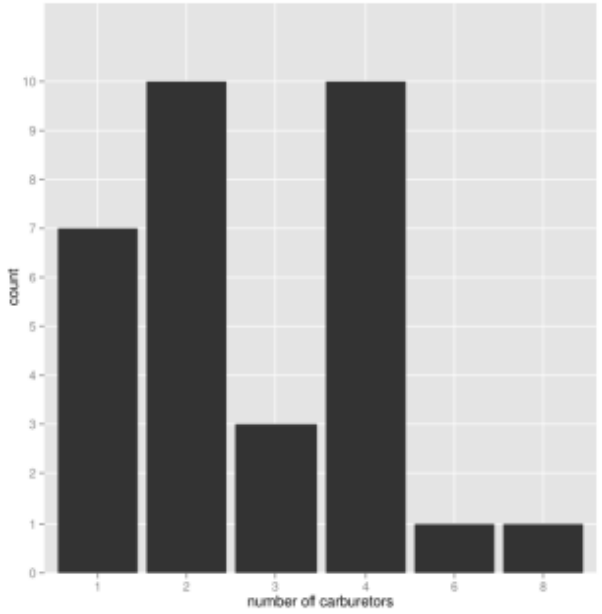
Basic statistics, linear regression and correlation

Agenda

- Describe Univariate data
  - Central tendency
  - Spread
  - Distribution
- Describe multivariate data
  - Linear regression
  - Correlation

Univariate data

- Categorical data
  - Nominal
  - Ordinal (categorical variables that can be sorted or ordered. E.g. t-shirt size)
- Continuous data
  - Continuous variables can be discretized to become categorical data
- Frequency distributions



Central tendency

- Categorical data
  - Mode
- Continuous data
  - Mean
  - Median

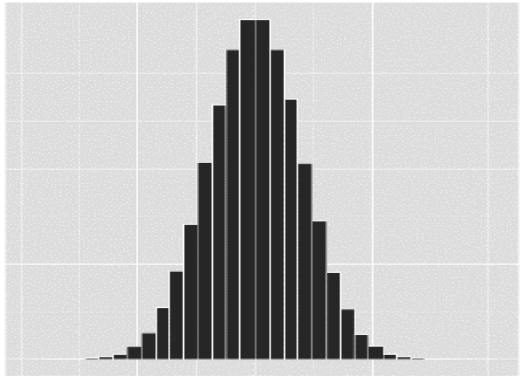


Figure 2.3: A normal distribution

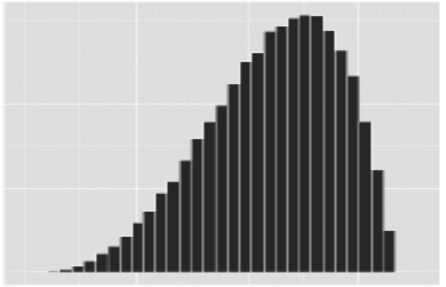


Figure 2.4a: A negatively skewed distribution

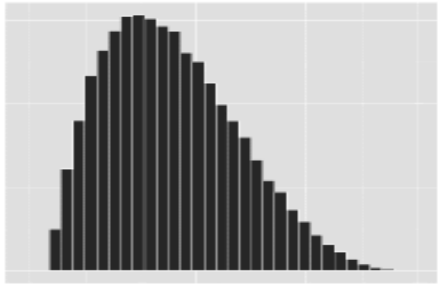


Figure 2.4b: A positively skewed distribution

Degree of skewness

Spread

- Variance
- Standard deviation

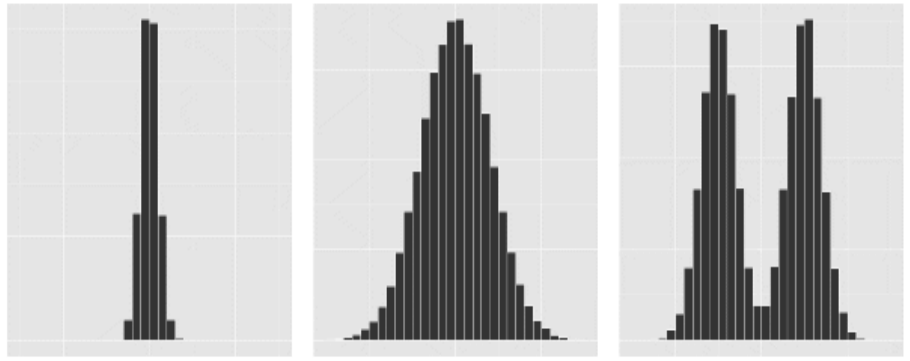


Figure 2.5: three distributions with the same mean and median

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

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

$$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1} \equiv s^2,$$

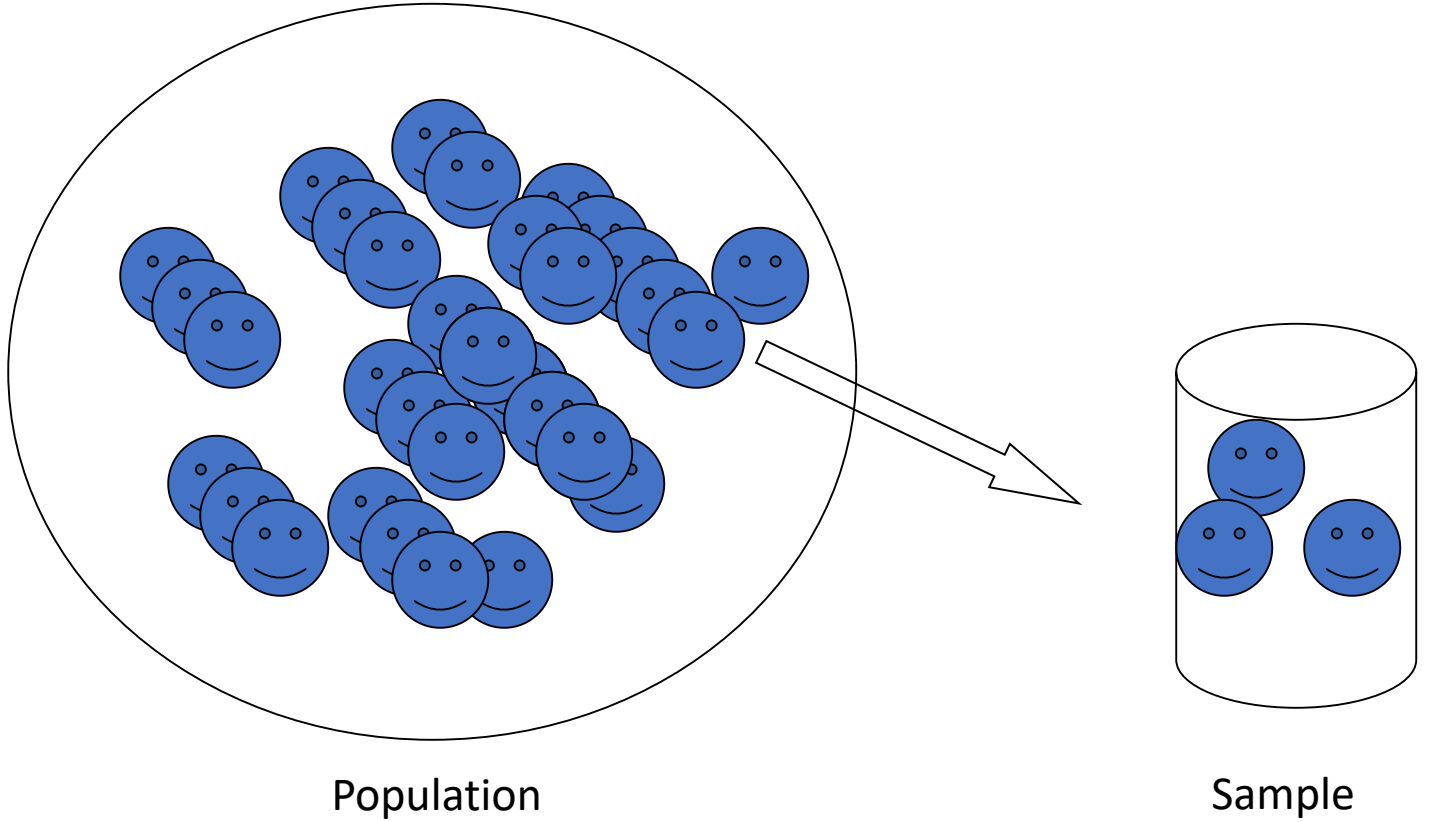
Degrees of freedom

$$\sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n-1}} \equiv s$$

SD of a sample

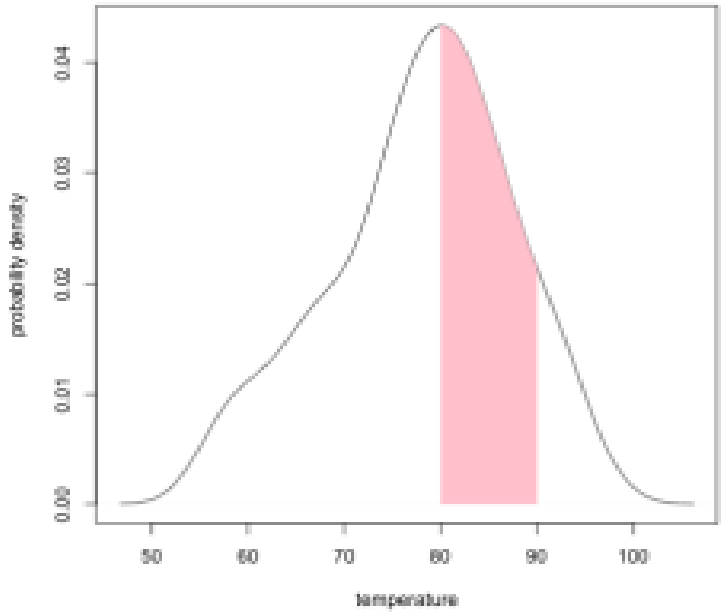
Populations,samples and estimation

- One of the core ideas of statistics is that we can use a subset of a group, study it and then make inferences or conclusions about that much larger group.



Probability density function (PDF)

- In probability theory, a **probability density function** (**PDF**), or **density** of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value. The probability of the random variable falling within a particular range of values is given by the integral of this variable's density over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one.



Descriptive Statistics

Summarizing Data:

- ✓ Central Tendency (or Groups' "Middle Values")
  - ✓ Mean
  - ✓ Median
  - ✓ Mode
- ✓ Variation (or Summary of Differences Within Groups)
  - ✓ Range
  - ✓ Interquartile Range
  - ✓ Variance
  - ✓ Standard Deviation
- …Wait! There's more

Box-Plots

A way to graphically portray almost all the descriptive statistics at once is the box-plot.

A box-plot shows: Upper and lower quartiles

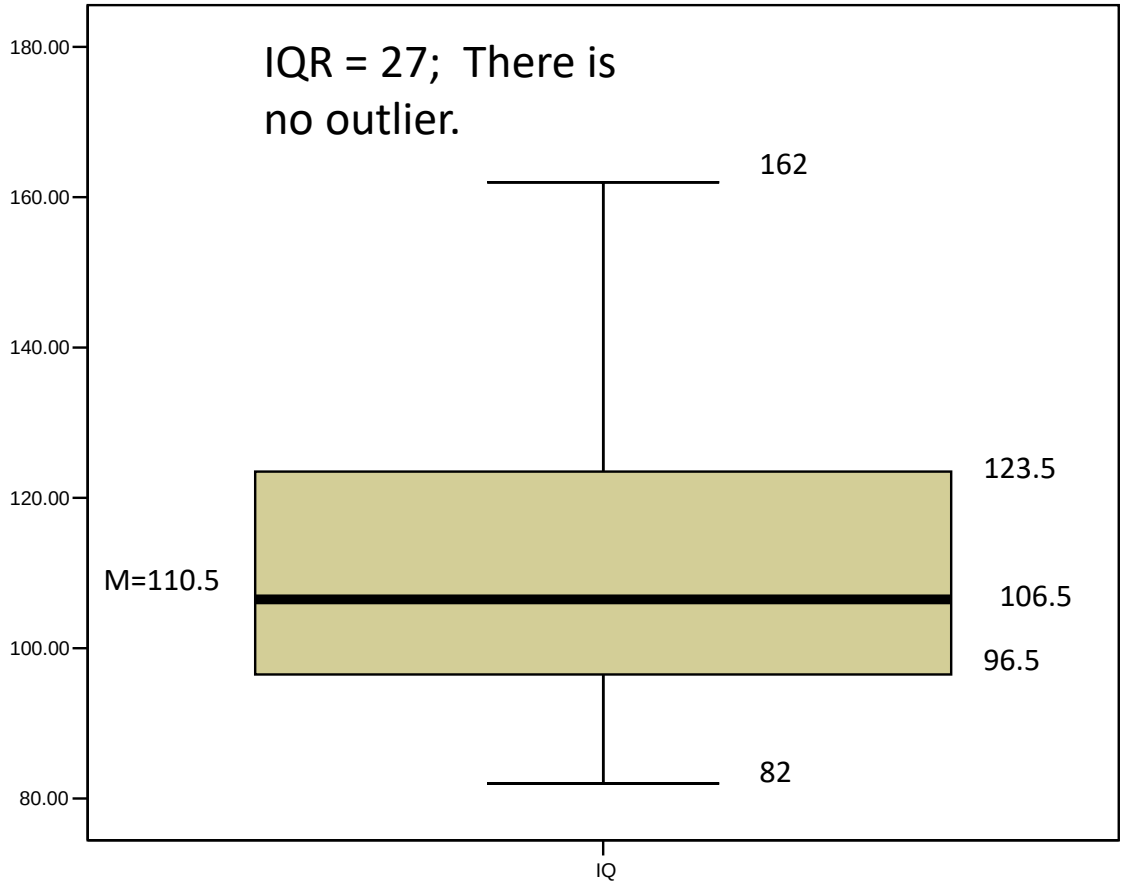
Mean

Median

Range

Outliers (1.5 IQR)

Example: Box-Plots



Multivariate data

- To determine the relationship among different variables
- The focus in this chapter is the relationship among continuous varaiables

Topics Covered:

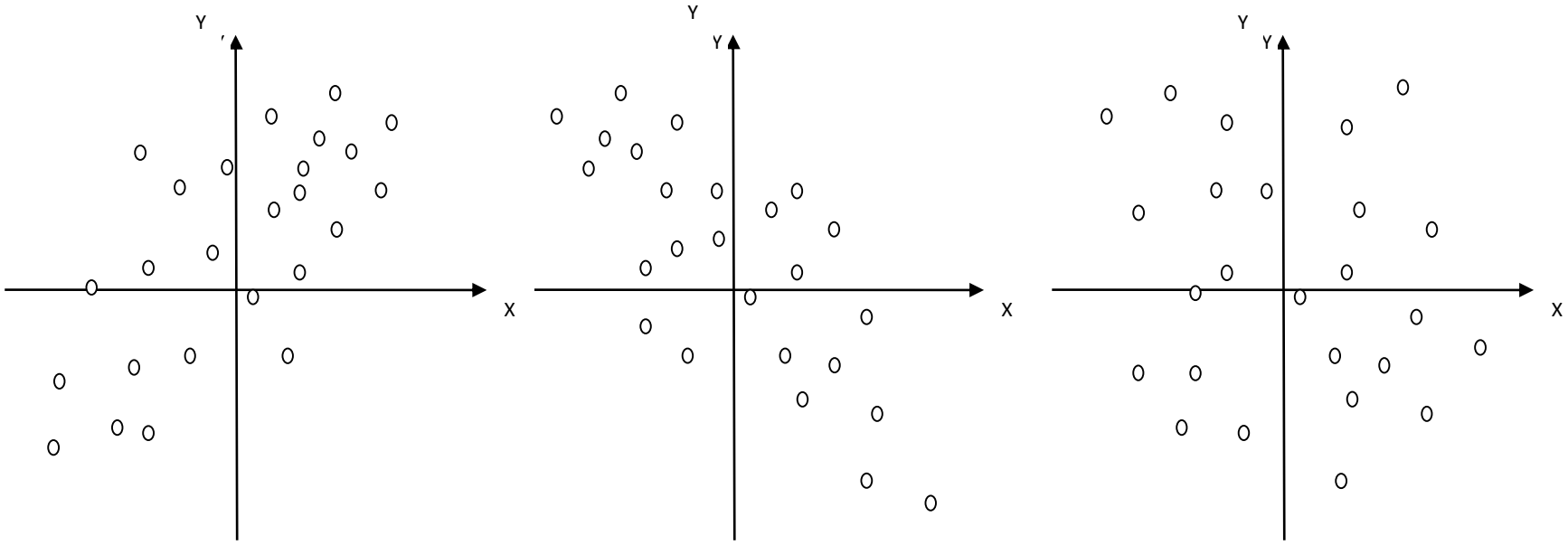
- Is there a relationship between *x* and *y*?
- What is the strength of this relationship
  - Pearson's r
- Can we describe this relationship and use this to predict *y* from *x*?
  - Regression

The relationship between *x* and *y*

- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?

- CORRELATION ≠ CAUSATION
  - In order to infer causality: manipulate independent variable and observe effect on dependent variable
  - For example, there may be a strong association between mortality and time per day spent watching movies, but before doctors should start recommending that we all should match more movies, we need to rule out another explanation- younger people watch more movies and are less likely to die.

Scattergrams



Positive correlation

Negative correlation

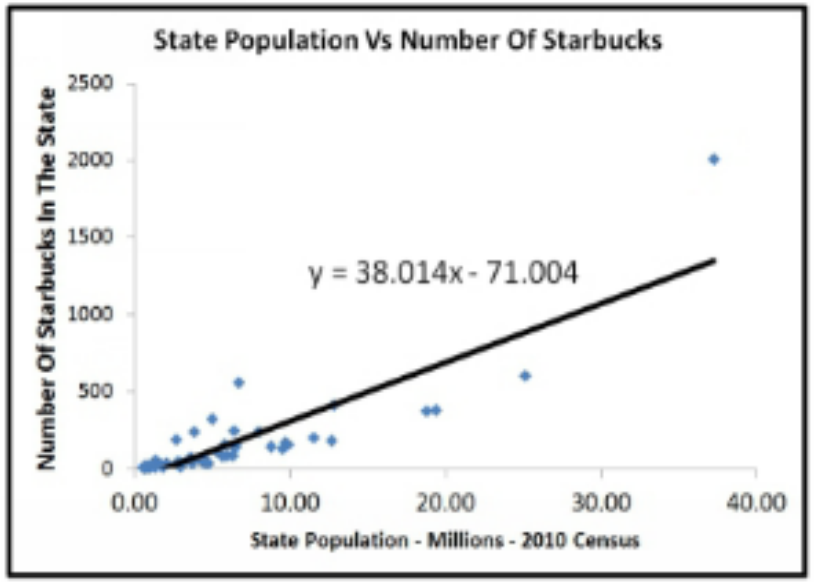
No correlation

What is linear regression

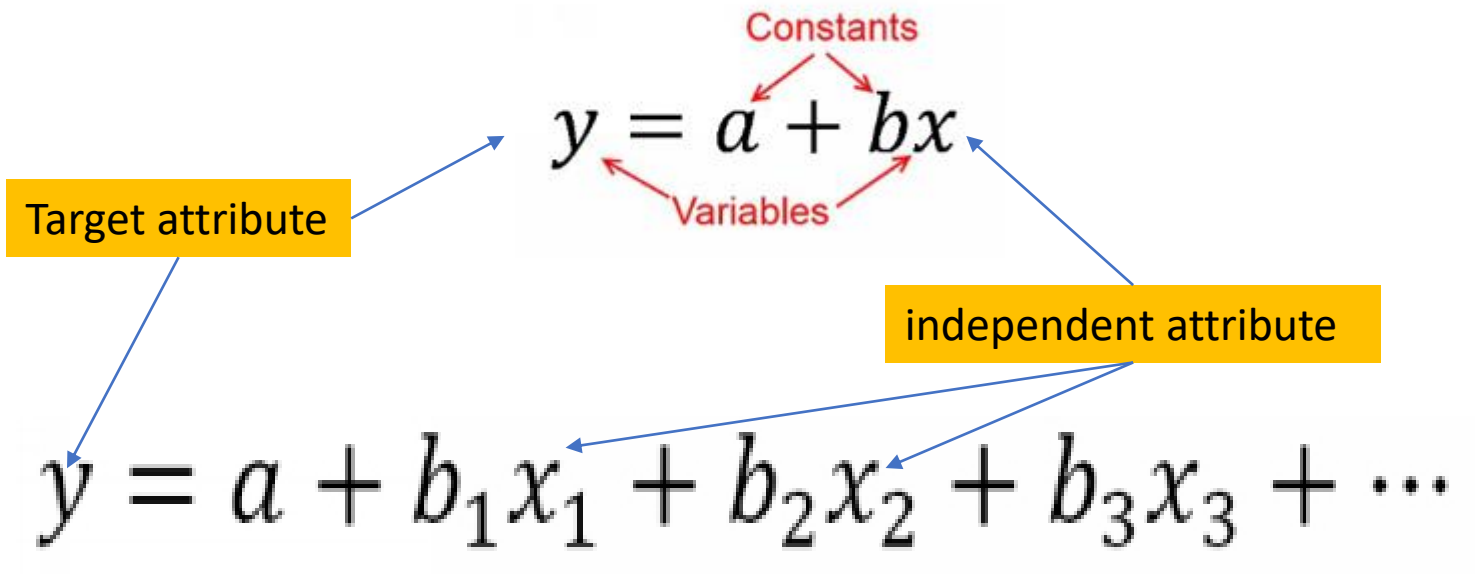
- Linear regression is a way of predicting an unknown variable using results that you do know.
- If you have a set of x and y values, you can use a regression equation to make a straight line relating the x and y.
- The reason you might want to do this is if you know some information, and want to estimate other information.
- For instance, you might have measured the fuel economy in your car when you were driving 30 miles per hour, when you were driving 40 miles per hour, and when you were driving 75 miles per hour.
- Now you are planning a cross country road trip and plan to average 60 miles per hour, and want to estimate what fuel economy you will have so that you can budget how much money you will need for gas.

Example

- The chart on the right shows an example of linear regression using real world data.
- It shows the relationship between the population of states within the United States, and the number of Starbucks (a coffee chain restaurant) within that state.



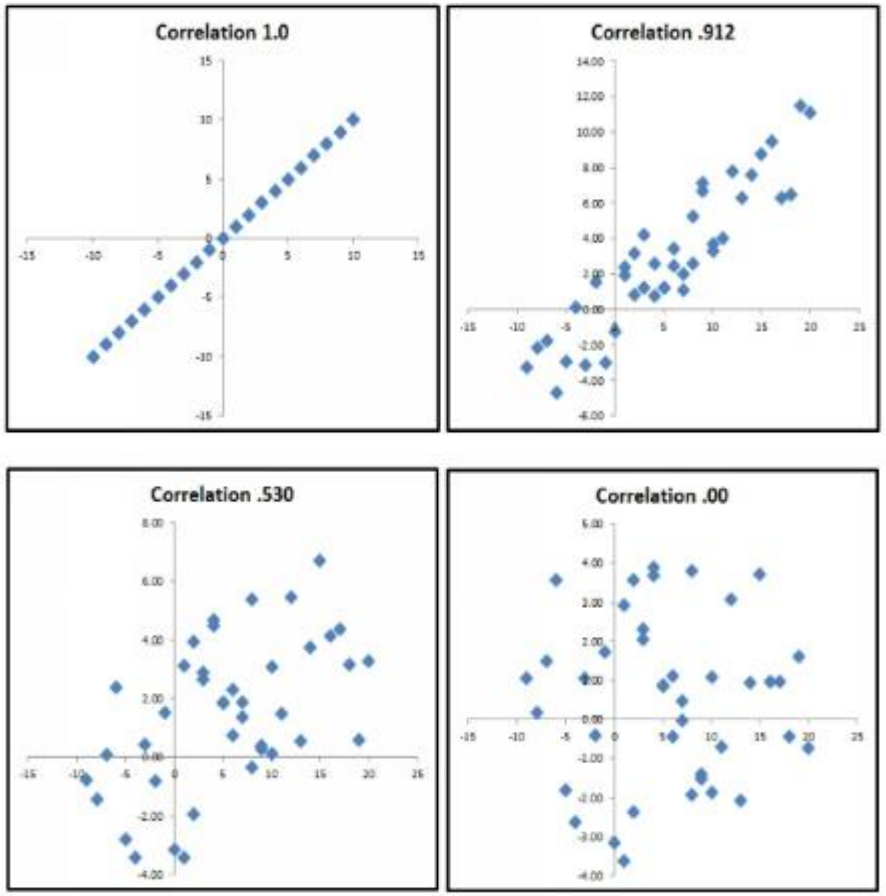
Equation of the linear regression

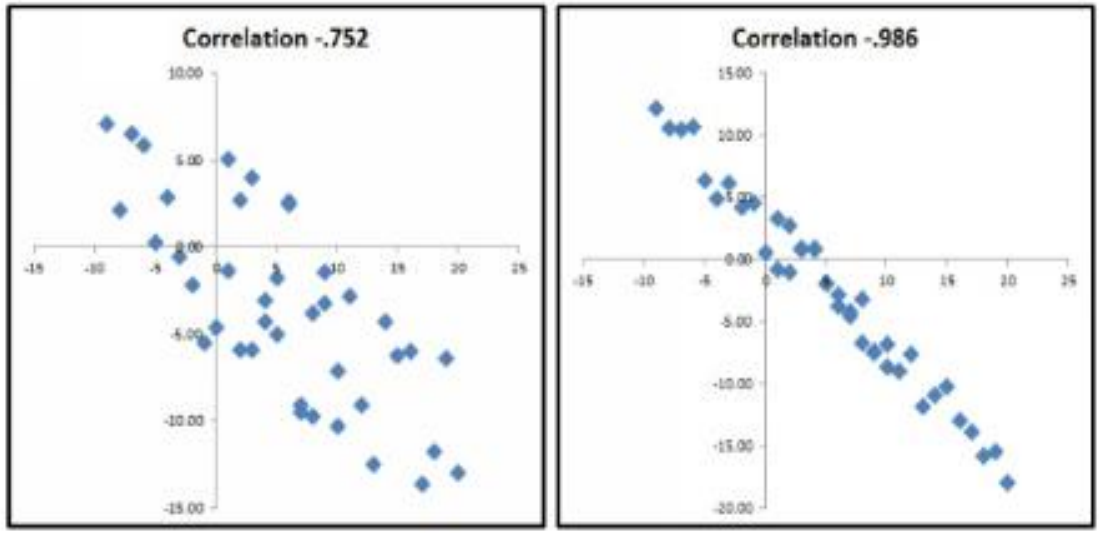


Correlation

- Correlation is a measure of how closely two variables move together.
- Pearson’s correlation coefficient is a common measure of correlation, and it ranges from +1 for two variables that are perfectly in sync with each other, to 0 when they have no correlation, to -1 when the two variables are moving opposite to each other.

Example: Correlation





Pearson's correlation coefficient

Sum Over All Data Points

x & y values of each point minus x & y mean values

$$r = \frac{\sum((x - \bar{x}) * (y - \bar{y}))}{(n - 1) * s_x * s_y}$$

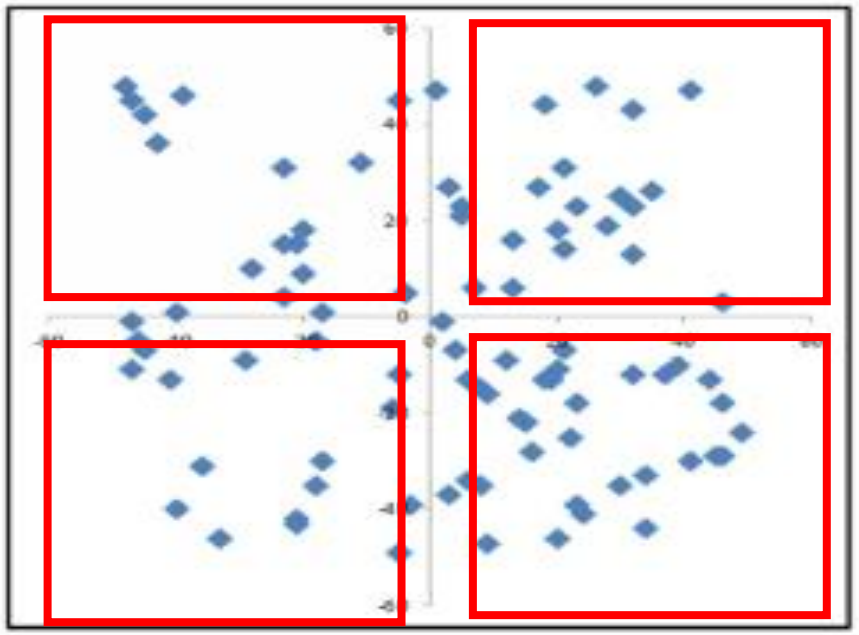
Pearson's Correlation

# of Data Points

Standard Deviation of x & y

Quadrant Four:
(x-mean(x))*(y-mean(y)) negative

Quadrant One:
(x-mean(x))*(y-mean(y)) positive



Quadrant Three:
(x-mean(x))*(y-mean(y)) positive

Quadrant Two:
(x-mean(x))*(y-mean(y)) negative