean

Example: Suppose a company has five departments with 24, 13, 19, 26, and 11 workers each. The population mean number of workers in each department is 18.6 workers.

Arithmetic Mean



$$\mu = \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

$$= \frac{24 + 13 + 19 + 26 + 11}{5}$$

$$= \frac{93}{5}$$

$$= 18.6$$

Median

- Middle value in an ordered array of numbers.

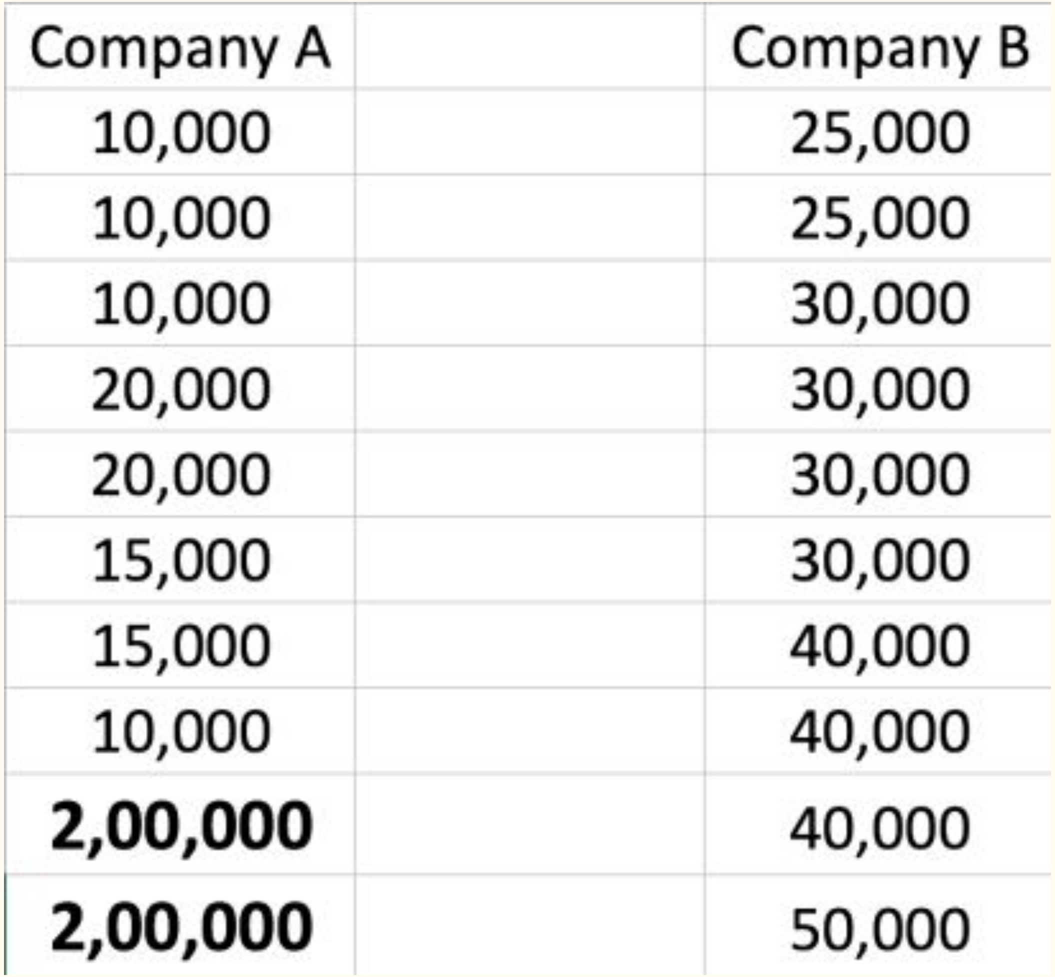
- Arrange the observations in an ordered array.
- If there is an odd number of terms, the median is the middle term of the ordered array.
- If there is an even number of terms, the median is the average of the middle two terms.

Median

- 3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

- There are 16 terms in the ordered array.
- Position of median = (n+1)/2 = (16+1)/2 = 8.5
- The median is between the 8th and 9th terms, 14.5.

- If the 21 is replaced by 100, the median is 14.5.
- If the 3 is replaced by -88, the median is 14.5.



Company A		Company B
10,000		25,000
10,000		25,000
10,000		30,000
20,000		30,000
20,000		30,000
15,000		30,000
15,000		40,000
10,000		40,000
2,00,000		40,000
2,00,000		50,000

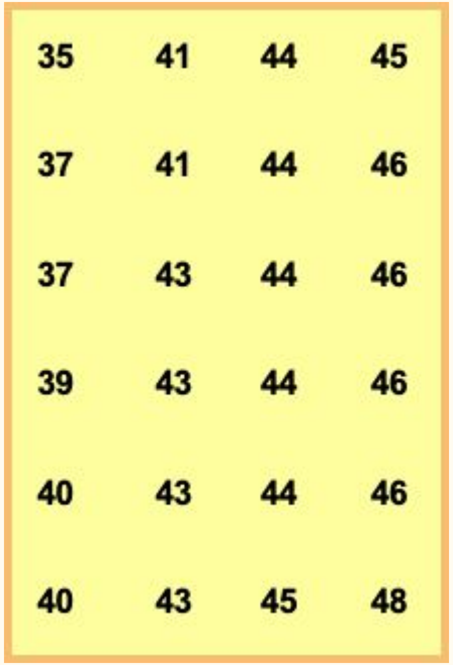
Mode

- The most frequently occurring value in a data set
- Applicable also to Categorical data

- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes

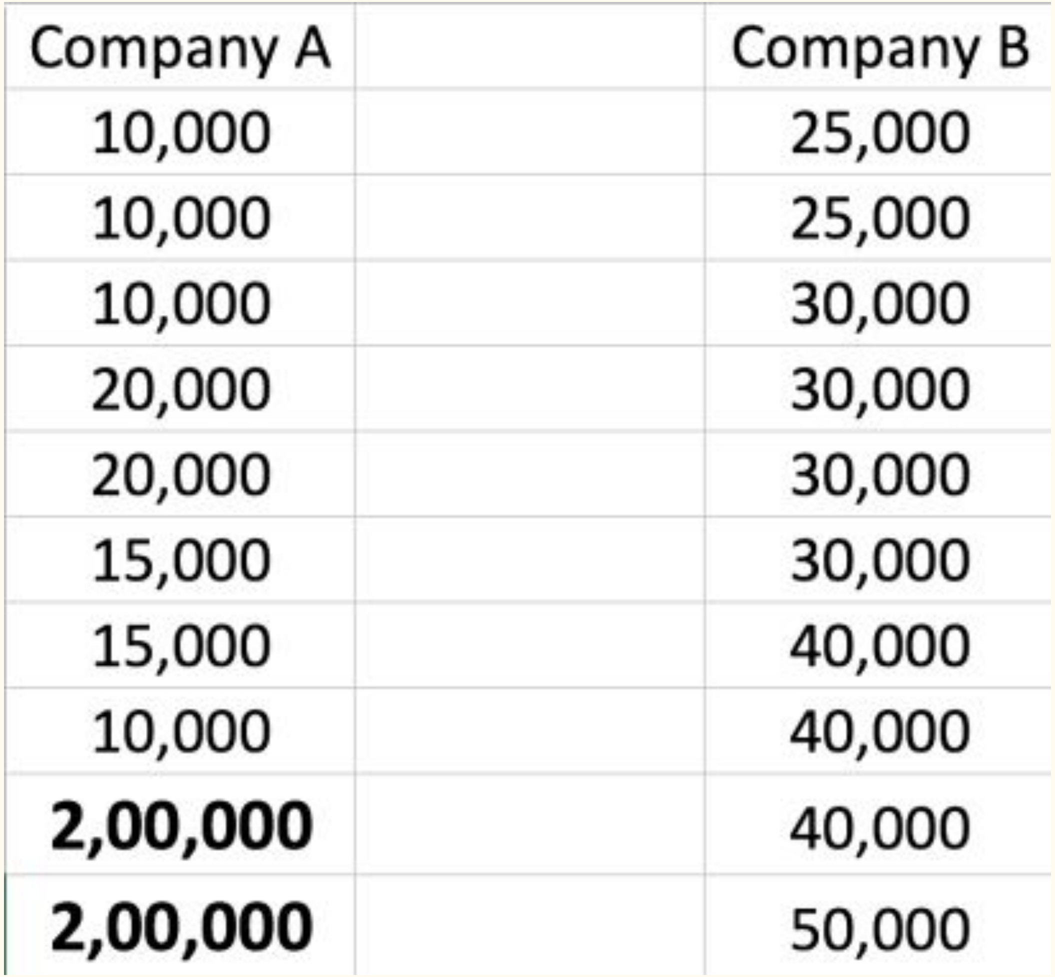
Mode

- The mode is 44.
- There are more 44s than any other value.



35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

Is one point summary of data conclusive ?



Company A		Company B
10,000		25,000
10,000		25,000
10,000		30,000
20,000		30,000
20,000		30,000
15,000		30,000
15,000		40,000
10,000		40,000
2,00,000		40,000
2,00,000		50,000

No Variability in Cash Flow

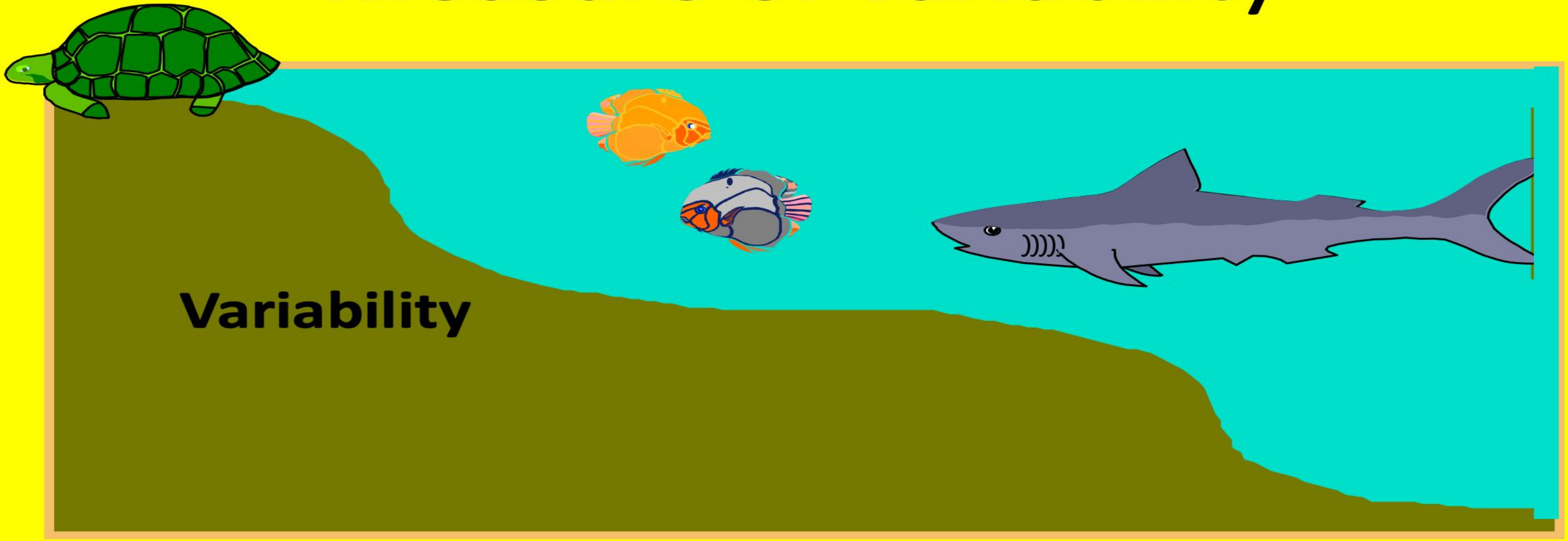
Mean

Variability in Cash Flow

Mean

Variability (Dispersion)

Measure of variability



No Variability

Measure of Variability or Dispersion

Common Measures of Variability

- Range
- Mean Absolute Deviation
- Variance
- Standard Deviation

Measure of Variability or Dispersion

Common Measures of Variability

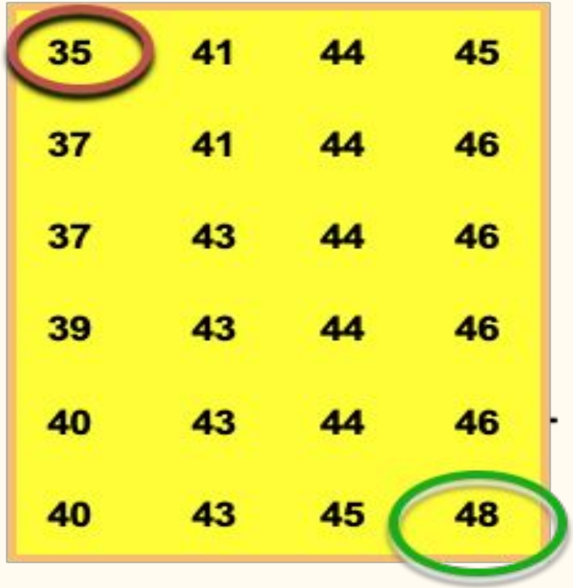
- Range
- Mean Absolute Deviation
- Variance
- Standard Deviation

Range

- The difference between the largest and the smallest values in a set of data
- Simple to compute
- Ignores all data points except extremes
- Example:

  = Largest - Smallest

  35 = 13



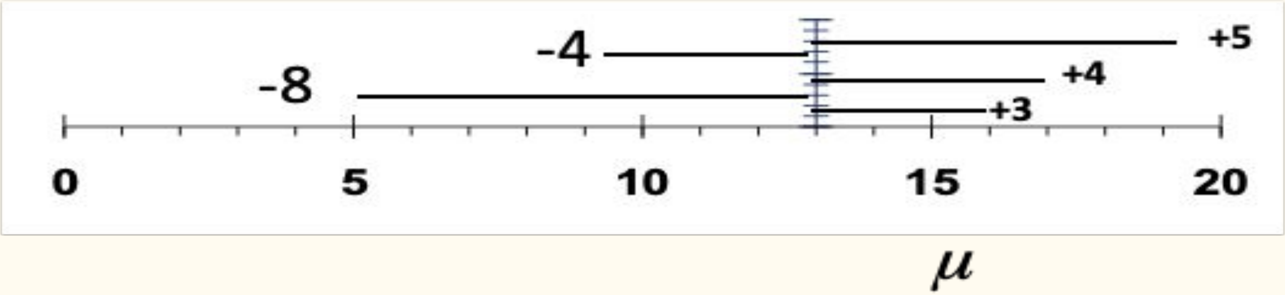
35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

Deviation from the Mean

- Data set: 5, 9, 16, 17, 18
- Mean:

$$\boxed{\mu = \frac{\sum X}{N} = \frac{65}{5} = 13}$$

- Deviations from the mean: -8, -4, 3, 4, 5



Mean Absolute Deviation

- Average of the absolute deviations from the mean

X	$X - \mu$	$\lvert X - \mu \rvert$
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$
\begin{aligned}
M.A.D. &= \frac{\sum \lvert X - \mu \rvert}{N} \\
&= \frac{24}{5} \\
&= 4.8
\end{aligned}
$$

Variance

- Average of the <u>squared</u> deviations from the arithmetic mean

X	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

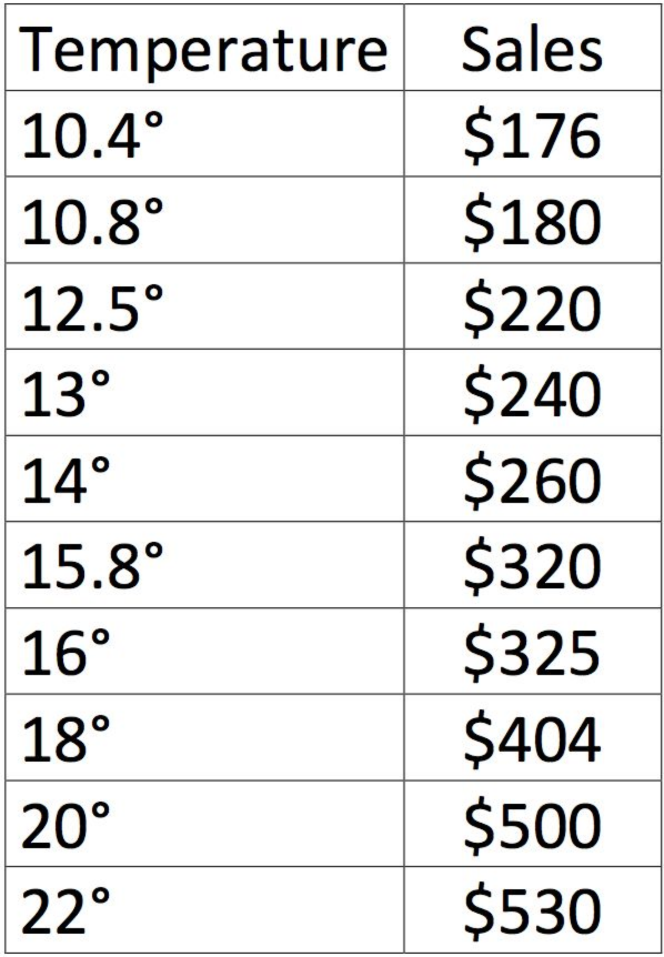
Inference about population using Mean Sample & Variance

Concept about ‘Variables’

Till now we only looked at one variable at a time.

Like salary, let’s say age, and so on.

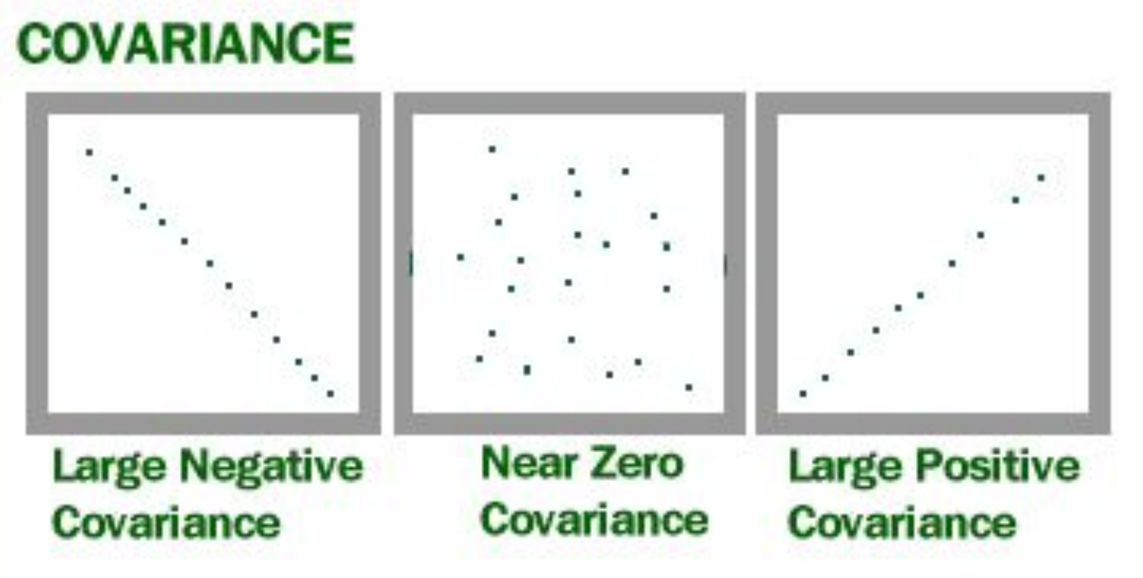
What if we want to compare two (and more) variables at a time?

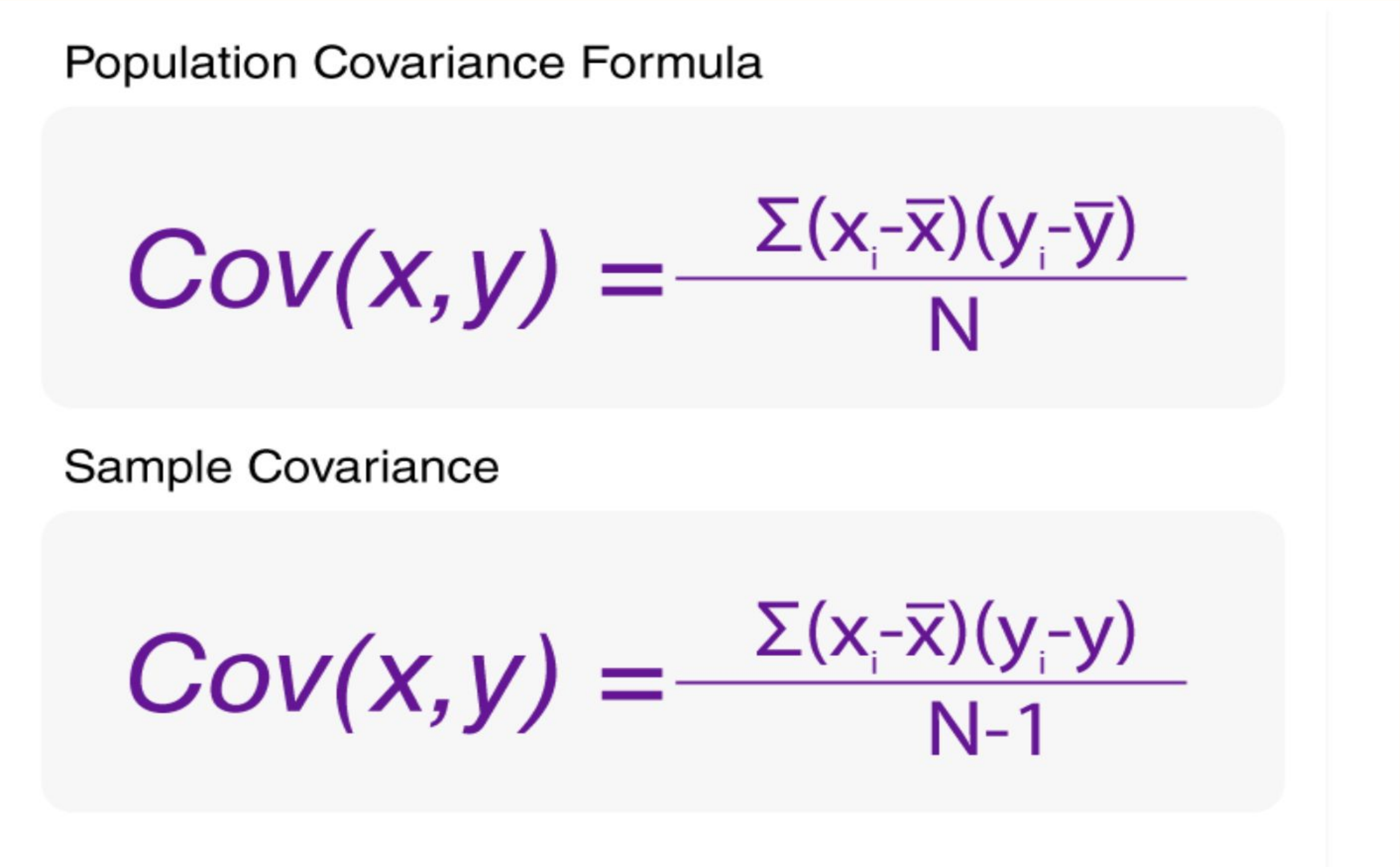


Temperature	Sales
10.4°	$176
10.8°	$180
12.5°	$220
13°	$240
14°	$260
15.8°	$320
16°	$325
18°	$404
20°	$500
22°	$530

Covariance

- How two variables are related to each other.



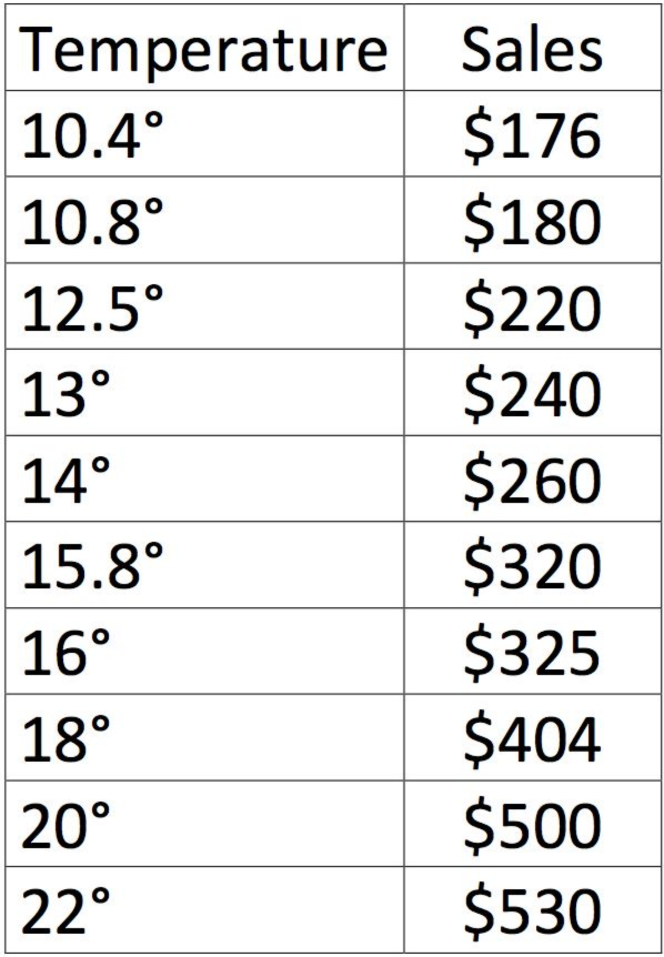


Population Covariance Formula

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$



Temperature	Sales
10.4°	$176
10.8°	$180
12.5°	$220
13°	$240
14°	$260
15.8°	$320
16°	$325
18°	$404
20°	$500
22°	$530

Does a greater number mean stronger relation between the two variables?

What if the sales were reported in Rupees?

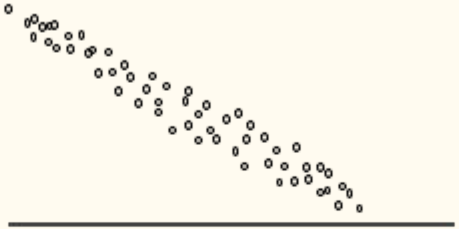
$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

$$\mathrm{r_{xy}} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2 \sum(\mathrm{y_i} - \bar{\mathrm{y}})^2}}$$

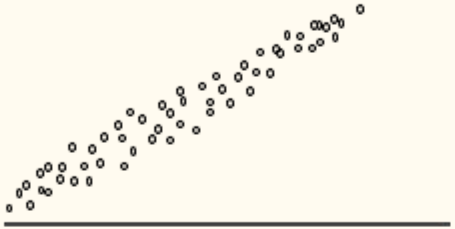
Correlation

- How two things are related to each other.
- Could be positively related, negatively related, or not related at all.
- Uses a score from -1 to 1.
- Both direction and the strength of the relationship is captured in score.

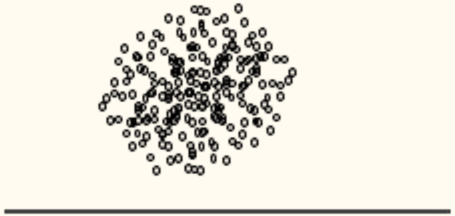
Three Degrees of Correlation



$r < 0$

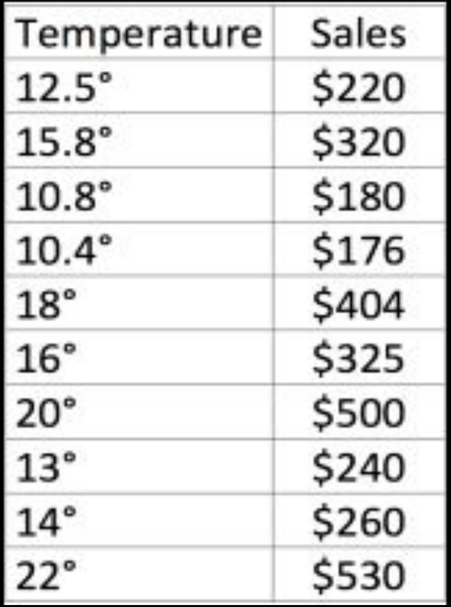


$r > 0$

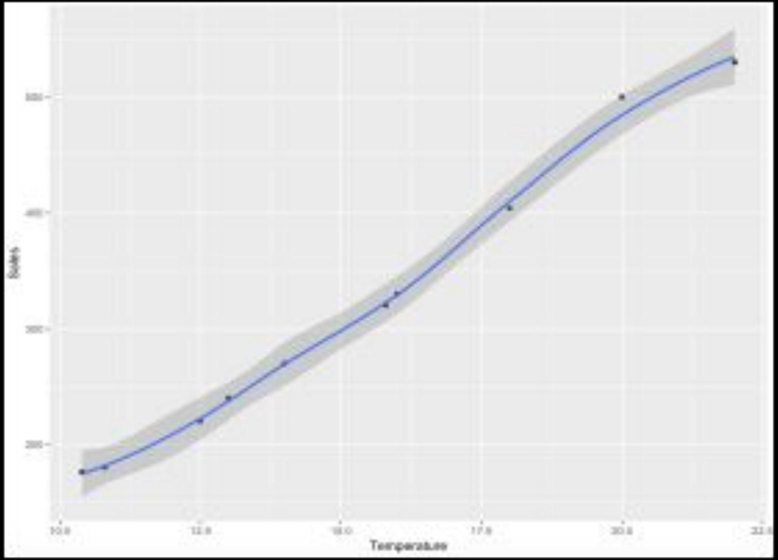


$r = 0$

Is increase in Temperature causing the increase in Ice Cream sales?

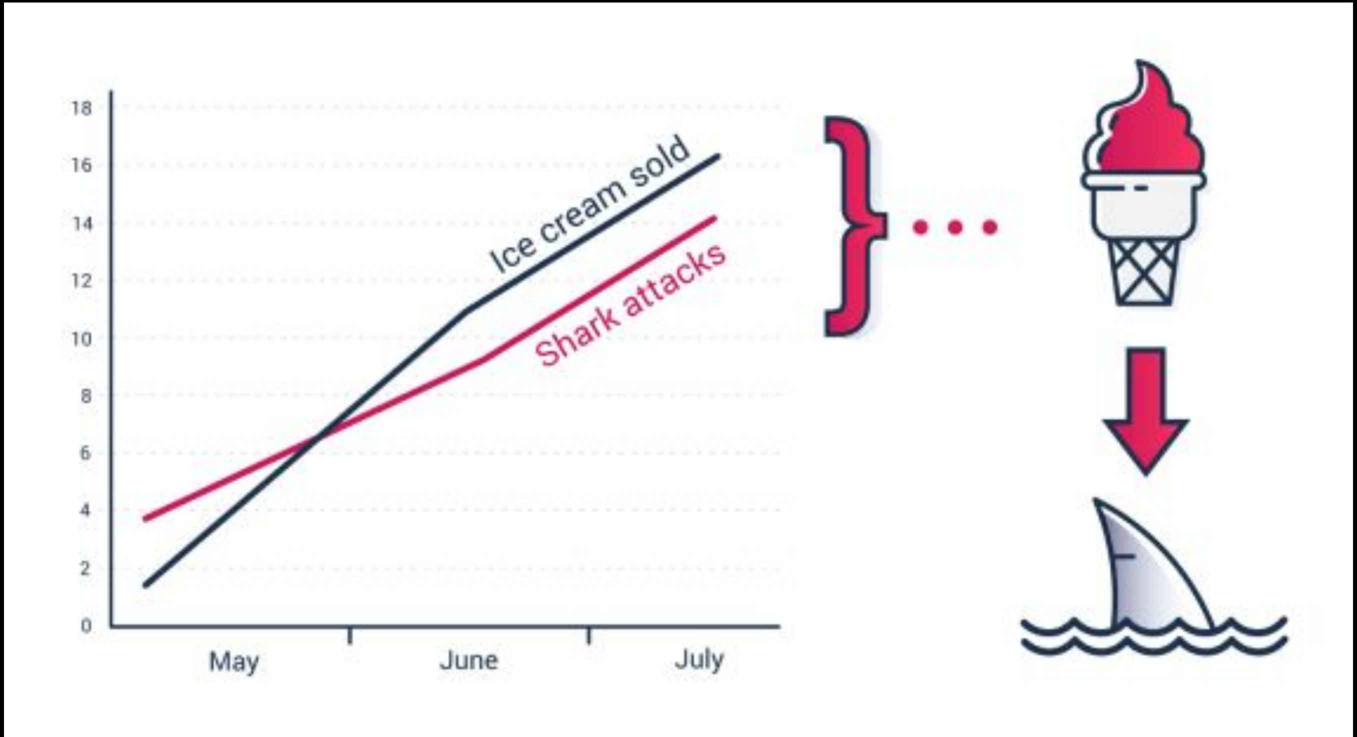


Temperature	Sales
12.5°	$220
15.8°	$320
10.8°	$180
10.4°	$176
18°	$404
16°	$325
20°	$500
13°	$240
14°	$260
22°	$530



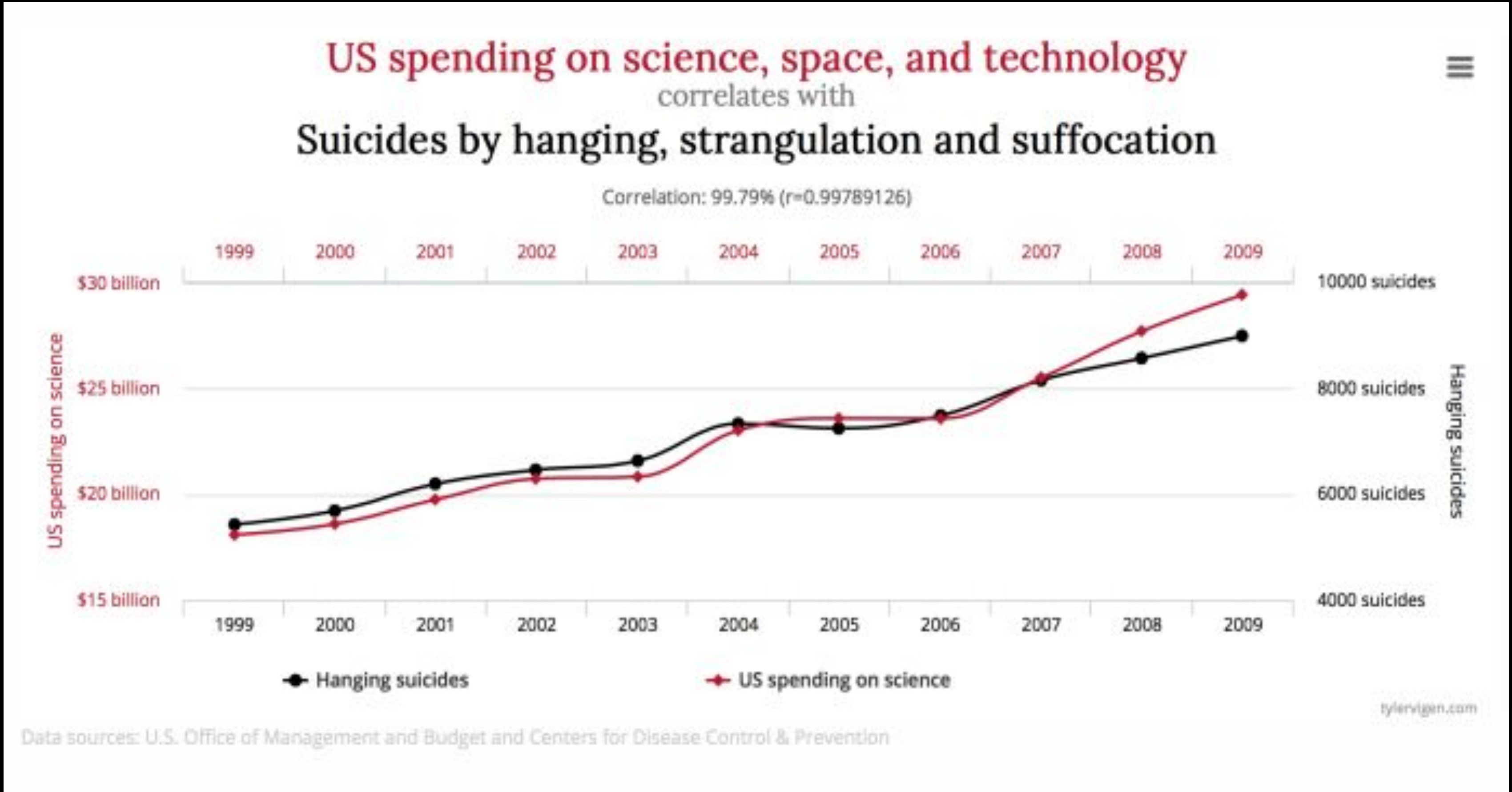
Correlation vs Causation

Correlation vs Causation



Increase in Temperature -> Increase in Ice Cream Sales

Increase in Temperature -> Increase in number of people going to the beach -> More shark attacks



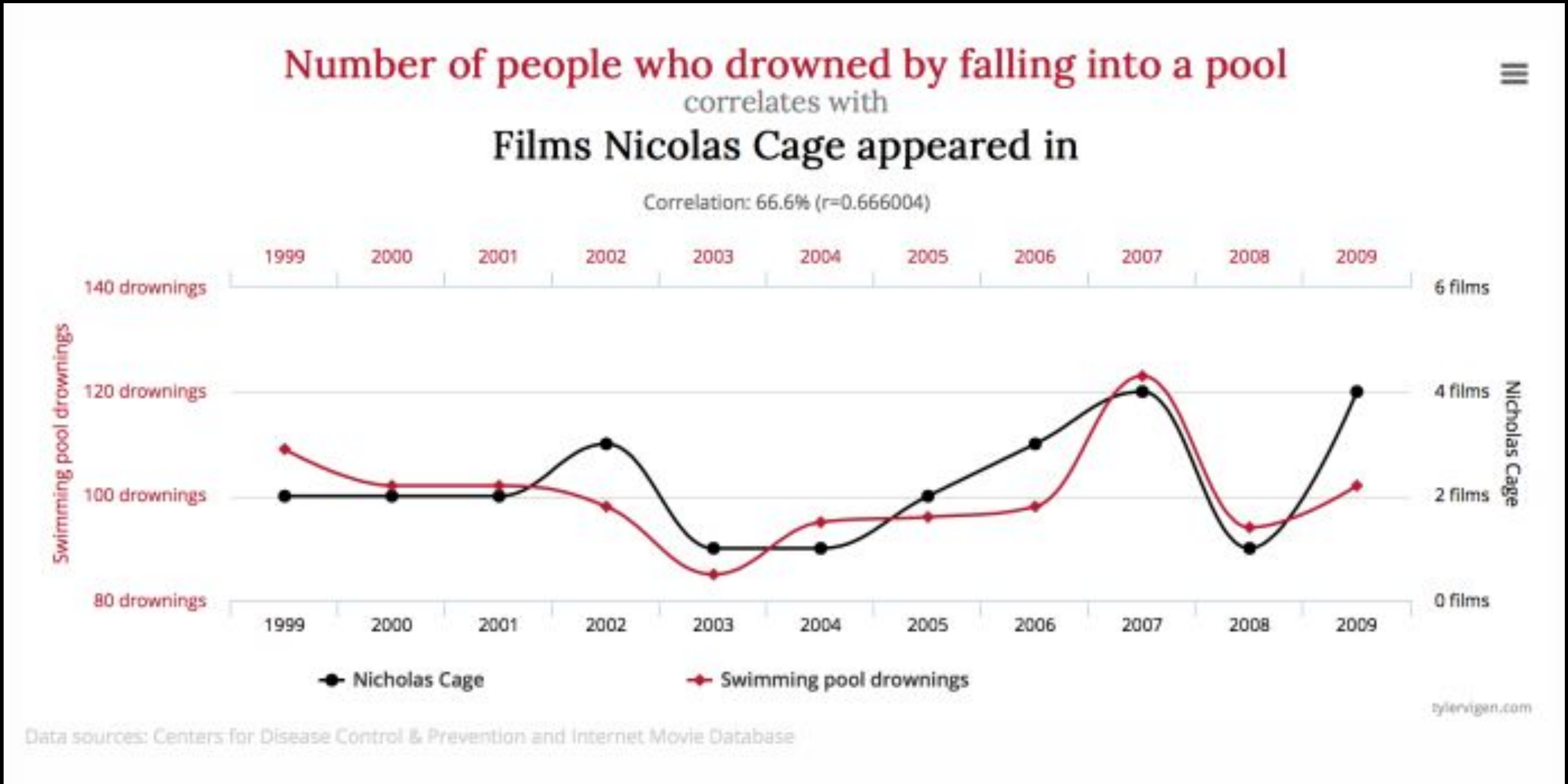
US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention



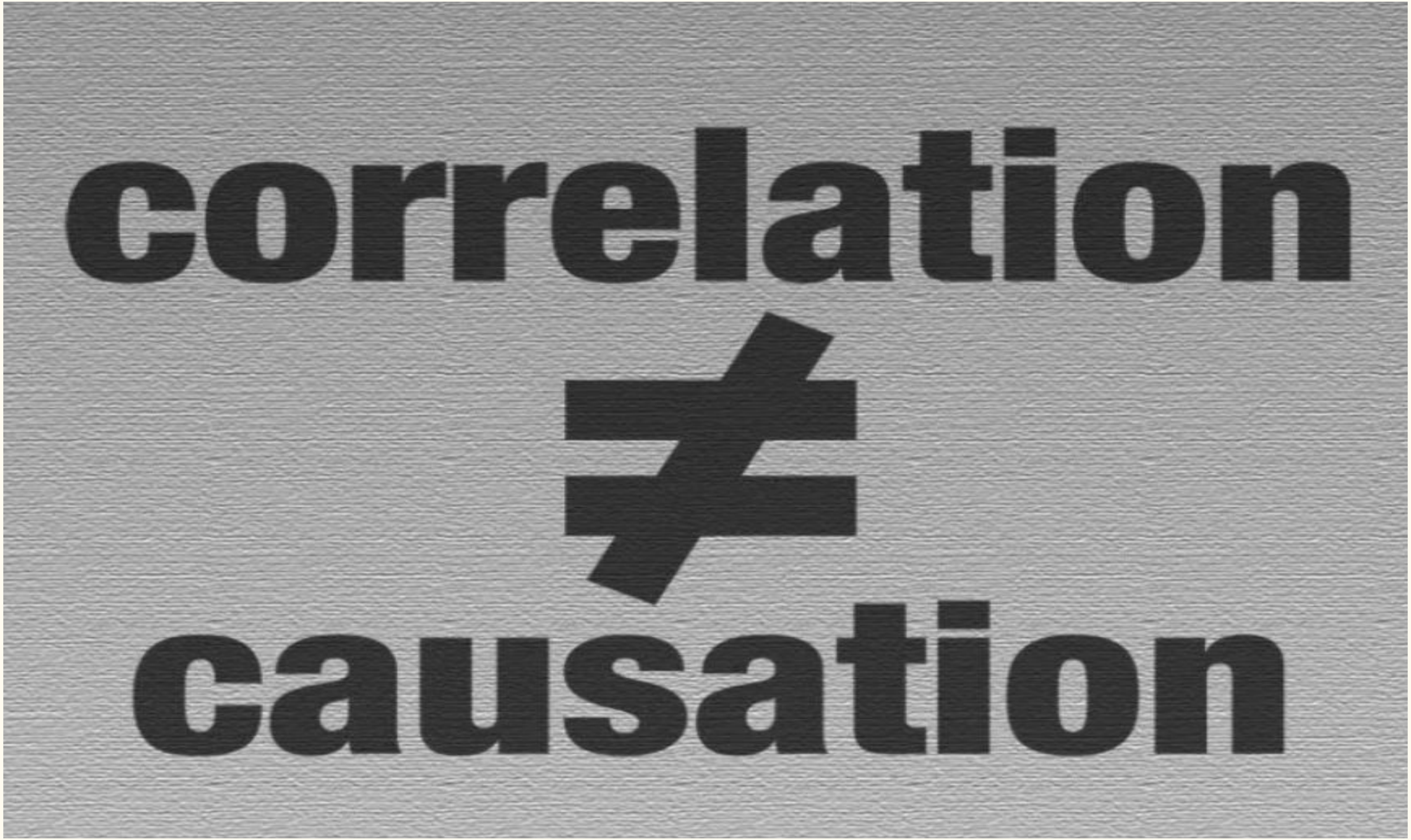
Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database



correlation

≠

causation

How to do Causality tests?

Before that, we will learn something more basic and fundamental!

Probability & Probability Distributions

End Note!

Thank You!