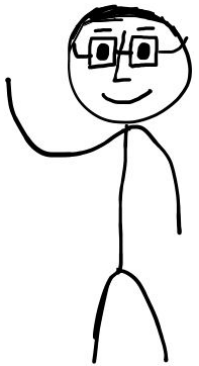
Data Analysis

Mohit Deshpande

Data Analysis

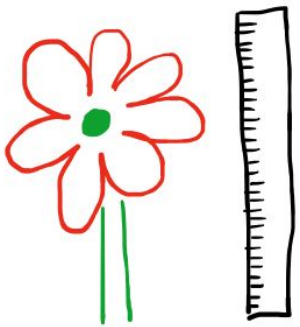
1. Introduction to Statistics
2. Mean and Standard Deviation
3. Linear Regression
4. Correlation Coefficient

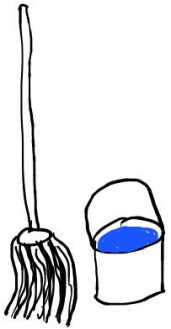


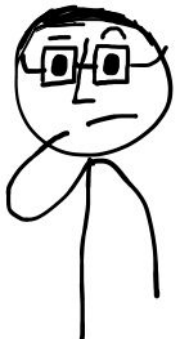
Introduction to Statistics

Statistics

branch of mathematics that deals with *collection*, *cleaning*, and *analysis* of data





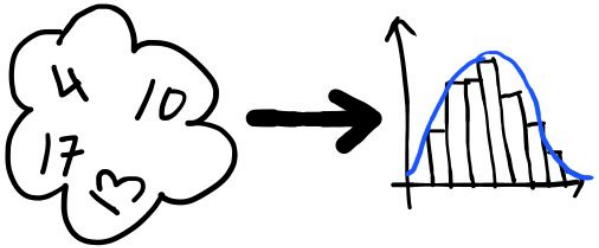


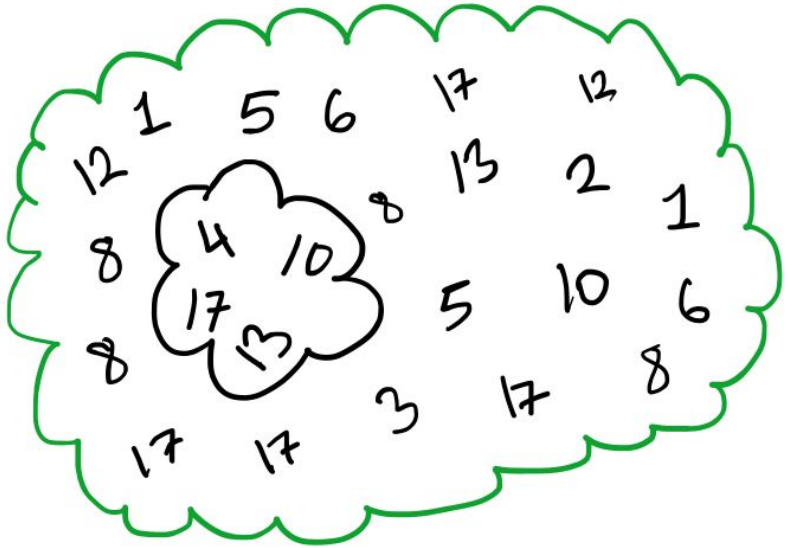
Introduction to Statistics

Statistics

Descriptive Statistics

Inferential Statistics



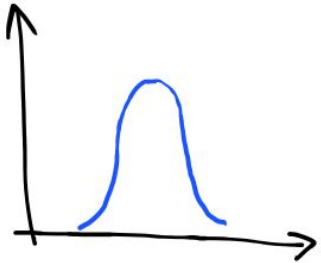


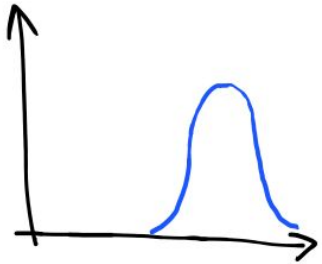
Descriptive Statistics

Central Tendency

Data distribution’s central value

Also called mean, average, expected value

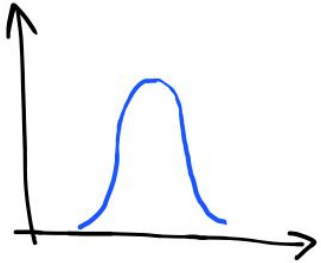


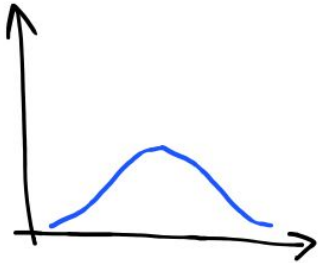


Dispersion

Data distribution’s spread

Also called variance



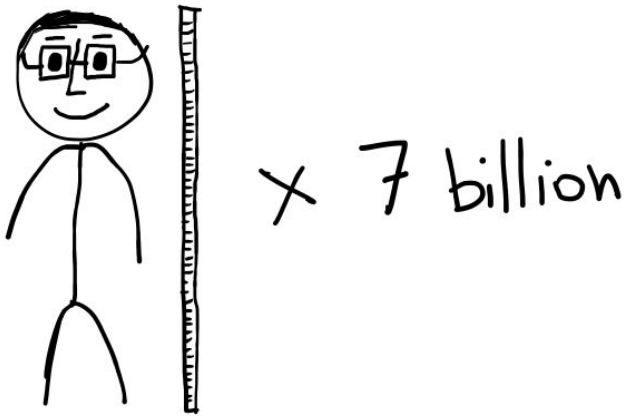


Inferential Statistics

Population: group of all of the items or events we're studying.

Usually impractical or impossible to get data.

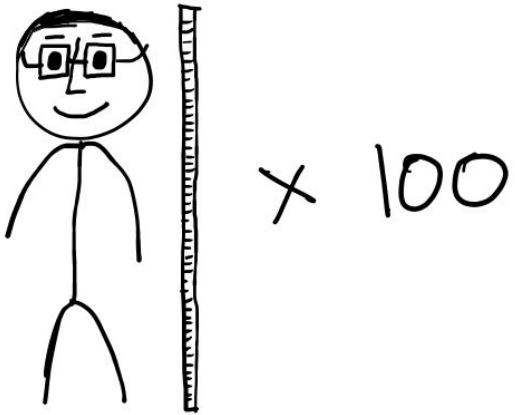
Example: height of humans



Sample: subset of the population selected via a defined method.

Much easier to obtain!

Example: measure 100 humans



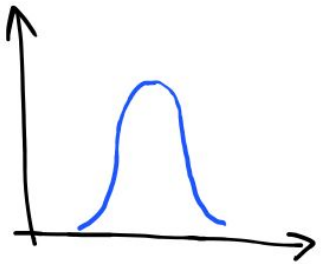
Data Analysis

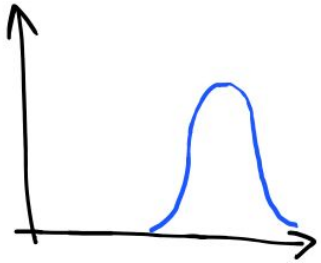
1. Introduction to Statistics 
2. Mean and Standard Deviation
3. Linear Regression
4. Correlation Coefficient

Mean and Standard Deviation

Mean

- Also called average or expected value
- If our dataset is x_1, x_2, …, x_n, then the mean is denoted by
  - μ_x for population mean
  - $\bar{x}$ for sample mean





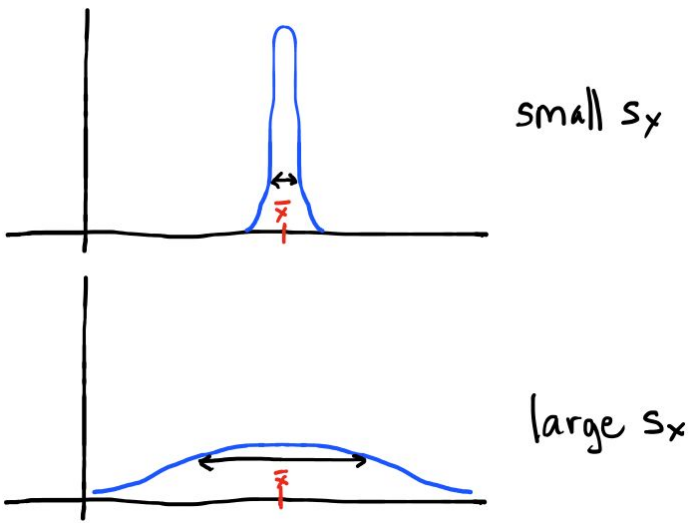
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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

Standard Deviation

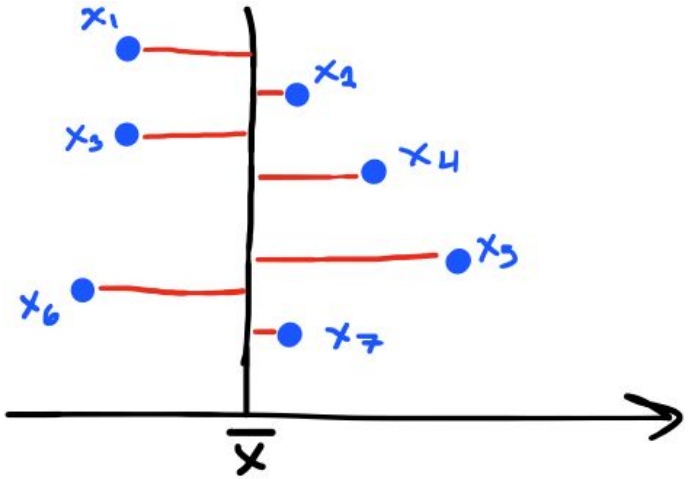
- Measures the spread or dispersion of the data
- If our dataset is x_1, x_2, …, x_n, then the standard deviation is denoted by
  - σ_x for population standard deviation
  - s_x for sample standard deviation
- Variance denoted
  - σ_x^2 for population variance
  - s_x^2 for sample variance

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



Variance

- **Variance**: the “average” of the squared differences of the data from the mean



$$s_x^2 = \frac{(\quad)^2 + (\quad)^2 + \cdots + (\quad)^2}{6}$$

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

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

List of Symbols

	Population	**Sample**
Mean	μ_x	$\overline{x}$
Standard Deviation	σ_x	s_x
Variance	σ_x^2	s_x^2

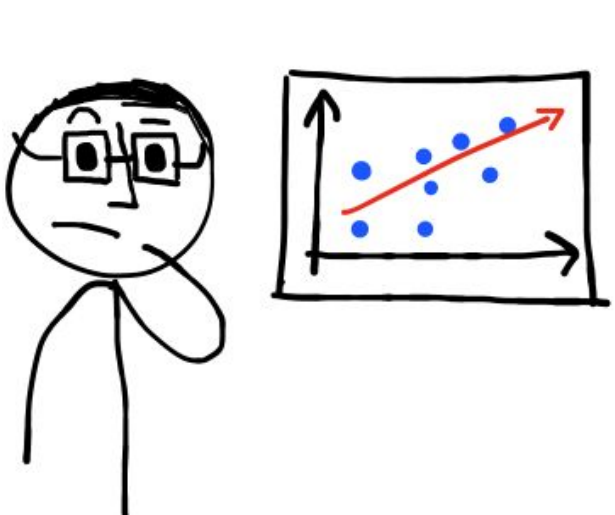
Data Analysis

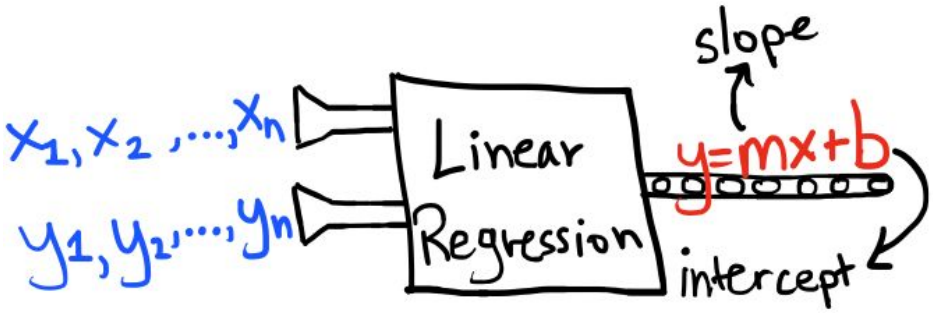
1. Introduction to Statistics 
2. Mean and Standard Deviation 
3. Linear Regression
4. Correlation Coefficient

Linear Regression

Linear Regression

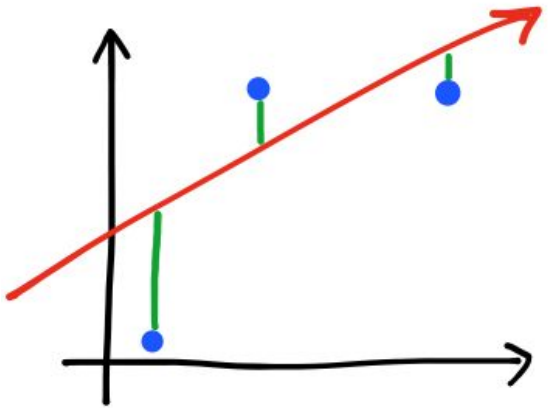
- Given our data, what is the slope and y-intercept of the line that “best” represents the trend of the data
- Use this **line-of-best-fit** to make predictions

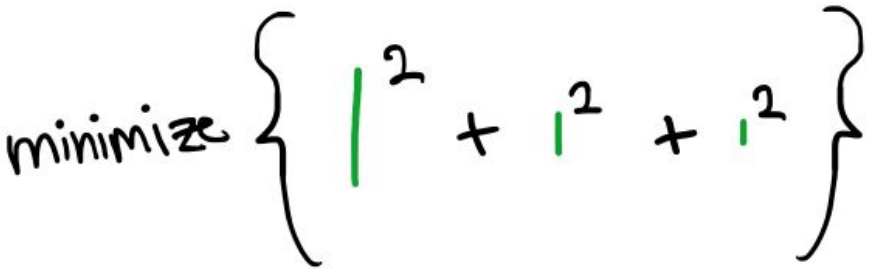




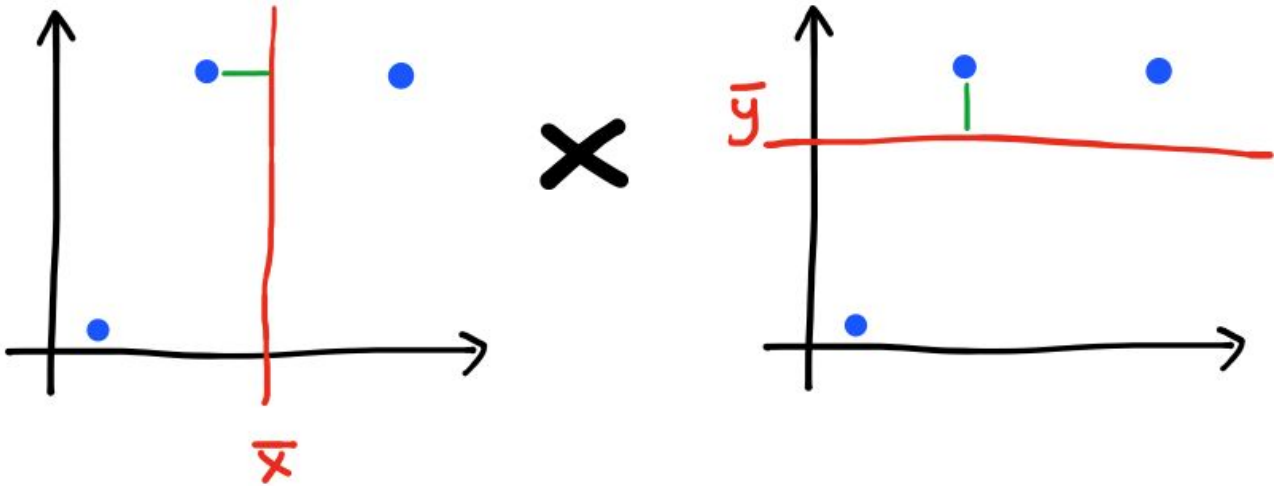
Least Squares Linear Regression

- Pick a line such that the sum of the squared vertical differences is minimal
- Falls under field of mathematical optimization problems
- Closed-form exact solution exists

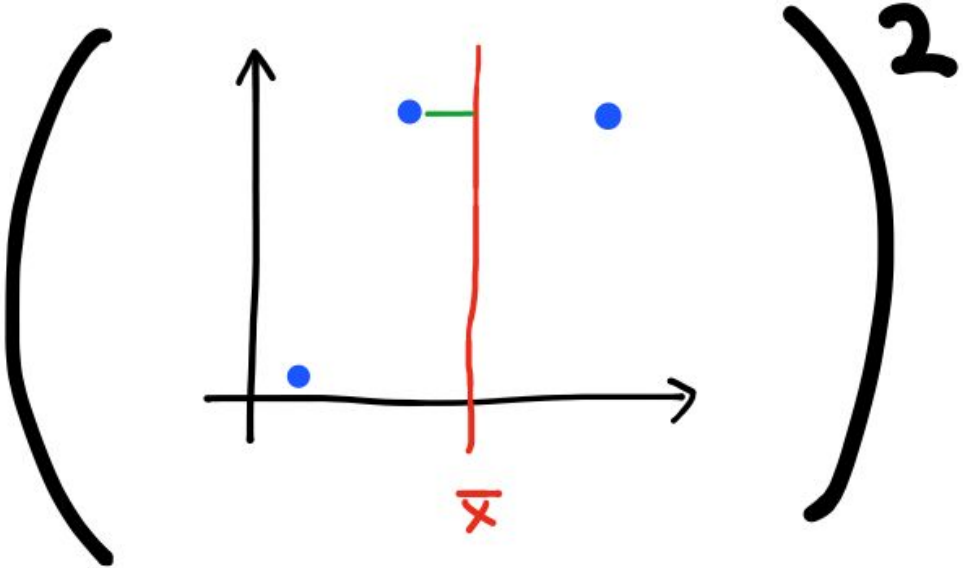




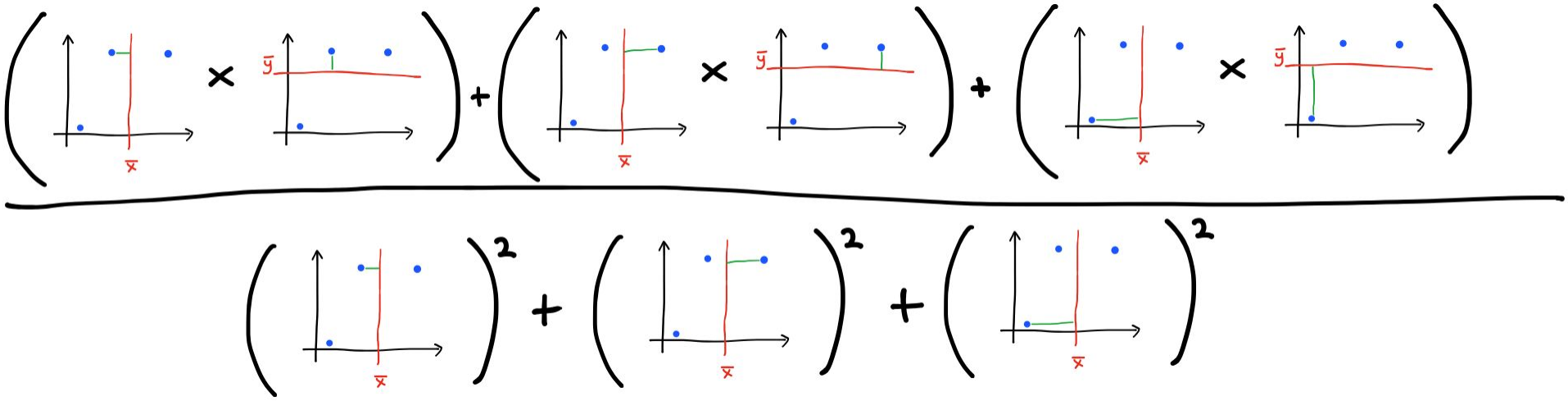
Least Squares Linear Regression Solution



Least Squares Linear Regression Solution



Least Squares Linear Regression Solution



Least Squares Linear Regression Solution

$$m = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}$$

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

$$b = \bar{y} - m\bar{x}$$

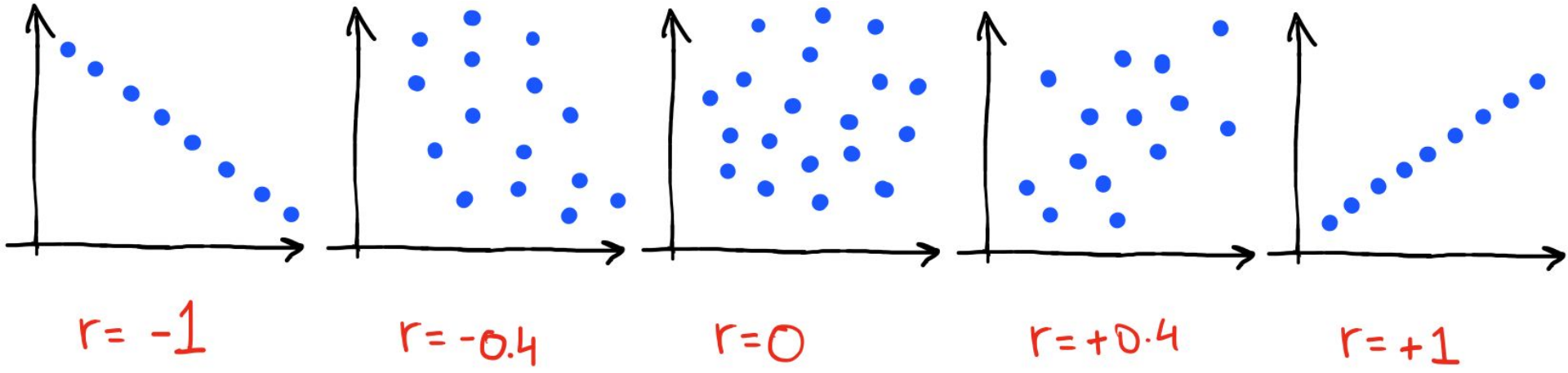
Data Analysis

1. Introduction to Statistics 
2. Mean and Standard Deviation 
3. Linear Regression 
4. Correlation Coefficient

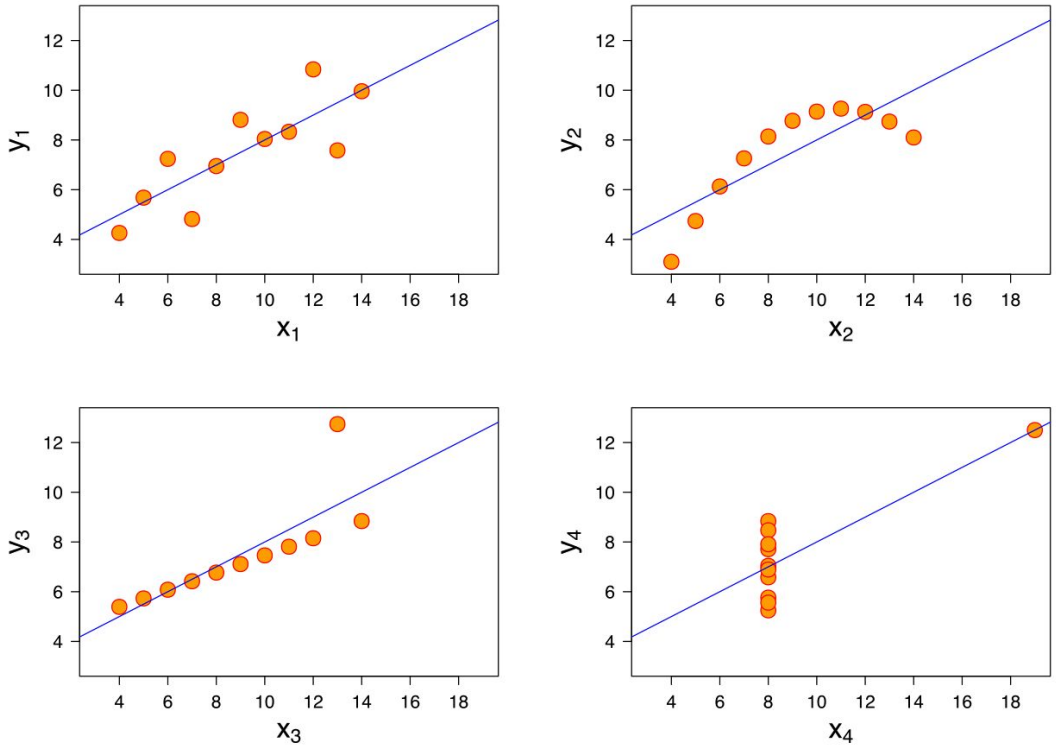
Correlation Coefficient

Correlation Coefficient

- Measure of strength and linearity of the line-of-best-fit
- Pearson Correlation Coefficient (PCC) denoted by r
  - r > 0 means X correlates positively with Y; r < 0 means X correlates negatively with Y
  - |r| ≈ 0 means there is a weak correlation; |r| ≈ 1 means there is a strong correlation

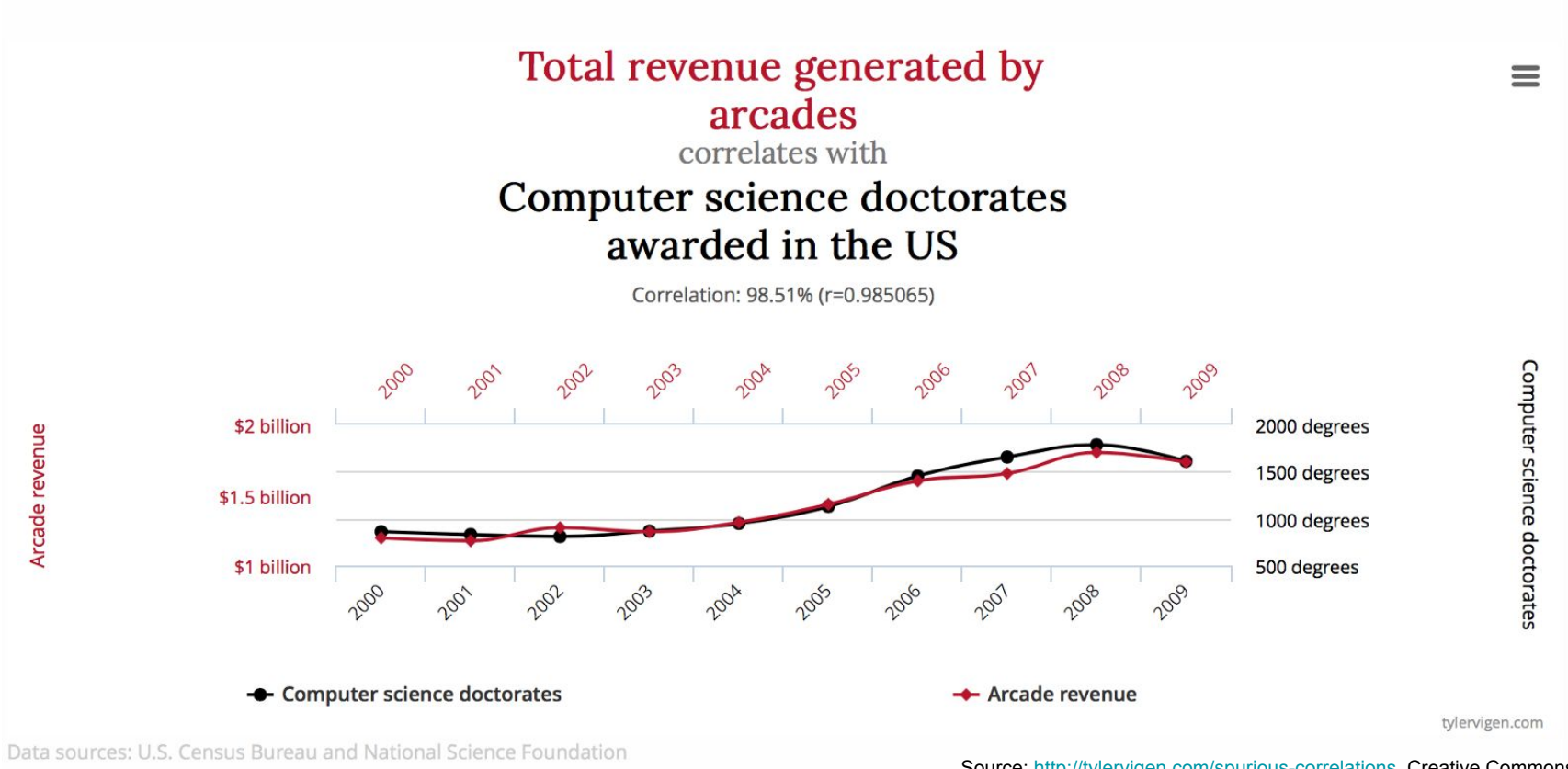


ALWAYS PLOT YOUR DATA!



Source: https://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg

CORRELATION DOES NOT IMPLY CAUSATION!



Source: http://tylervigen.com/spurious-correlations. Creative Commons License

Data Analysis

1. Introduction to Statistics 
2. Mean and Standard Deviation 
3. Linear Regression 
4. Correlation Coefficient