Statistics for Business and Economics
8th Edition

Chapter 2

Describing Data: Numerical

Chapter Goals

After completing this chapter, you should be able to:

- Compute and interpret the mean, median, and mode for a set of data
- Find the range, variance, standard deviation, and coefficient of variation and know what these values mean
- Apply the empirical rule to describe the variation of population values around the mean
- Explain the weighted mean and when to use it
- Explain how a least squares regression line estimates a linear relationship between two variables

Chapter Topics

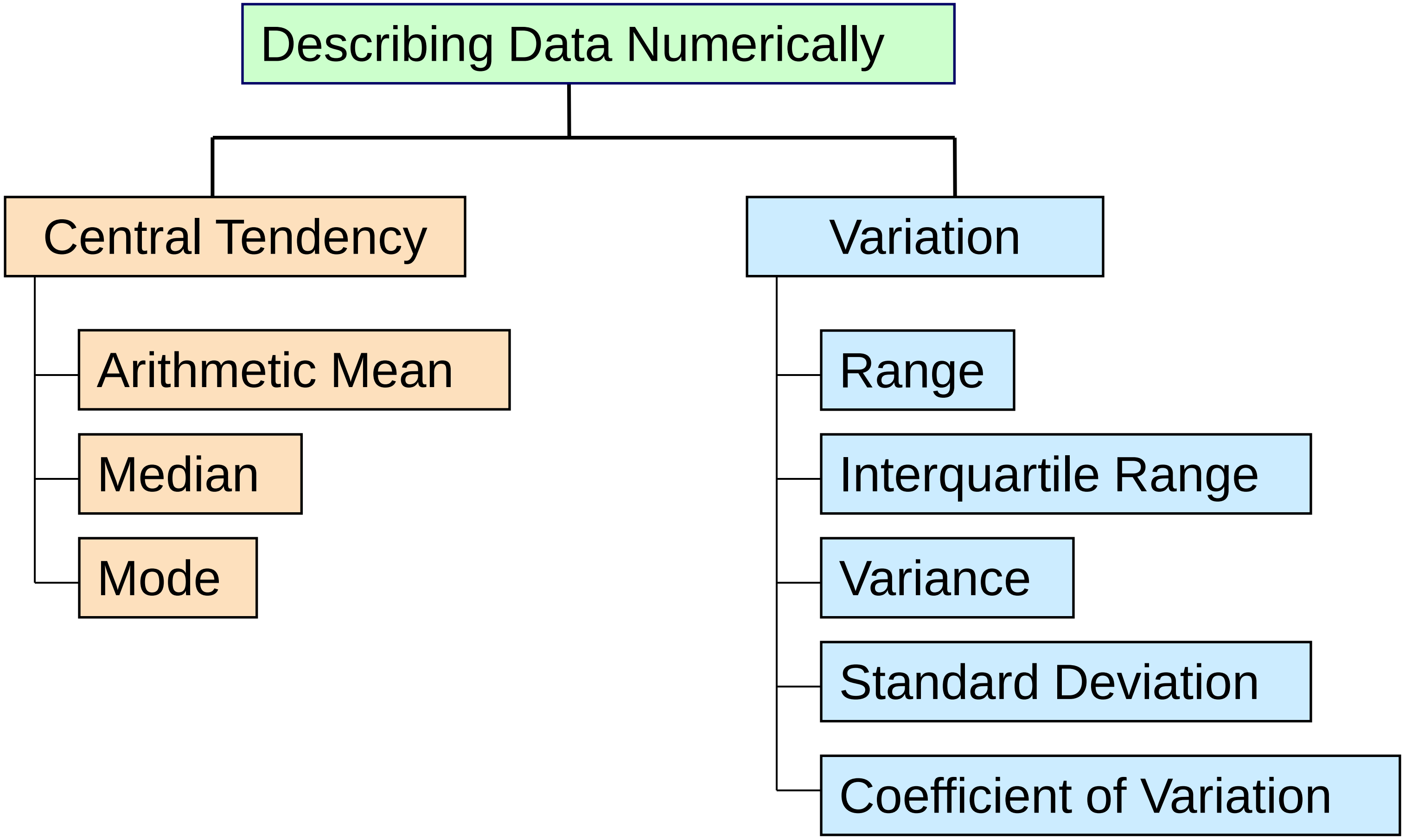
- Measures of central tendency, variation, and shape
  - Mean, median, mode, geometric mean
  - Quartiles
  - Range, interquartile range, variance and standard deviation, coefficient of variation
  - Symmetric and skewed distributions
- Population summary measures
  - Mean, variance, and standard deviation
  - The empirical rule and Chebyshev’s Theorem

Chapter Topics

(continued)

- Five number summary and box-and-whisker plots
- Covariance and coefficient of correlation
- Pitfalls in numerical descriptive measures and ethical considerations

Describing Data Numerically



2.1

Measures of Central Tendency

Overview

Central Tendency

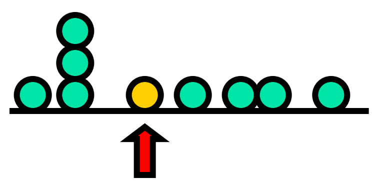
- Mean
- Median
- Mode

Mean

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

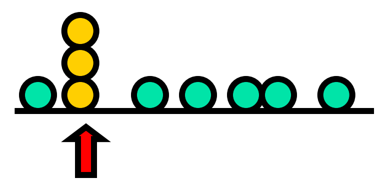
Arithmetic average

Median



Midpoint of ranked values

Mode



Most frequently observed value

(if one exists)

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:

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

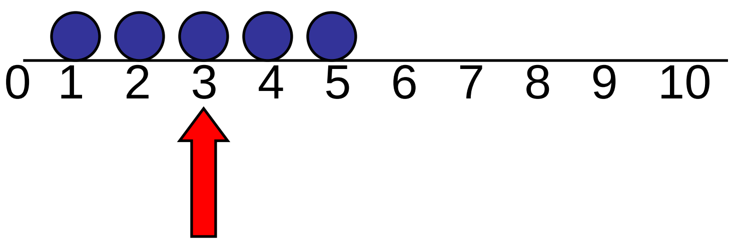
Observed values

Sample size

Arithmetic Mean

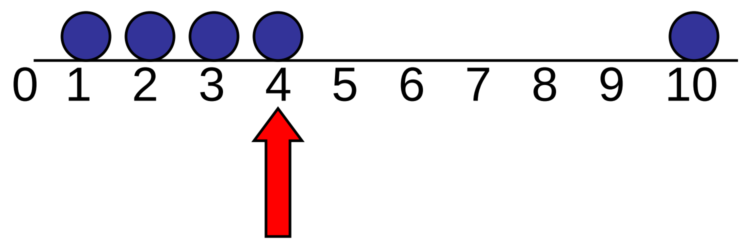
(continued)

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



Mean = 3

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

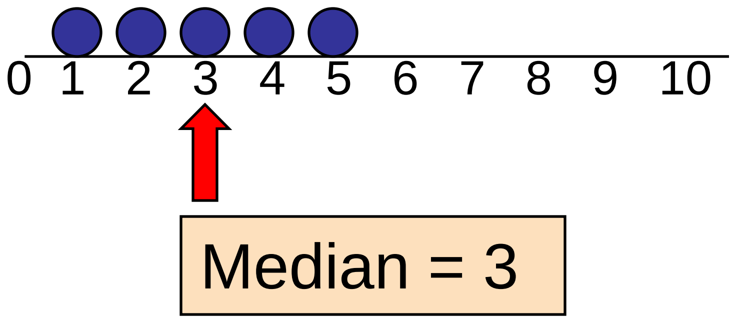


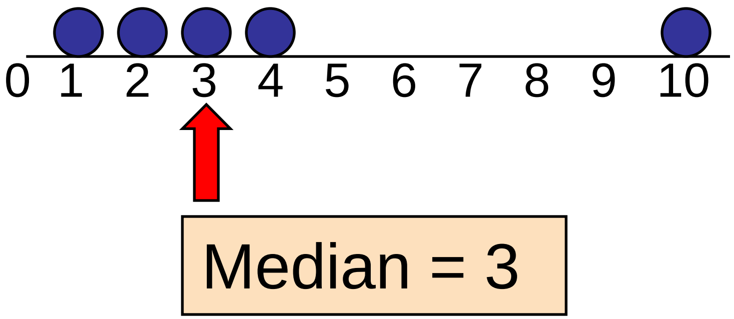
Mean = 4

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

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:

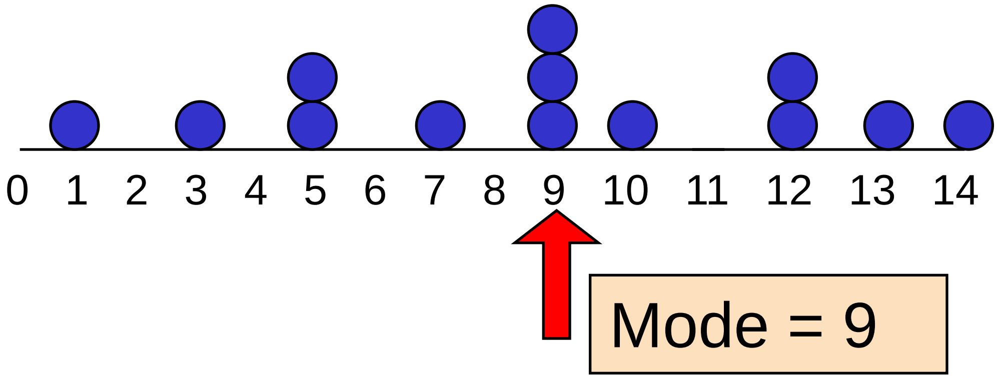
$$\text{Median position} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ position in the ordered data}$$

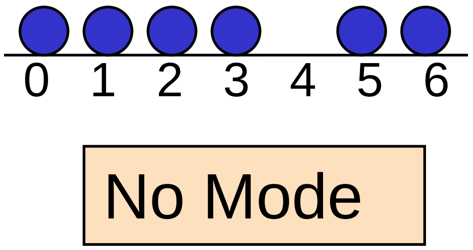
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- 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



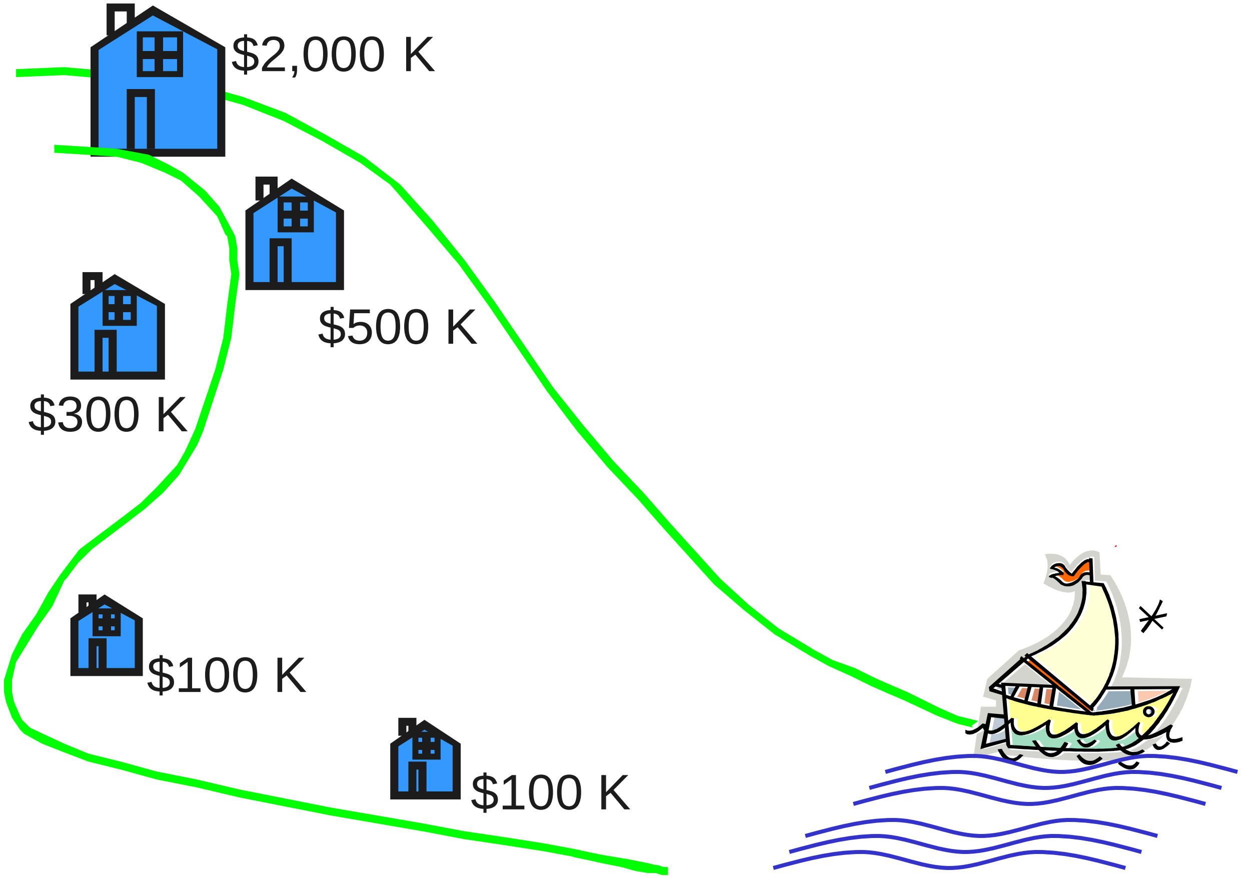


Review Example

- 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"?

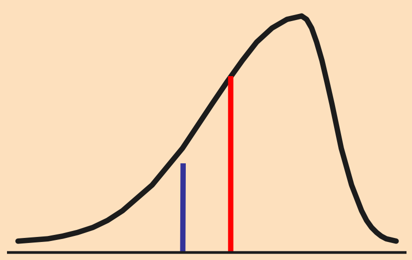
- **Mean** is generally used, unless extreme values (outliers) exist . . .
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Shape of a Distribution

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

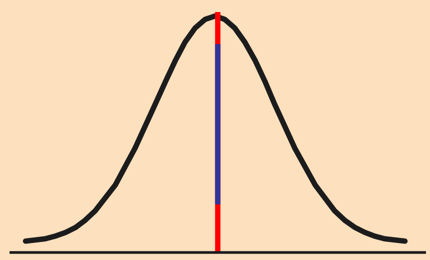
Left-Skewed

Mean < Median



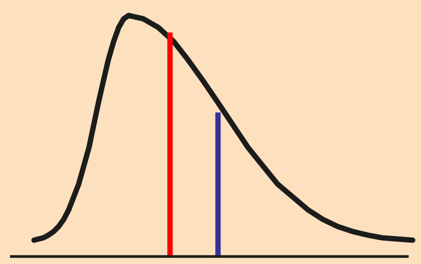
Symmetric

Mean = Median



Right-Skewed

Median < Mean



Geometric Mean

- Geometric mean
  - Used to measure the rate of change of a variable over time

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

- Geometric mean rate of return
  - Measures the status of an investment over time

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

  - Where X_i is the rate of return in time period i

Example

An investment of $100,000 rose to $150,000 at the end of year one and increased to $180,000 at end of year two:

$$X_1 = \$100{,}000 \qquad X_2 = \$150{,}000 \qquad X_3 = \$180{,}000$$

50% increase

20% increase

What is the mean percentage return over time?

Example

(continued)

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

$$\bar{X} = \frac{(50\%) + (20\%)}{2} = 35\%$$

Misleading result

Geometric mean rate of return:

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

$$= [(50) \times (20)]^{1/2} - 1$$

$$= (1000)^{1/2} - 1 = 31.623 - 1 = 30.623\%$$

Accurate result

Percentiles and Quartiles

- Percentiles and Quartiles indicate the position of a value relative to the entire set of data

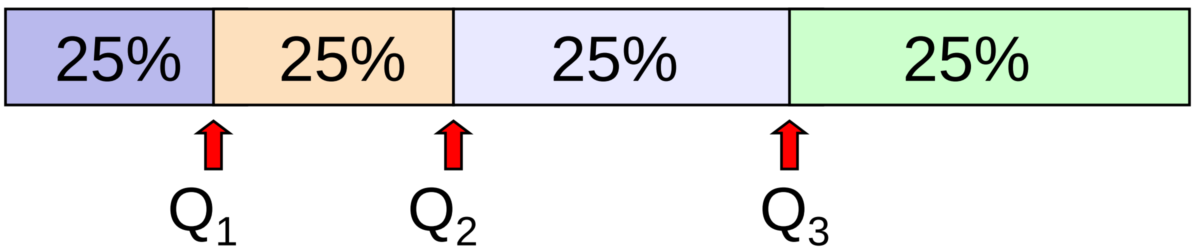
- Generally used to describe large data sets

- Example: An IQ score at the 90^{th} percentile means that 10% of the population has a higher IQ score and 90% have a lower IQ score.

P^{th} percentile = value located in the $(P/100)(n + 1)^{th}$ ordered position

Quartiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment (note that the widths of the segments may be different)



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Formulas

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position:	$Q_1 = 0.25(n+1)$
Second quartile position: (the median position)	$Q_2 = 0.50(n+1)$
Third quartile position:	$Q_3 = 0.75(n+1)$

where n is the number of observed values

Quartiles

- Example: Find the first quartile

Sample Ranked Data:	11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 = is in the 0.25(9+1) = 2.5 position of the ranked data

so use the value half way between the 2nd and 3rd values,

so $Q_1 = 12.5$

Five-Number Summary

The **five-number summary** refers to five descriptive measures:

minimum
first quartile
median
third quartile
maximum

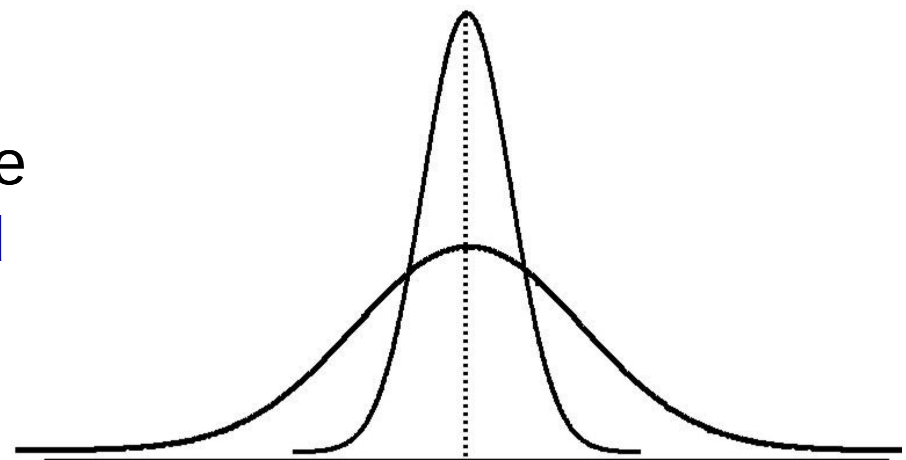
$$\text{minimum} < Q_1 < \text{median} < Q_3 < \text{maximum}$$

2.2

Measures of Variability

- Variation
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

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



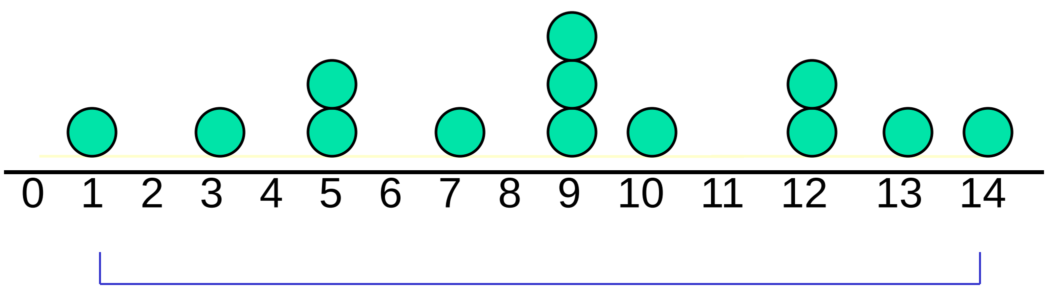
Same center, different variation

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

- Can eliminate some outlier problems by using the interquartile range

- Eliminate high- and low-valued observations and calculate the range of the middle 50% of the data

- Interquartile range = 3rd quartile – 1st quartile

$$IQR = Q_3 – Q_1$$

Interquartile Range

- The interquartile range (IQR) measures the spread in the middle 50% of the data

- Defined as the difference between the observation at the third quartile and the observation at the first quartile

$$IQR = Q_3 - Q_1$$

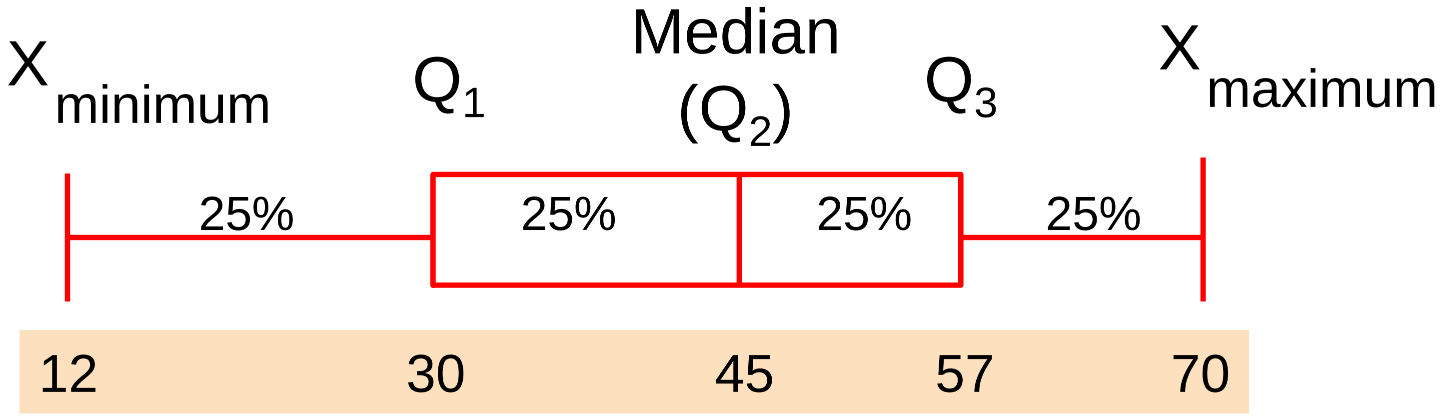
Box-and-Whisker Plot

- A box-and-whisker plot is a graph that describes the shape of a distribution
- Created from the five-number summary: the minimum value, Q_1, the median, Q_3, and the maximum
- The inner box shows the range from Q_1 to Q_3, with a line drawn at the median
- Two "whiskers" extend from the box. One whisker is the line from Q_1 to the minimum, the other is the line from Q_3 to the maximum value

Box-and-Whisker Plot

The plot can be oriented horizontally or vertically

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

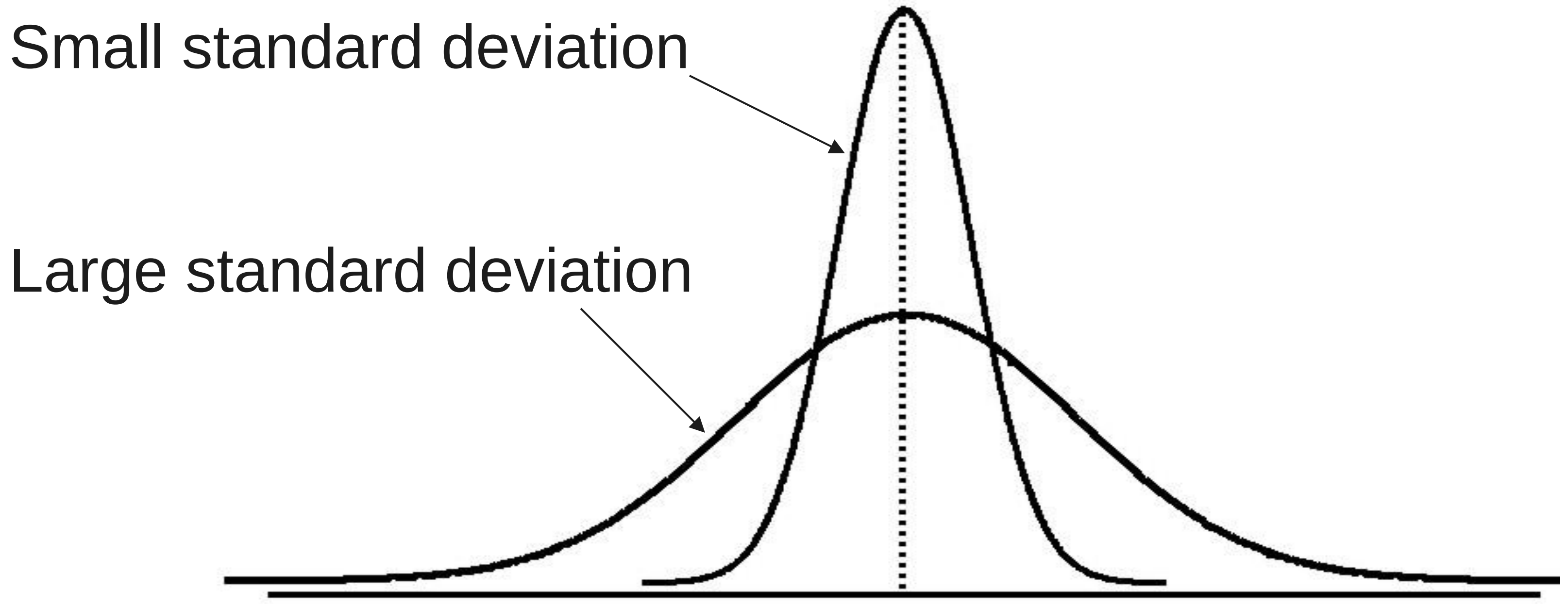
$$s = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

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

$$= \sqrt{\frac{130}{7}} = 4.3095$$

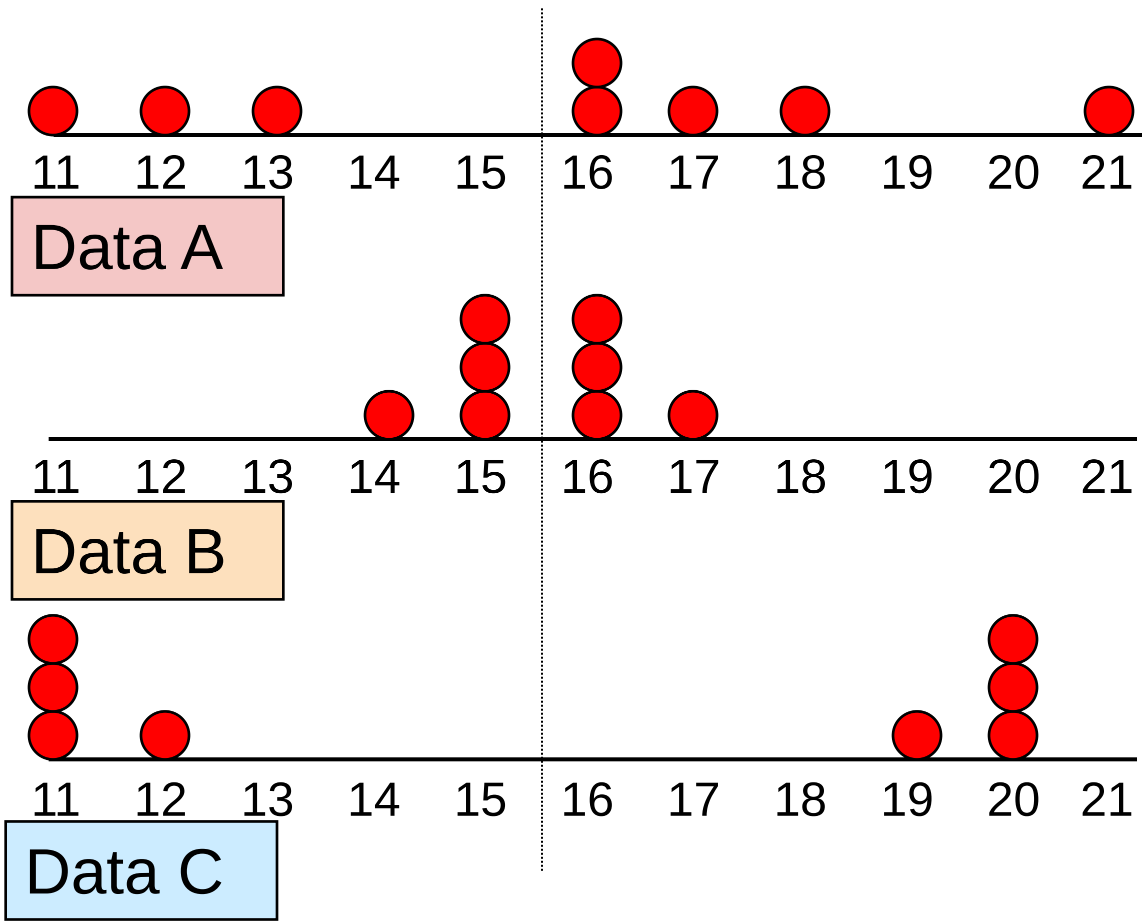
→ A measure of the “average” scatter around the mean

Measuring variation



Comparing Standard Deviations

Mean = 15.5 for each data set



s = 3.338
(compare to the two cases below)

s = 0.926
(values are concentrated near the mean)

s = 4.570
(values are dispersed far from the mean)

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)

Using Microsoft Excel

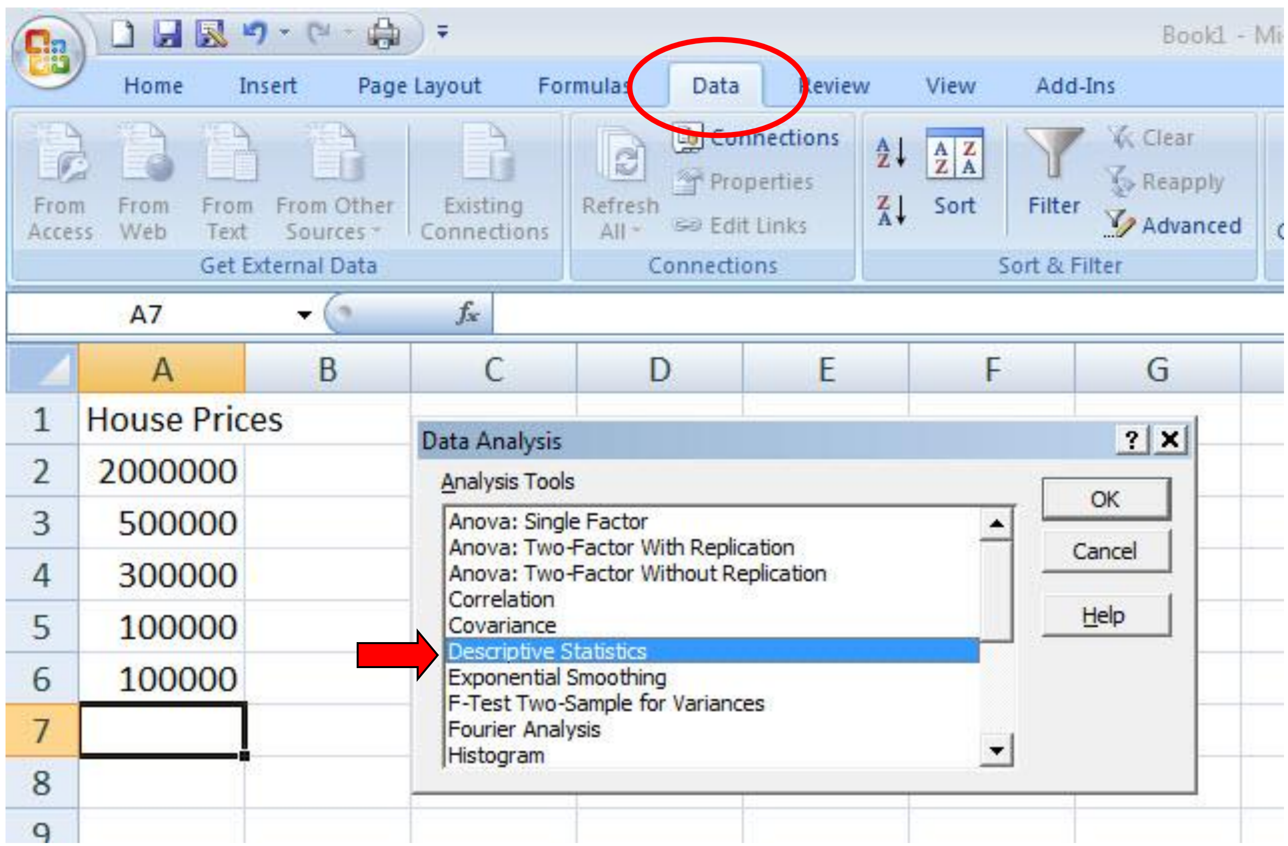
- Descriptive Statistics can be obtained from Microsoft® Excel
  - Select:

    data / data analysis / descriptive statistics

  - Enter details in dialog box

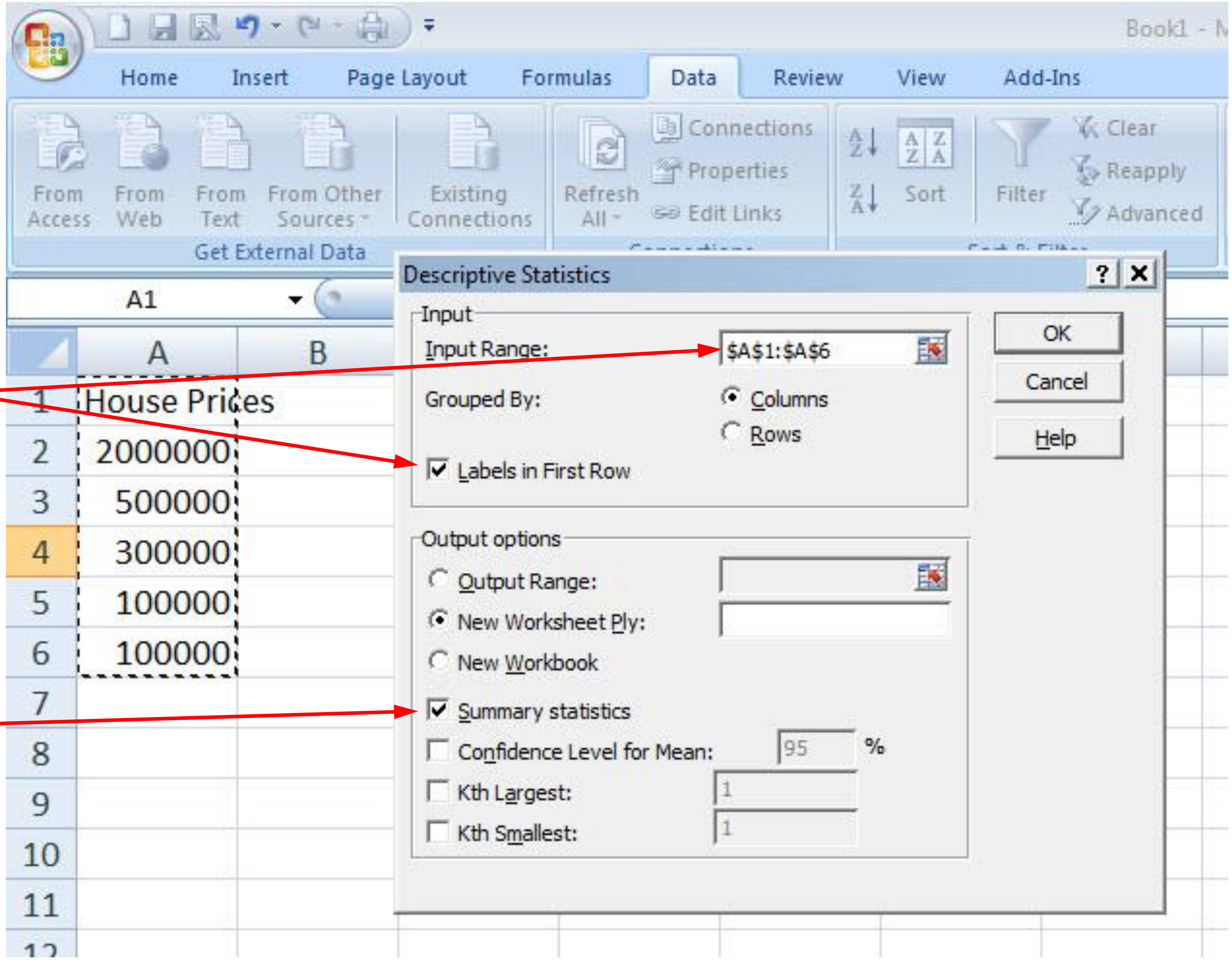
Using Excel

- Select data / data analysis / descriptive statistics



Using Excel

- Enter input range details
- Check box for summary statistics
- Click OK



Excel output

Microsoft Excel descriptive statistics output, using the house price data:

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

	A	B
1	*House Prices*	
2		
3	Mean	600000
4	Standard Error	357770.8764
5	Median	300000
6	Mode	100000
7	Standard Deviation	800000
8	Sample Variance	6.4E+11
9	Kurtosis	4.130126953
10	Skewness	2.006835938
11	Range	1900000
12	Minimum	100000
13	Maximum	2000000
14	Sum	3000000
15	Count	5
16		

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

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu} \right) \cdot 100\%$$

Sample coefficient of variation:

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

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

Chebychev's Theorem

- For any population with mean μ and standard deviation σ , and $k > 1$, the percentage of observations that fall within the interval

$$[\mu + k\sigma]$$

Is *at least*

$$100[1-(1/k^2)]\%$$

Chebychev's Theorem

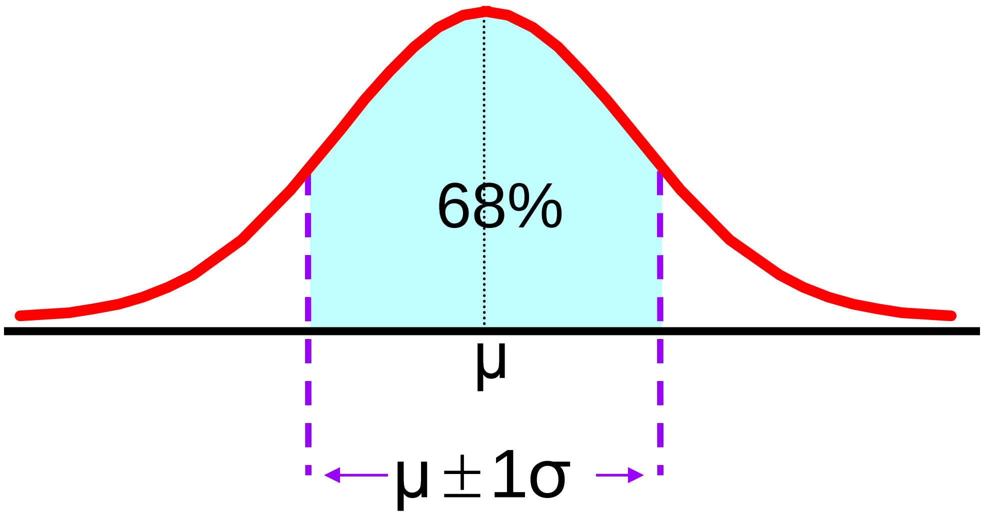
(continued)

- Regardless of how the data are distributed, at least $(1 - 1/k^2)$ of the values will fall within k standard deviations of the mean (for $k > 1$)
  - Examples:

At least		within	
$(1 - 1/1.5^2) =$	55.6%	.... $k = 1.5$	$(\mu \pm 1.5\sigma)$
$(1 - 1/2^2) =$	75%	.... $k = 2$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) =$	89%	.... $k = 3$	$(\mu \pm 3\sigma)$

The Empirical Rule

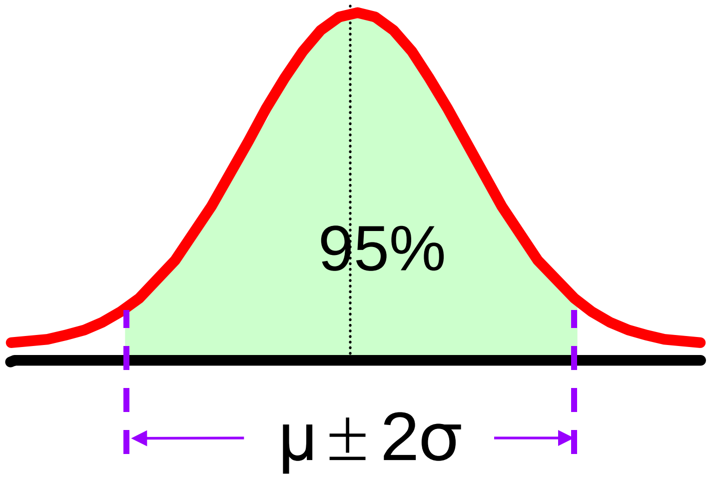
- If the data distribution is bell-shaped, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample

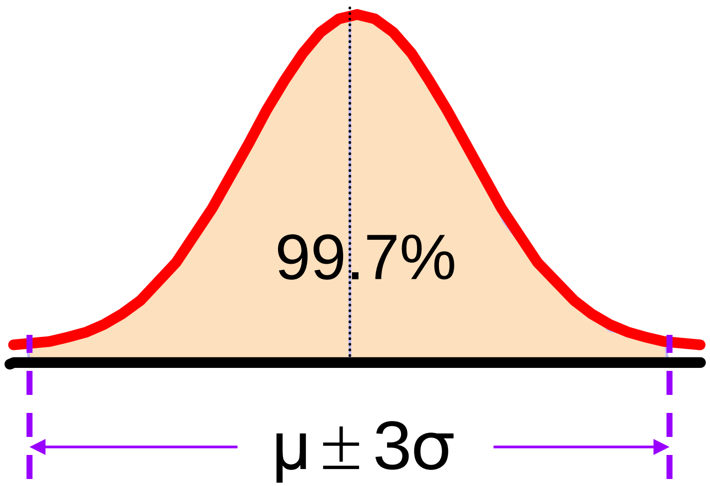


The Empirical Rule

(continued)

- $\mu \pm 2\sigma$ 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





z-Score

A z-score shows the position of a value relative to the mean of the distribution.

- indicates the number of standard deviations a value is from the mean.
  - A z-score greater than zero indicates that the value is greater than the mean
  - a z-score less than zero indicates that the value is less than the mean
  - a z-score of zero indicates that the value is equal to the mean.

z-Score

(continued)

- If the data set is the entire population of data and the population mean, μ, and the population standard deviation, σ, are known, then for each value, x_i, the z-score associated with x_i is

$$z = \frac{x_i - \mu}{\sigma}$$

z-Score

(continued)

- If intelligence is measured for a population using an IQ score, where the mean IQ score is 100 and the standard deviation is 15, what is the z-score for an IQ of 121?

$$z = \frac{x_i - \mu}{\sigma} = \frac{121 - 100}{15} = 1.4$$

A score of 121 is 1.4 standard deviations above the mean.

Weighted Mean and Measures of Grouped Data

- The weighted mean of a set of data is

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

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

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

Approximations for Grouped Data

Suppose data are grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, and the midpoints of the classes are $m_1, m_2, \ldots, m_K$

- For a sample of n observations, the mean is

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}$$

where $n = \sum_{i=1}^{K} f_i$

Approximations for Grouped Data

Suppose data are grouped into K classes, with frequencies $f_1, f_2, \ldots, f_K$, and the midpoints of the classes are $m_1, m_2, \ldots, m_K$

- For a sample of n observations, the variance is

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

Measures of Relationships Between Variables

2.4

Two measures of the relationship between variable are

- Covariance
  - a measure of the direction of a linear relationship between two variables
- Correlation Coefficient
  - a measure of both the direction and the strength of a linear relationship between two variables

Covariance

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

- The population covariance:

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

- The sample covariance:

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

  - Only concerned with the strength of the relationship
  - No causal effect is implied

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:

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

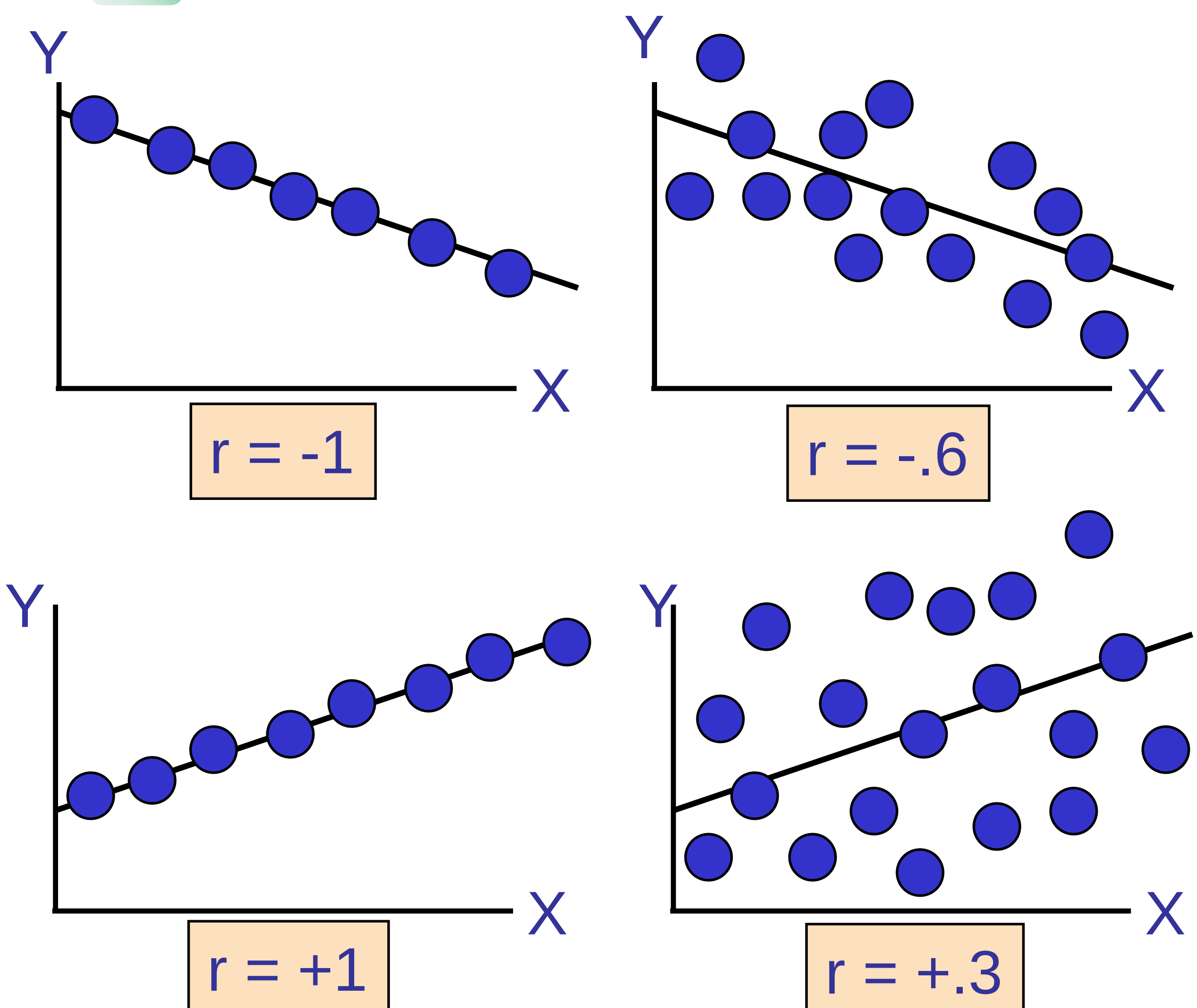
- Sample correlation coefficient:

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

Features of Correlation Coefficient, r

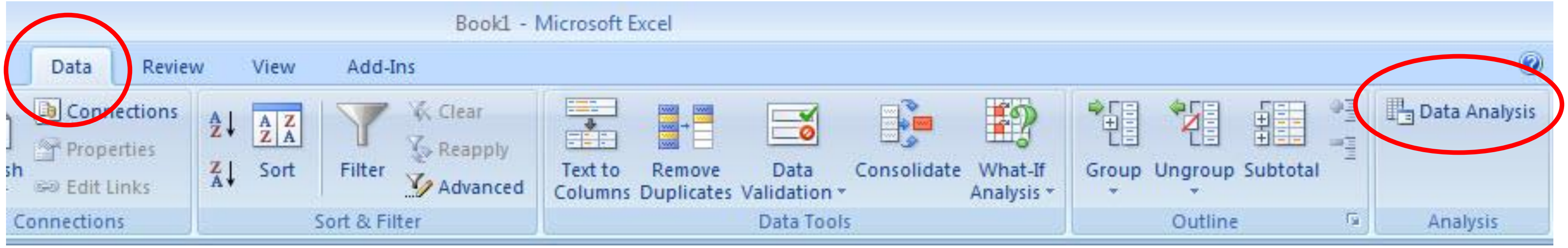
- Unit free
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

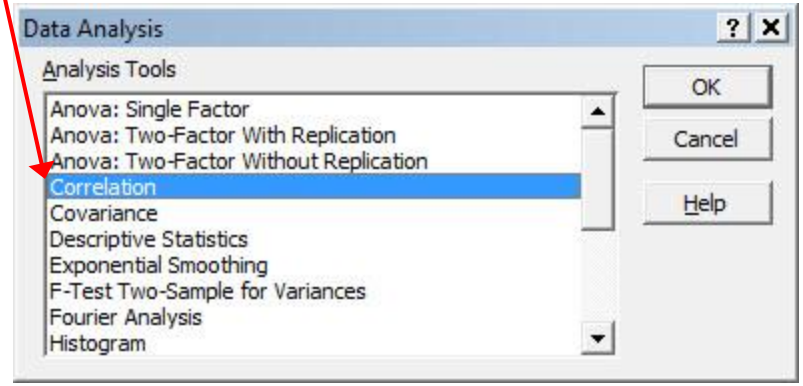


Using Excel to Find the Correlation Coefficient

- Select Data / Data Analysis

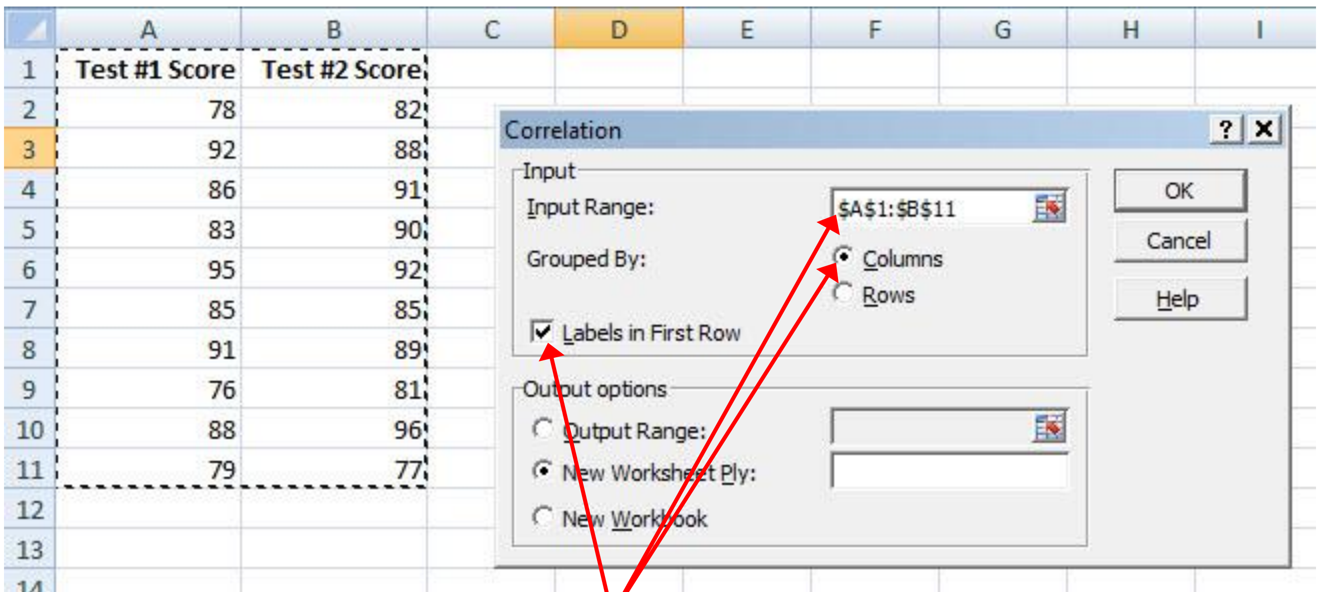


- Choose Correlation from the selection menu
- Click OK . . .

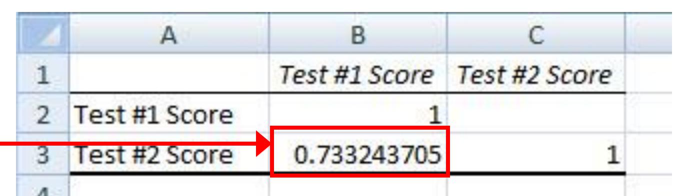


Using Excel to Find the Correlation Coefficient

(continued)

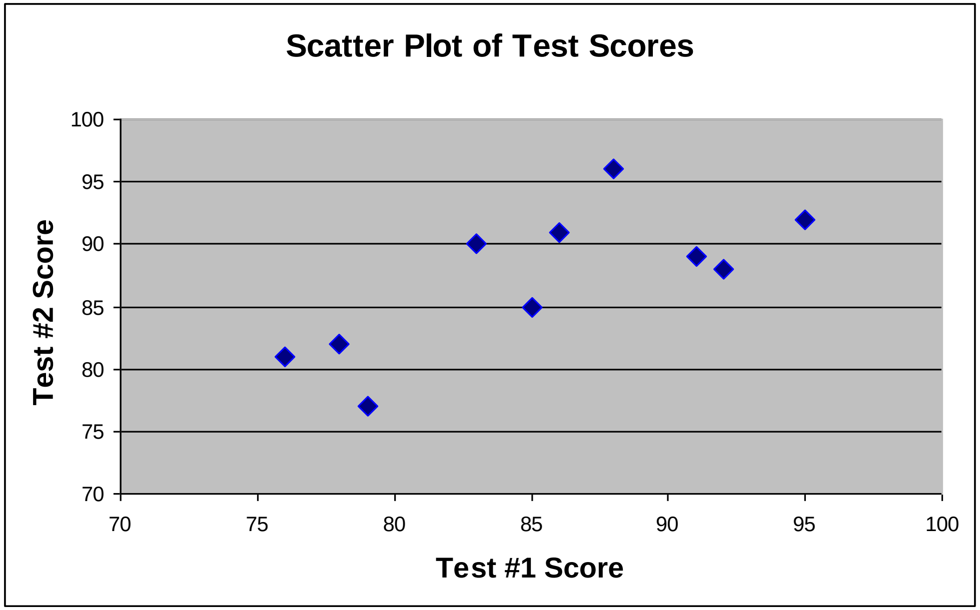


- Input data range and select appropriate options
- Click OK to get output



Interpreting the Result

- r = .733
- There is a relatively strong positive linear relationship between test score #1 and test score #2



- Students who scored high on the first test tended to score high on second test

Chapter Summary

- Described measures of central tendency
  - Mean, median, mode
- Illustrated the shape of the distribution
  - Symmetric, skewed
- Described measures of variation
  - Range, interquartile range, variance and standard deviation, coefficient of variation
- Discussed measures of grouped data
- Calculated measures of relationships between variables
  - covariance and correlation coefficient