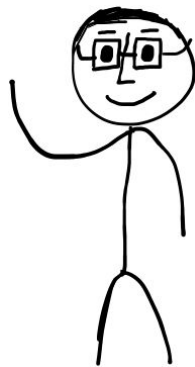


Data Analysis

Mohit Deshpande

Data Analysis

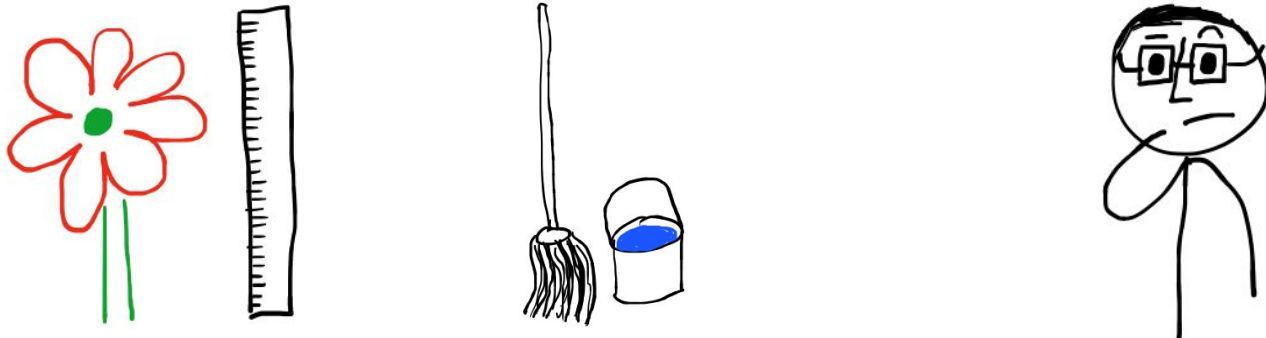
1. Introduction to Statistics
2. Mean and Standard Deviation
3. Linear Regression
4. Correlation Coefficient



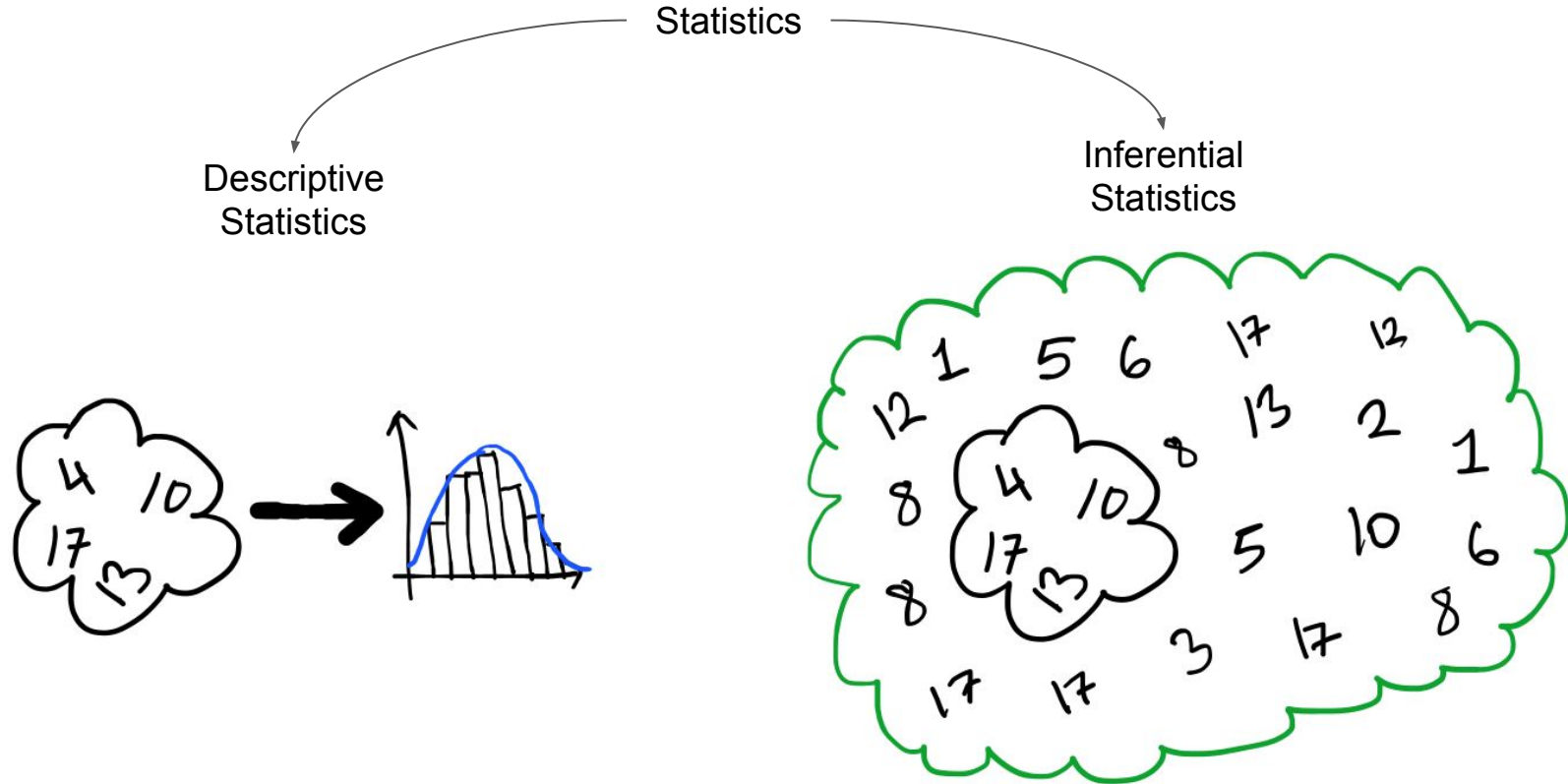
Introduction to Statistics

Statistics

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



Introduction to Statistics

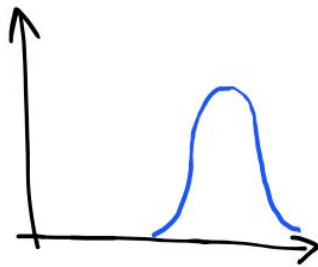
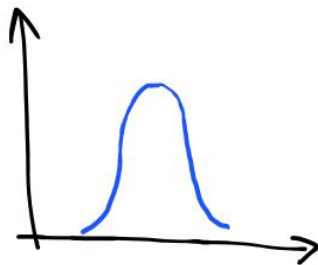


Descriptive Statistics

Central Tendency

Data distribution's central value

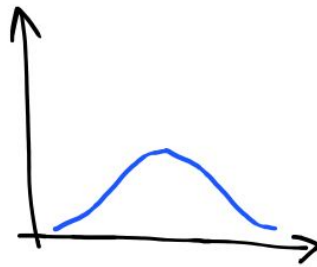
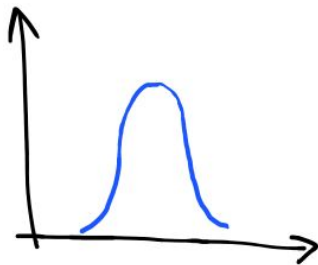
Also called mean, average, expected value



Dispersion

Data distribution's spread

Also called variance

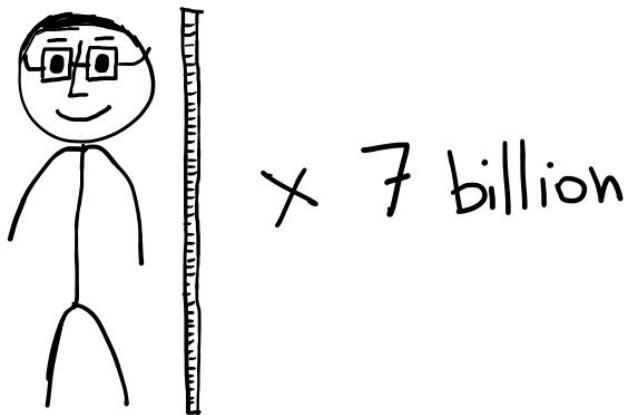


Inferential Statistics

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

Usually impractical or impossible to get data.

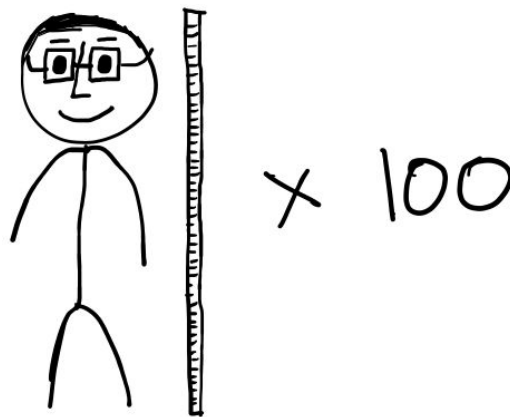
Example: height of humans




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

Much easier to obtain!

Example: measure 100 humans



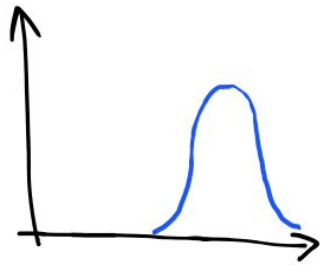
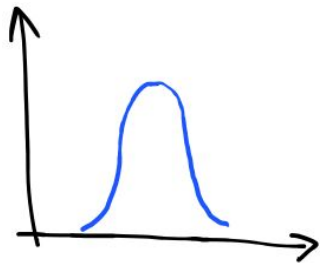
Data Analysis

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, \dots, x_n , then the mean is denoted by
 - μ_x for population mean
 - \bar{x} for sample mean



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

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

Standard Deviation

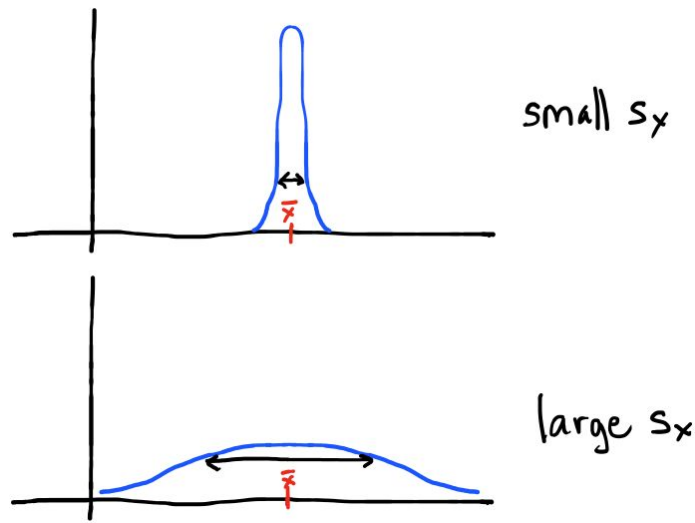
- Measures the spread or dispersion of the data
- If our dataset is x_1, x_2, \dots, x_n , then the standard deviation is denoted by

- σ_x for population standard deviation
- s_x for sample standard deviation

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

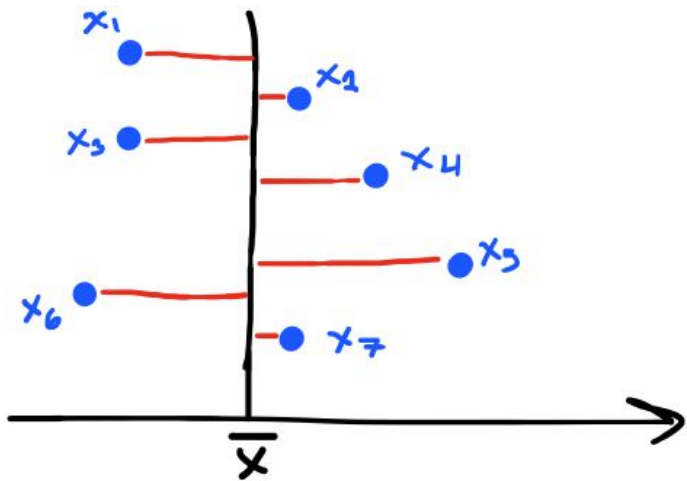
- Variance denoted

- σ_x^2 for population variance
- s_x^2 for sample variance



Variance

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



$$S_x^2 = \frac{(-)^2 + (-)^2 + \dots + (-)^2}{6}$$



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

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

List of Symbols

	Population	Sample
Mean	μ_x	\bar{x}
Standard Deviation	σ_x	s_x
Variance	σ_x^2	s_x^2

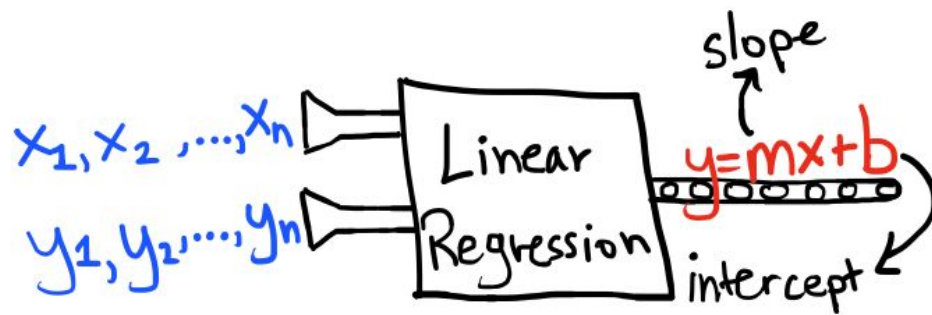
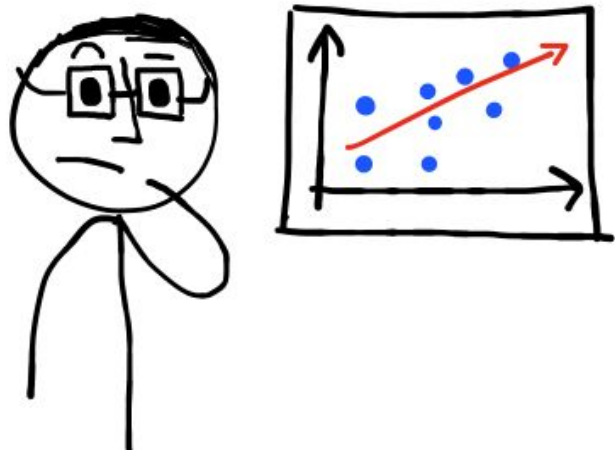
Data Analysis

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

Linear Regression

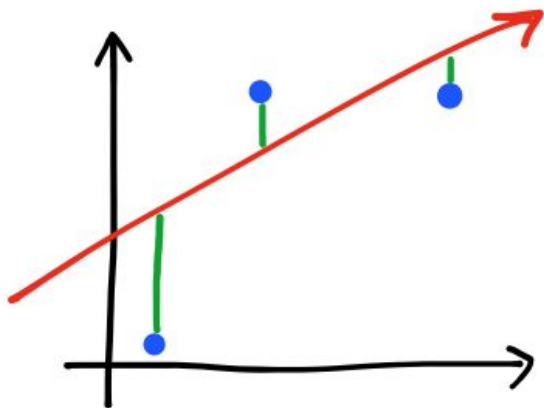
Linear Regression

- 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



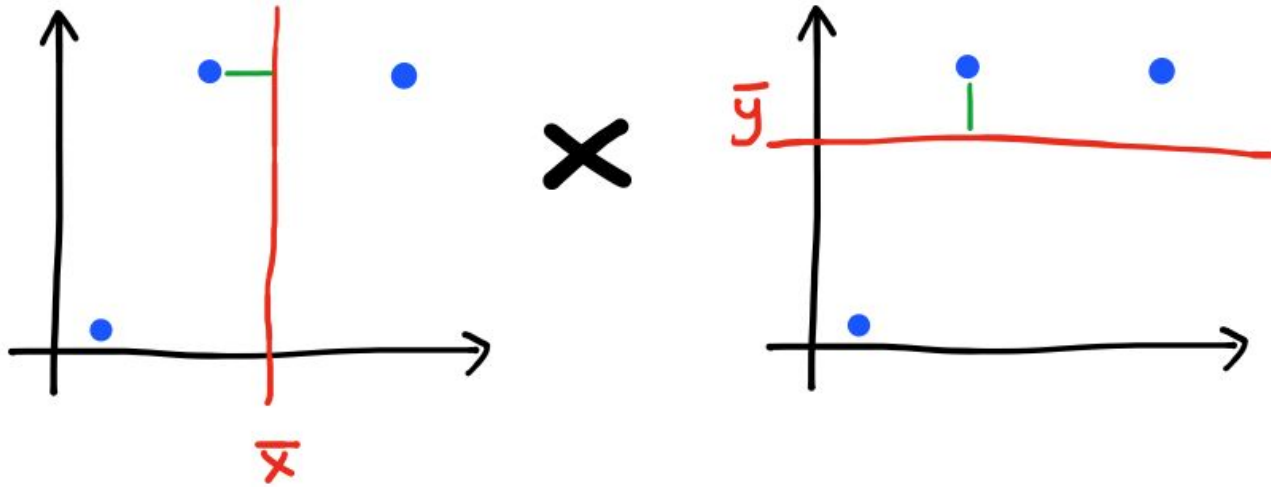
Least Squares Linear Regression

- 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

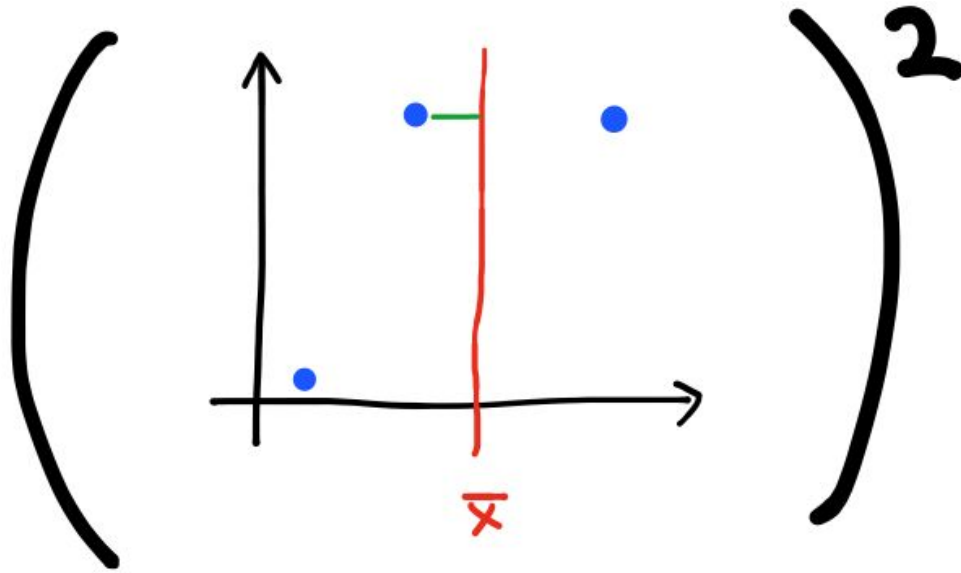


$$\text{minimize } \left\{ | \quad |^2 + | \quad |^2 + | \quad |^2 \right\}$$

Least Squares Linear Regression Solution



Least Squares Linear Regression Solution



Least Squares Linear Regression Solution

$$\begin{aligned}
 & \left(\begin{array}{c} \text{Graph 1} \times \text{Graph 2} \end{array} \right) + \left(\begin{array}{c} \text{Graph 1} \times \text{Graph 2} \end{array} \right) + \left(\begin{array}{c} \text{Graph 1} \times \text{Graph 2} \end{array} \right) \\
 & \hline
 & \left(\begin{array}{c} \text{Graph 1} \end{array} \right)^2 + \left(\begin{array}{c} \text{Graph 1} \end{array} \right)^2 + \left(\begin{array}{c} \text{Graph 1} \end{array} \right)^2
 \end{aligned}$$

The diagram illustrates the least squares linear regression solution through three visual components:




- Top Row:** Three pairs of graphs, each pair enclosed in large parentheses and separated by a plus sign. Each pair consists of a scatter plot (left) and a regression line (right). The scatter plots show three data points (blue dots) and a vertical red line at \bar{x} . The regression lines are red and show the predicted values for each data point. The vertical distance between each data point and its corresponding regression line is marked with a green vertical line segment.
- Bottom Row:** Three individual scatter plots, each enclosed in large parentheses and separated by a plus sign. Each scatter plot shows the three data points and the vertical red line at \bar{x} . The vertical distance between each data point and the regression line is marked with a green vertical line segment.

Least Squares Linear Regression Solution

$$\begin{aligned} m &= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

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

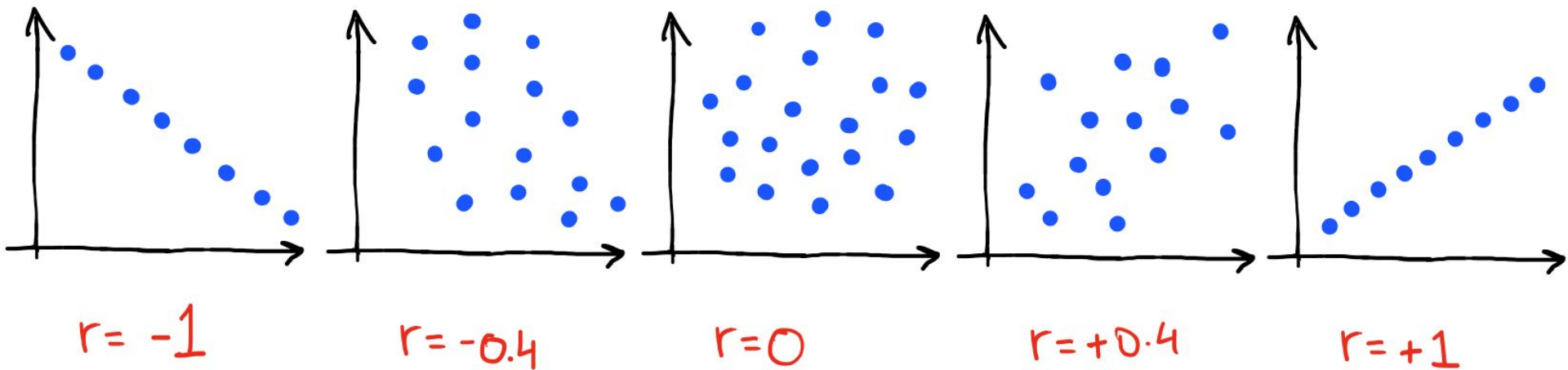
Data Analysis

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

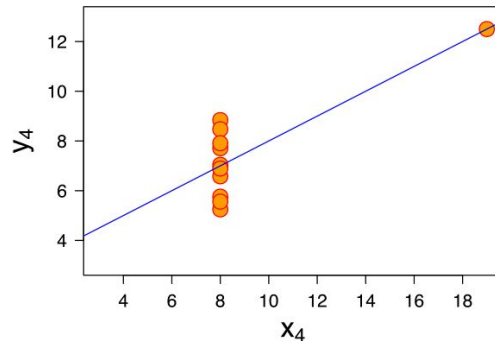
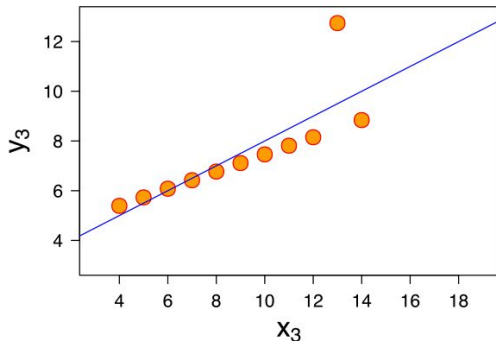
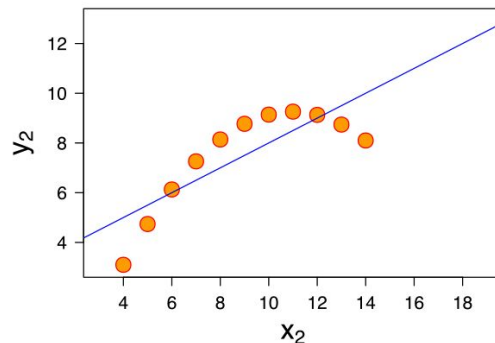
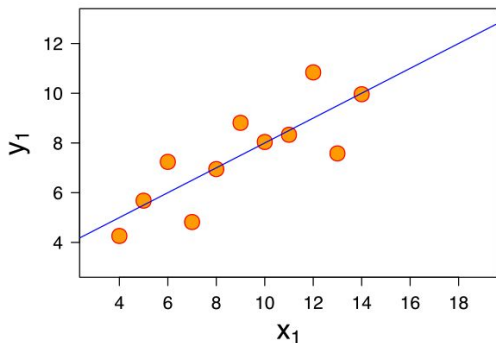
Correlation Coefficient

Correlation Coefficient

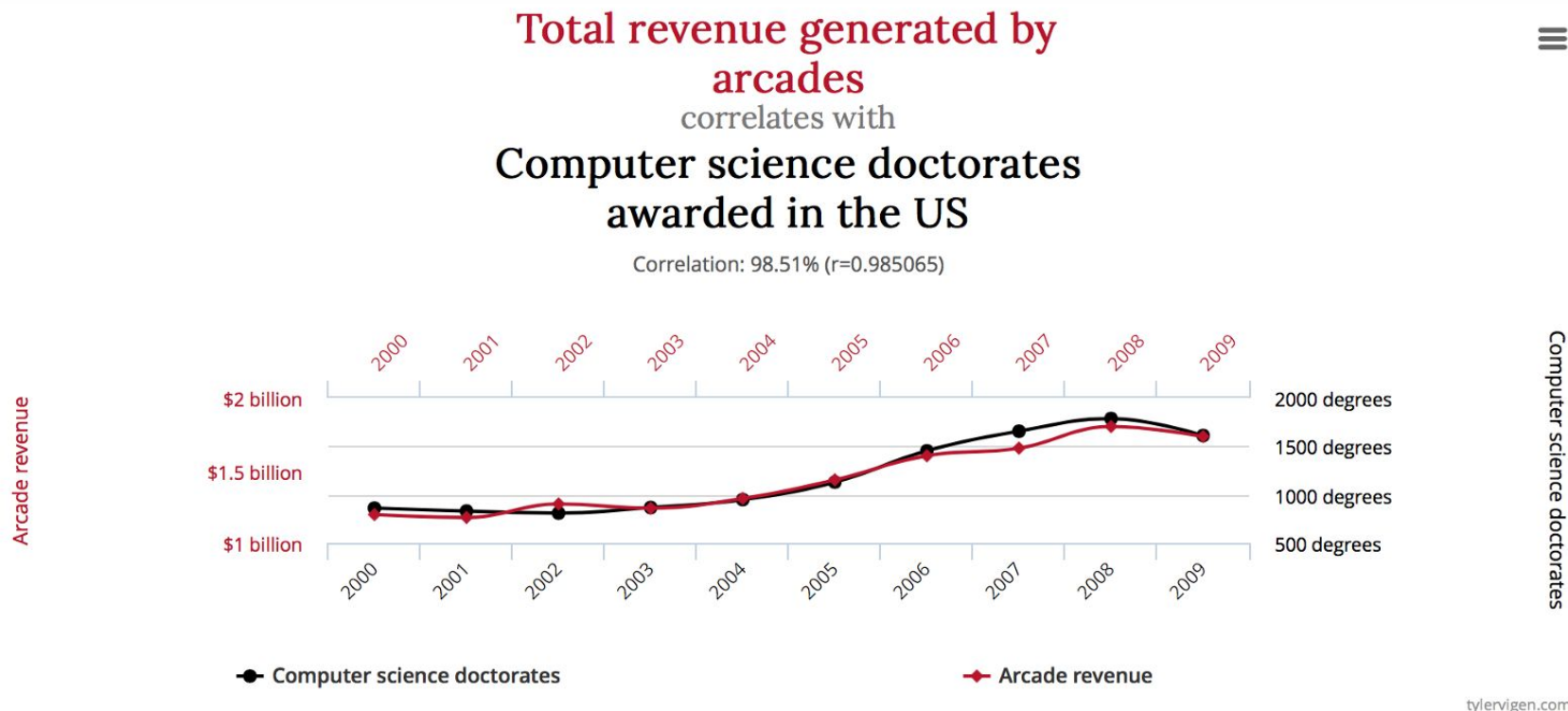
- Measure of strength and linearity of the line-of-best-fit
- Pearson Correlation Coefficient (PCC) denoted by r
 - $r > 0$ means X correlates positively with Y; $r < 0$ means X correlates negatively with Y
 - $|r| \approx 0$ means there is a weak correlation; $|r| \approx 1$ means there is a strong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



ALWAYS PLOT YOUR DATA!



CORRELATION DOES NOT IMPLY CAUSATION!



Data sources: U.S. Census Bureau and National Science Foundation

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 