Statistics for Business and Economics

Chapter 2

Describing Data: Numerical

Measures of Central Tendency

Overview

Central Tendency

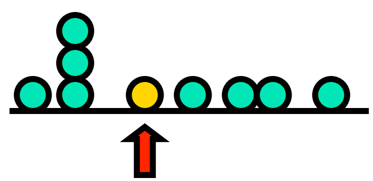
- Mean
- Median
- Mode

Mean

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

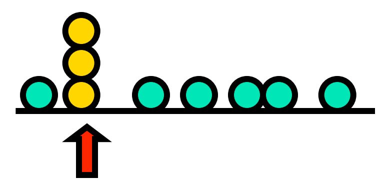
Arithmetic average

Median



Midpoint of ranked values

Mode



Most frequently observed value

Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency
  - For a population of N values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

  - For a sample of size n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

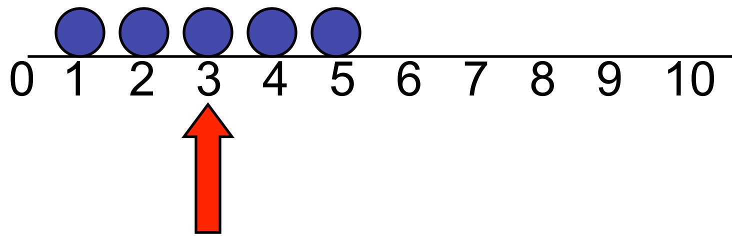
Observed values

Sample size

Arithmetic Mean

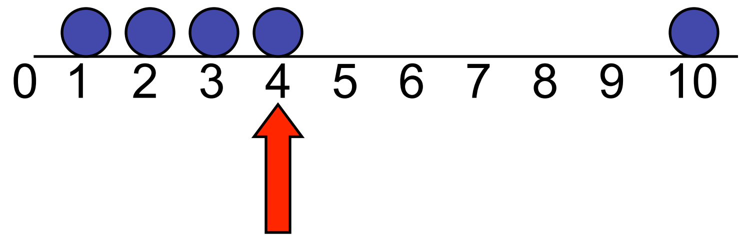
(continued)

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



Mean = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

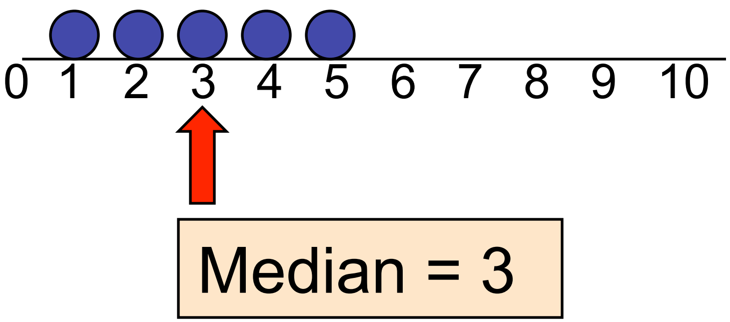


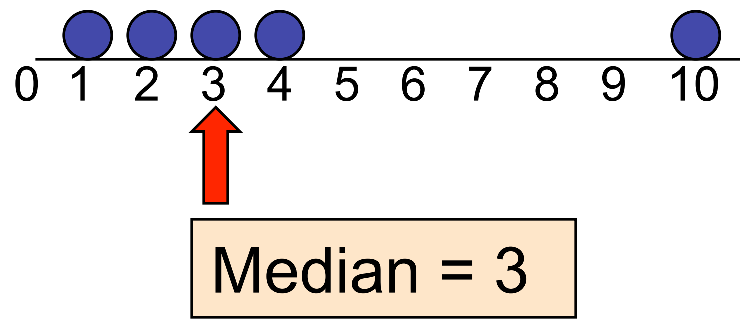
Mean = 4

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Median

- In an ordered list, the median is the “middle” number (50% above, 50% below)





- Not affected by extreme values

Finding the Median

- The location of the median:

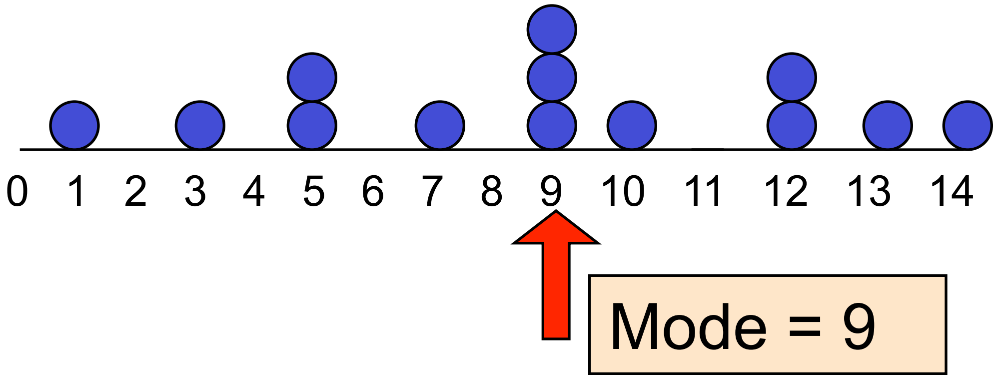
$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

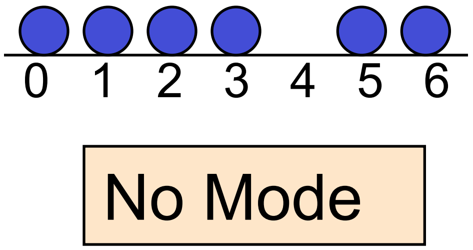
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may may be no mode
- There may be several modes



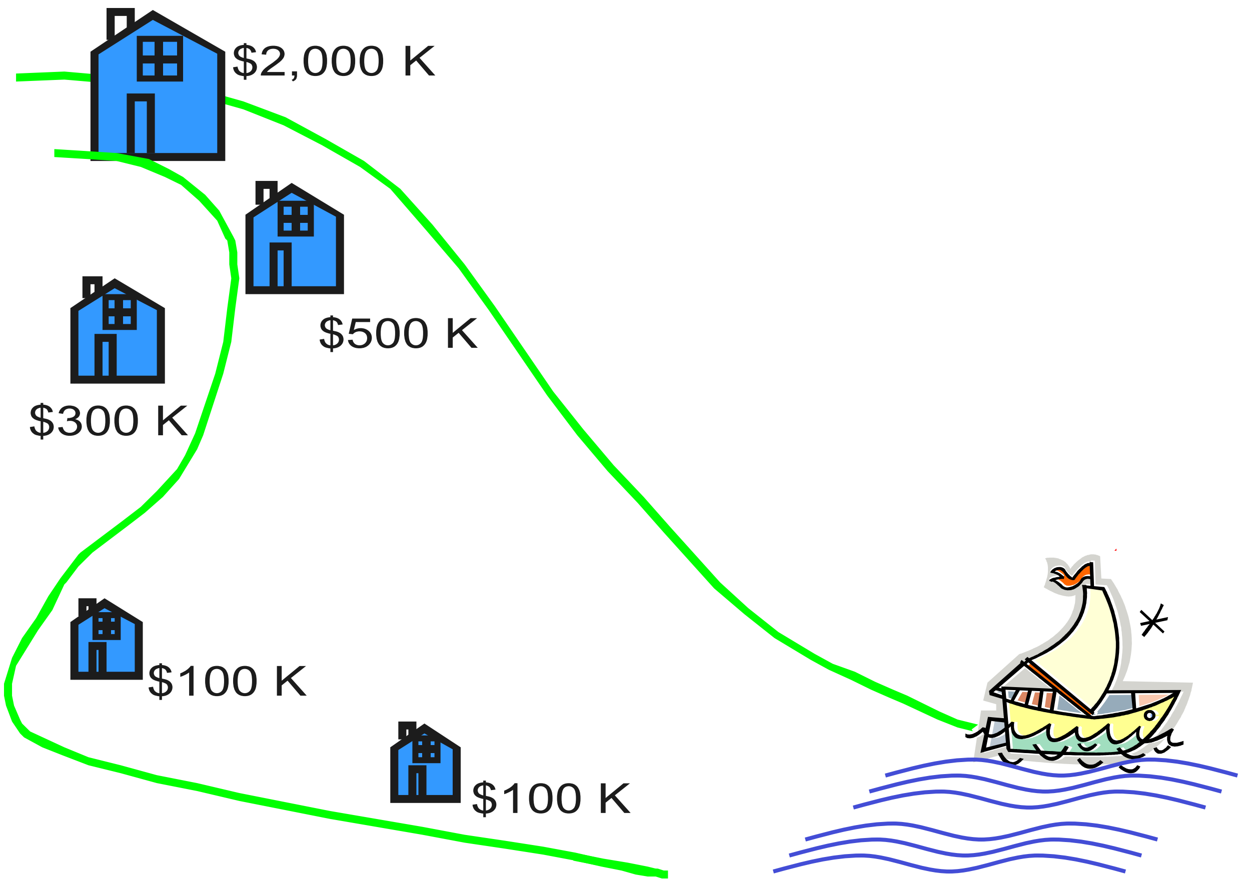


Review Example

- Five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



Review Example: Summary Statistics

House Prices:
$2,000,000
500,000
300,000
100,000
100,000
Sum 3,000,000

- **Mean:** ($3,000,000/5) = **$600,000**
- **Median:** middle value of ranked data = **$300,000**
- **Mode:** most frequent value = **$100,000**

Which measure of location is the "best"?

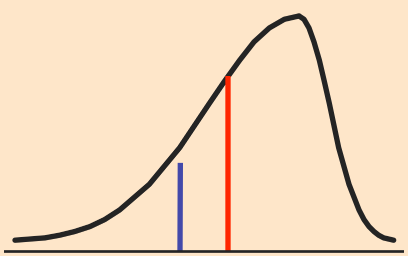
- **Mean** is generally used, unless extreme values (outliers) exist . . .
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Shape of a Distribution

- Describes how data are distributed
- Measures of shape
  - Symmetric or skewed

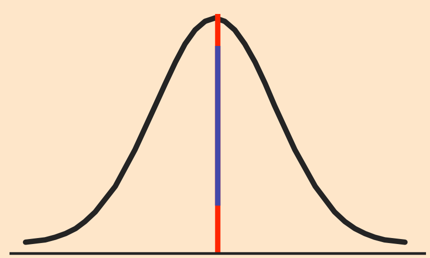
Left-Skewed

Mean < Median



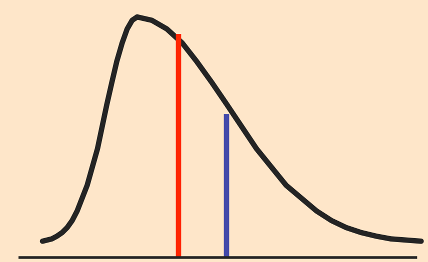
Symmetric

Mean = Median



Right-Skewed

Median < Mean



Geometric Mean

- Geometric mean

$$\bar{X}_g = \sqrt[n]{(X_1 \times X_2 \times \cdots \times X_n)} = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometric mean rate of return

$$\bar{r}_g = (X_1 \times X_2 \times \ldots \times X_n)^{1/n} - 1$$

Example

Initial Investment: $100

After 5 years: $125

Question: the average annual rate of returns?

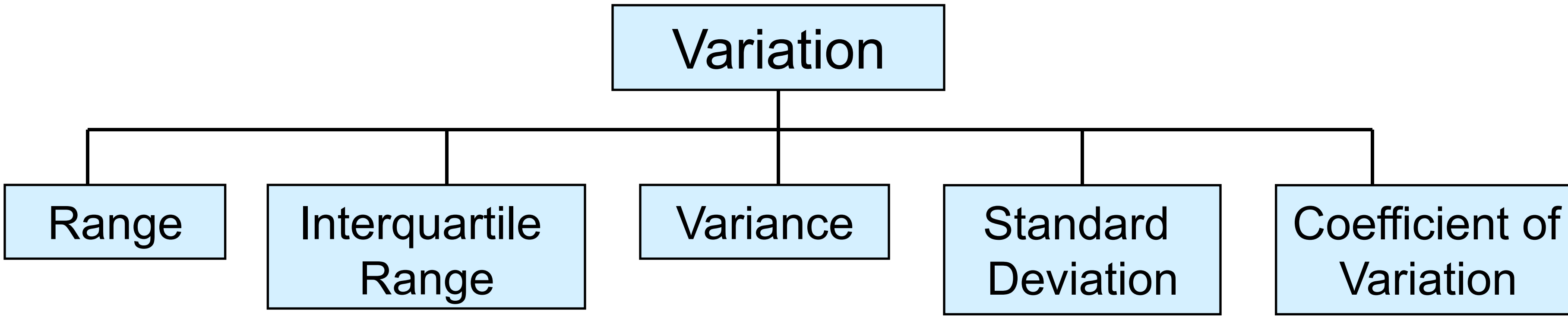
- 25 % divided by 5 years = 5%? This is Wrong!
- ``Compound Interest'' over 5 years

  $100 x (1+0.05)^5 = $127.63 > $125 after 5 years
- Answer:

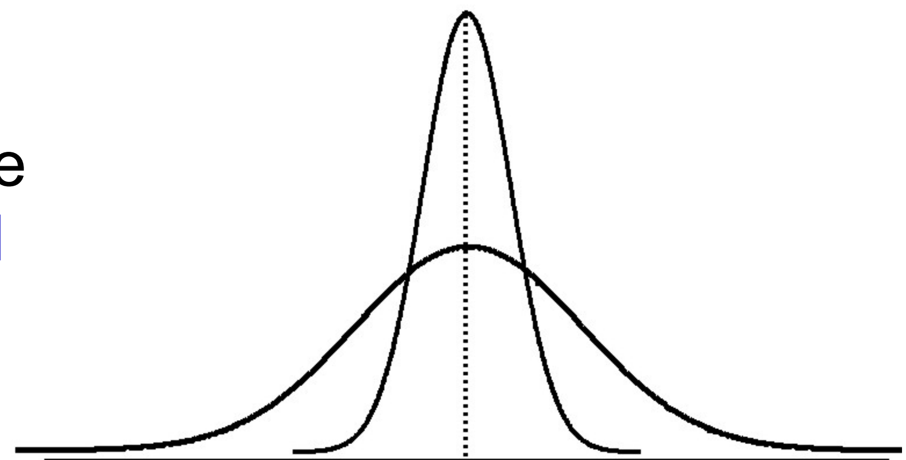
  $100 x (1+r)^5 = $125 → r = 4.6%

2.2

Measures of Variability



- Measures of variation give information on the spread or variability of the data values.



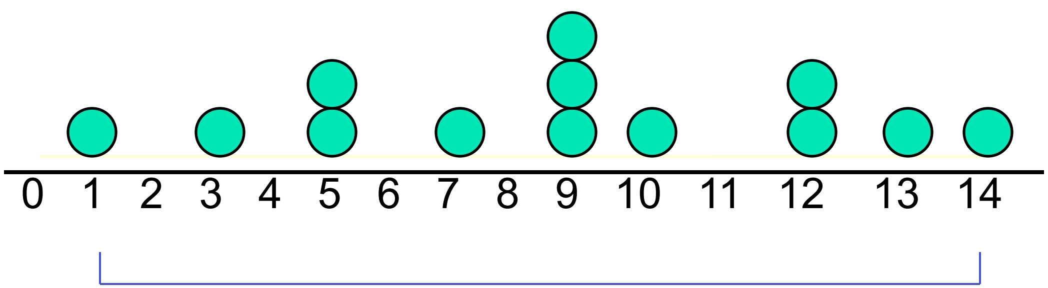
Same center, different variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{largest} - X_{smallest}$$

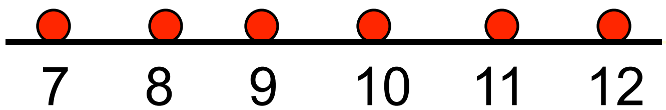
Example:



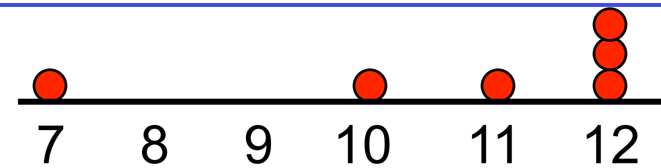
Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

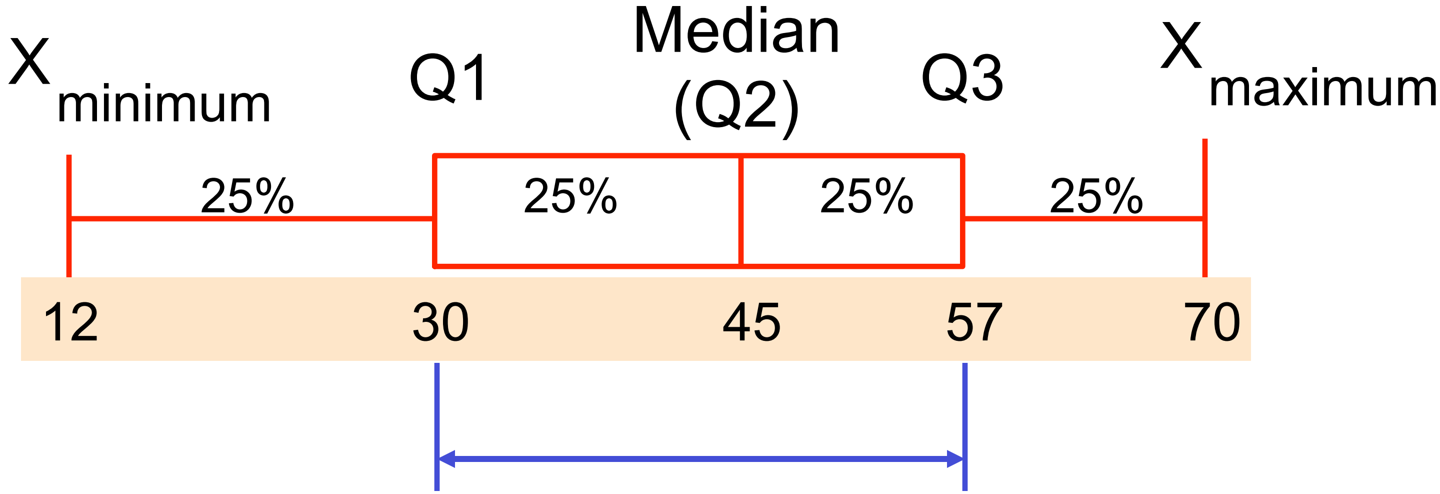
Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

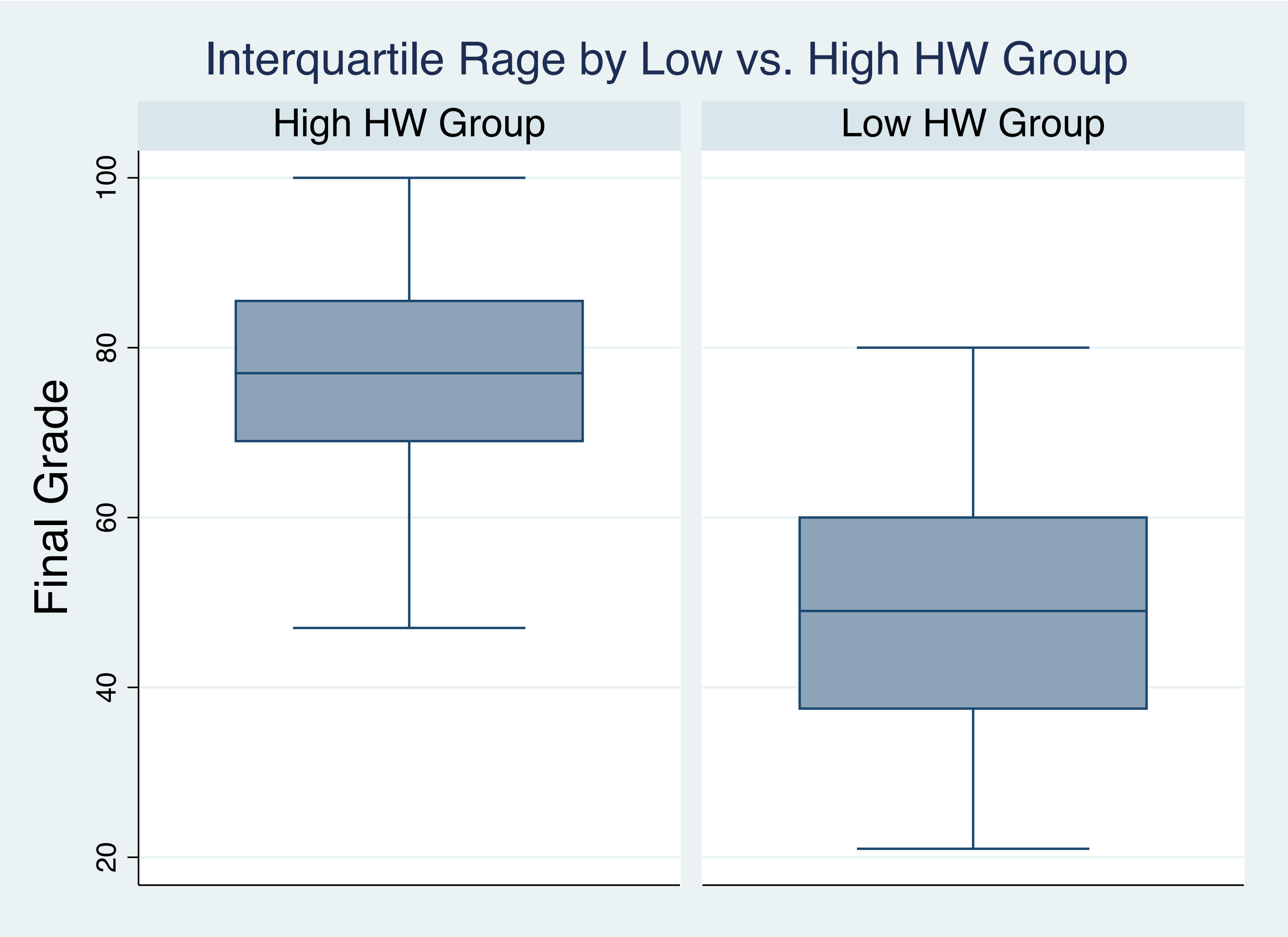
Interquartile Range

Example:



Interquartile range
= 57 – 30 = 27

STATA Example



Population Variance

- Average of squared deviations of values from the mean
  - Population variance:

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

Where μ = population mean

N = population size

x_i = i^{th} value of the variable x

Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Population standard deviation:

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

Sample Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Sample standard deviation:

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

Calculation Example: Sample Standard Deviation

Sample Data (x_i) :

10	12	14	15	17	18	18	24

n = 8 Mean = $\bar{x}$ = 16

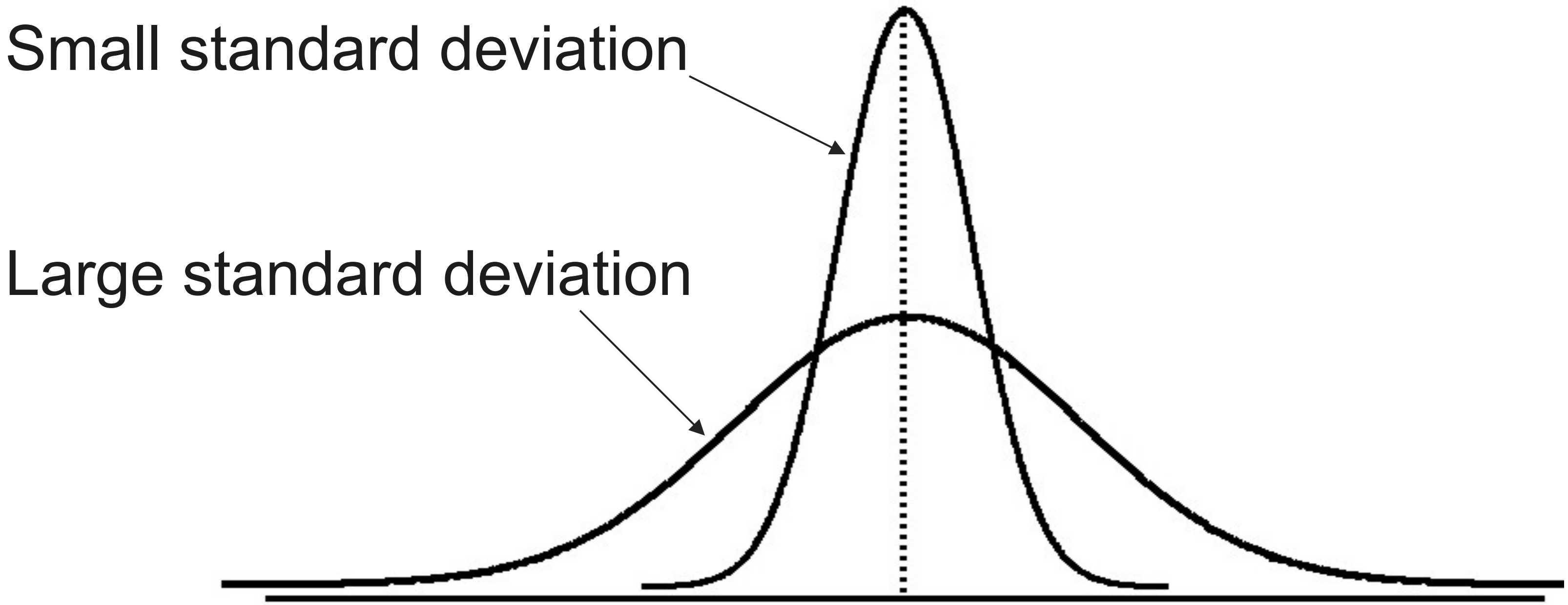
$$s = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

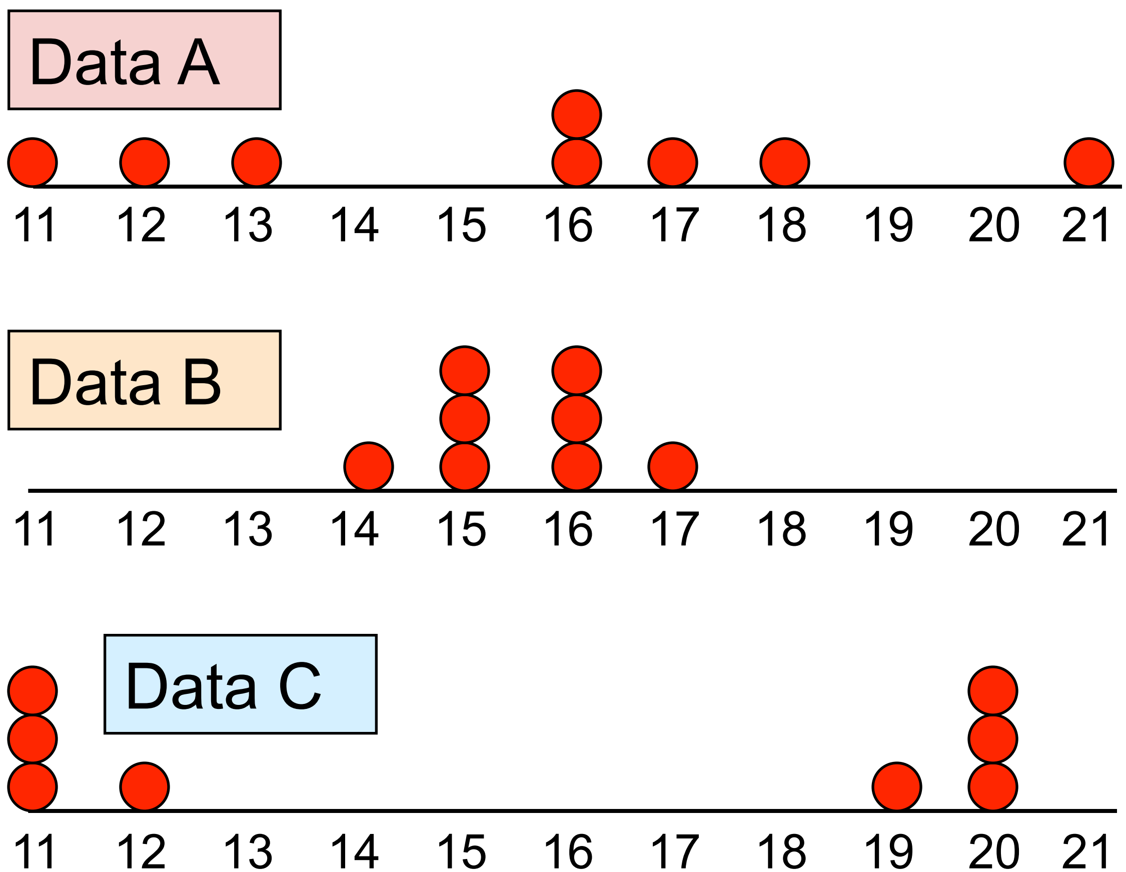
$$= \sqrt{\frac{126}{7}} = 4.2426$$

A measure of the "average" scatter around the mean

Measuring variation



Comparing Standard Deviations



Mean = 15.5
s = 3.338

Mean = 15.5
s = 0.926

Mean = 15.5
s = 4.570

Advantages of Variance and Standard Deviation

- Each value in the data set is used in the calculation
- Values far from the mean are given extra weight

  (because deviations from the mean are squared)

Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

$$CV = \left(\frac{s}{\bar{x}} \right) \cdot 100\%$$

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

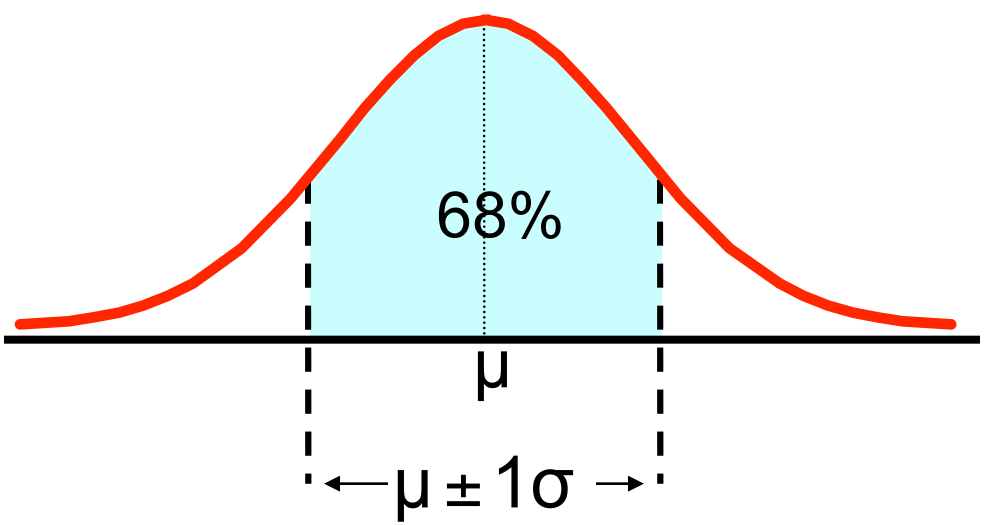
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

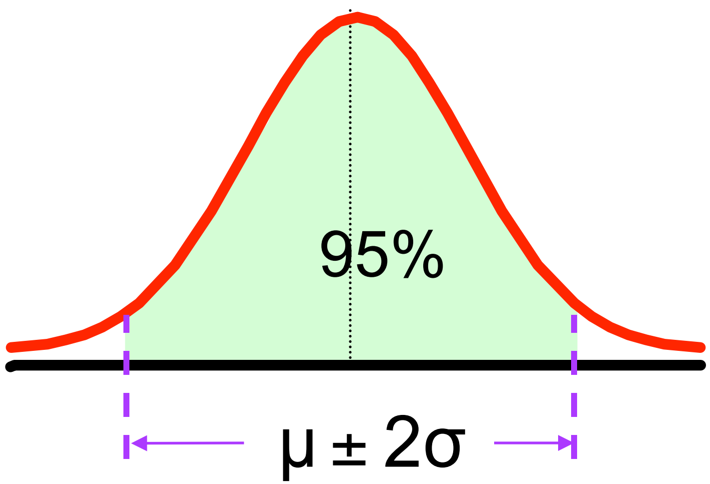
The Empirical Rule

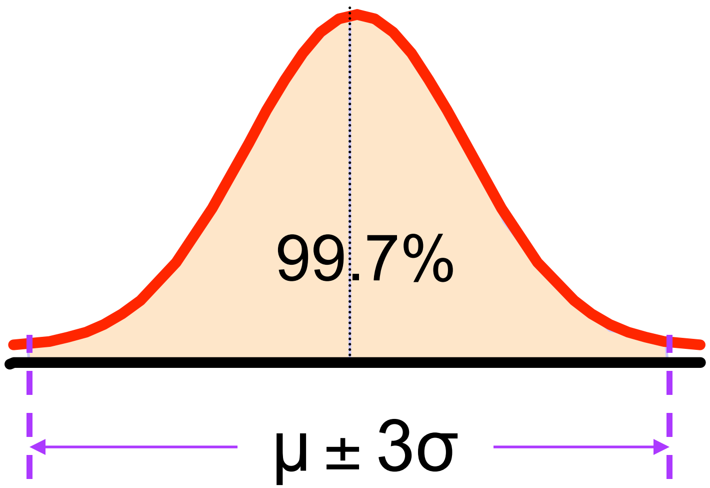
- If the data distribution is approximated by normal distribution, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample



The Empirical Rule

- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains almost all (about 99.7%) of the values in the population or the sample





2.3

Weighted Mean

- The weighted mean of a set of data is

$$\bar{X} = \sum_{i=1}^{n} w_i X_i = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$$

  - Where w_i is the weight of the i^{th} observation
    and $\sum w_i = 1$

- Use when data is already grouped into n classes, with w_i values in the i^{th} class

2.3

Example

- Consider a student with the scores of assignment (x1), midterm (x2), and final exam (x3) given by

  x1 = 90, x2 = 90, and x3 = 46.

- The weights:
- w1 = 0.1, w2 =0.3, and w3 = 0.6.
- The final grade for this student is

$$\sum_{i=1}^{3} w_i x_i = 0.1 \times 90 + 0.3 \times 90 + 0.6 \times 46 = 63.6$$

The Sample Covariance

- The covariance measures the strength of the linear relationship between **two variables**
- The population covariance:

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

- The sample covariance:

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

  - Only concerned with the strength of the relationship
  - No causal effect is implied
  - Depends on the unit of measurement

Interpreting Covariance

- **Covariance** between two variables:

Cov(x,y) > 0 → x and y tend to move in the same direction

Cov(x,y) < 0 → x and y tend to move in opposite directions

Cov(x,y) = 0 → x and y are independent

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two variables
- Population correlation coefficient:

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_X \sigma_Y}$$

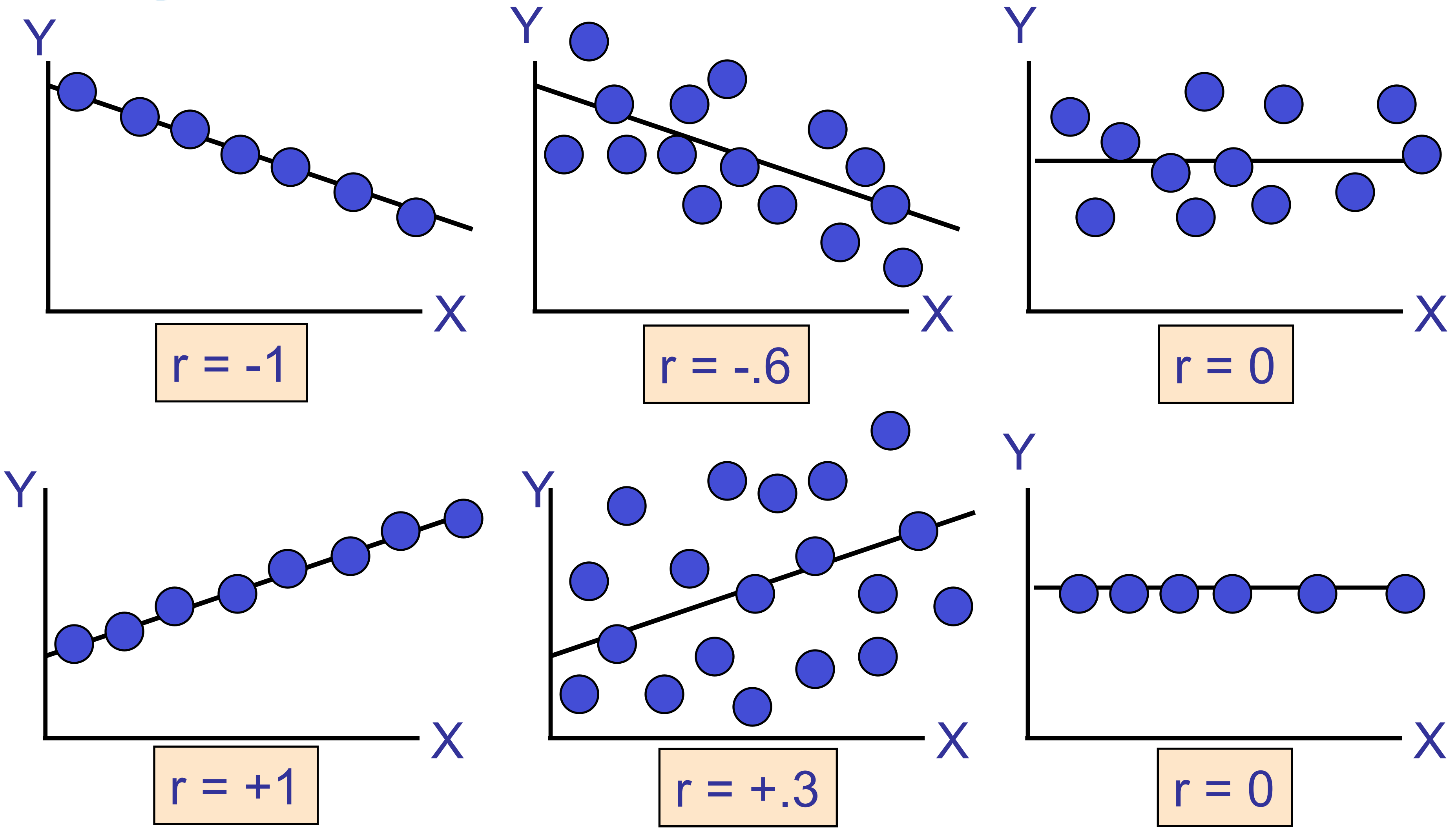
- Sample correlation coefficient:

$$r = \frac{\text{Cov}(x, y)}{s_X s_Y}$$

Features of Correlation Coefficient, r

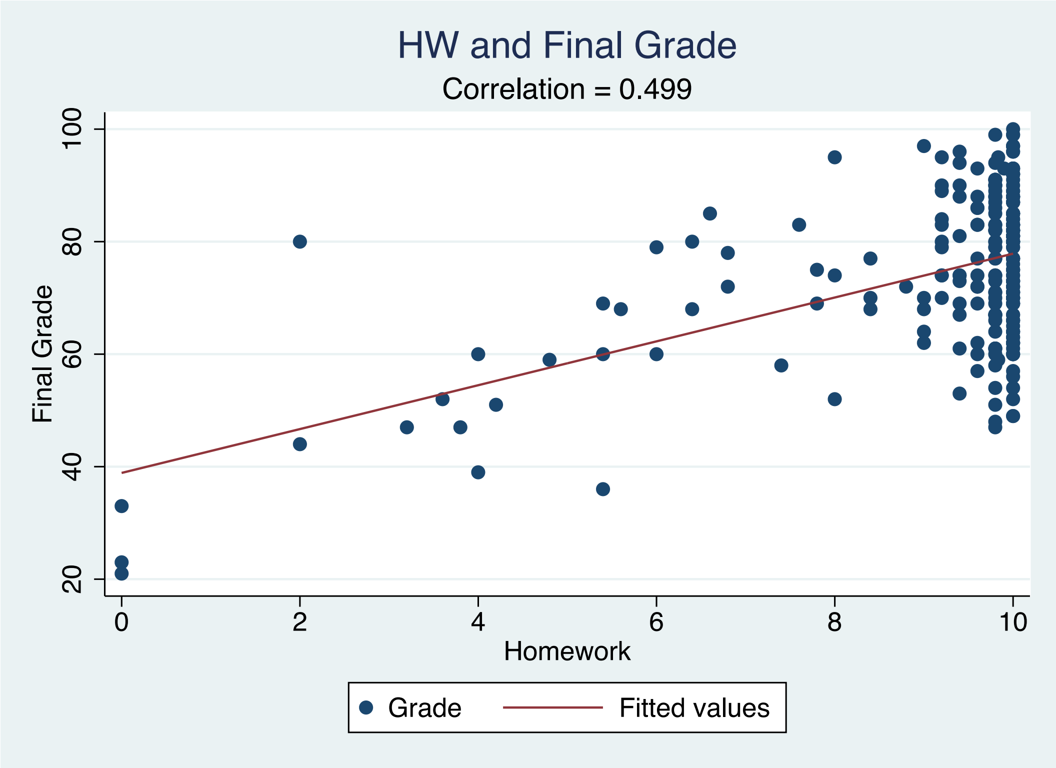
- Unit free
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients



Example: HW and Final Grade

- r = .0.499



- There is a relatively strong positive linear relationship between HW scores and Final Grades

- Students who scored high on HW assignment tended to have high final grades