A Course In Business Statistics
4th Edition

Chapter 3
Describing Data Using Numerical Measures

Chapter Goals

After completing this chapter, you should be able to:

- Compute and interpret the mean, median, and mode for a set of data
- Compute the range, variance, and standard deviation and know what these values mean
- Construct and interpret a box and whiskers plot
- Compute and explain the coefficient of variation and z scores
- Use numerical measures along with graphs, charts, and tables to describe data

Chapter Topics

- Measures of Center and Location
  - Mean, median, mode, geometric mean, midrange
- Other measures of Location
  - Weighted mean, percentiles, quartiles
- Measures of Variation
  - Range, interquartile range, variance and standard deviation, coefficient of variation

Summary Measures

- **Describing Data Numerically**
  - **Center and Location**
    - Mean
    - Median
    - Mode
    - Weighted Mean
  - **Other Measures of Location**
    - Percentiles
    - Quartiles
  - **Variation**
    - Range
    - Interquartile Range
    - Variance
    - Standard Deviation
    - Coefficient of Variation

Measures of Center and Location

Overview

Center and Location

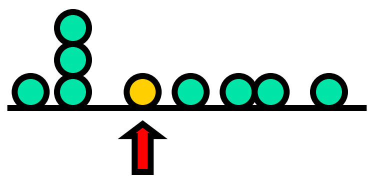
- **Mean**
- **Median**
- **Mode**
- **Weighted Mean**

Mean

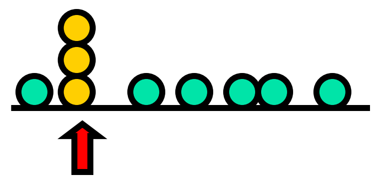
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Median



Mode



Weighted Mean

$$\bar{X}_W = \frac{\sum w_i x_i}{\sum w_i}$$

$$\mu_W = \frac{\sum w_i x_i}{\sum w_i}$$

Mean (Arithmetic Average)

- The Mean is the arithmetic average of data values
  - Sample mean

n = Sample Size

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

  - Population mean

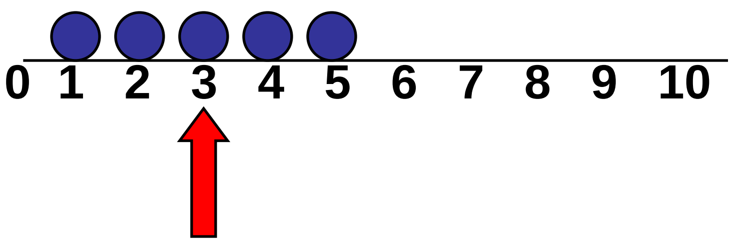
N = Population Size

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

Mean (Arithmetic Average)

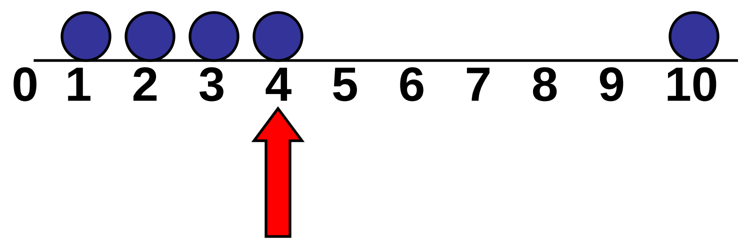
(continued)

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



Mean = 3

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

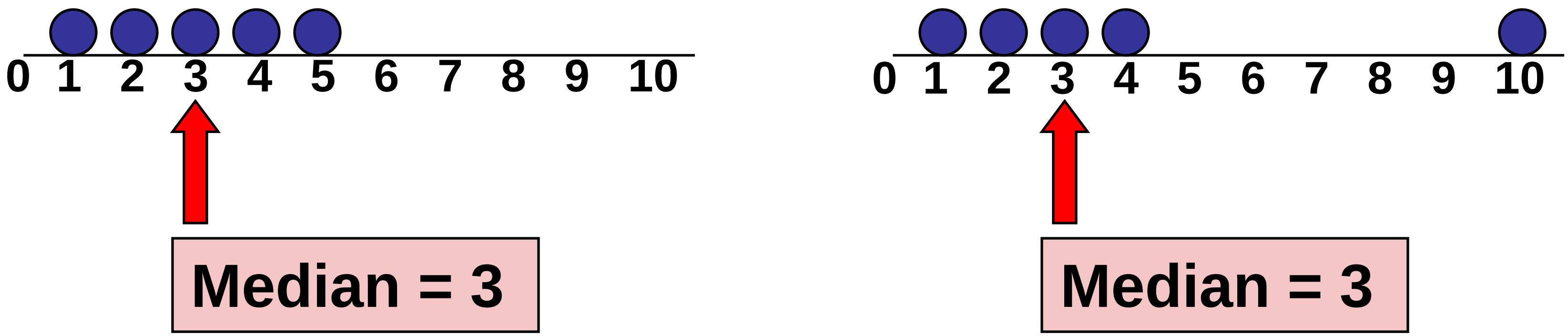


Mean = 4

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

Median

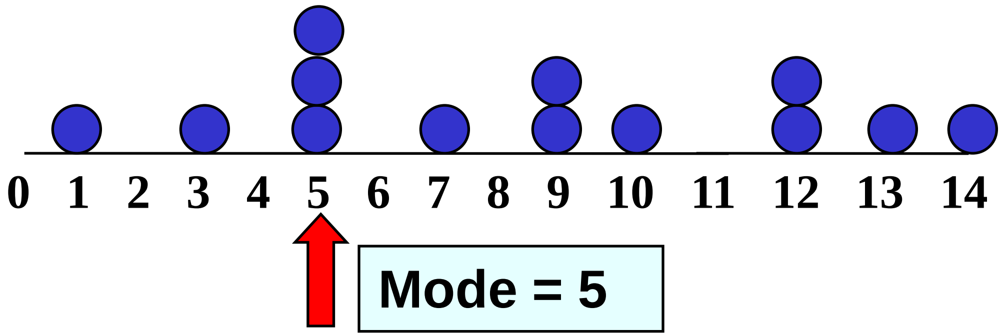
- Not affected by extreme values

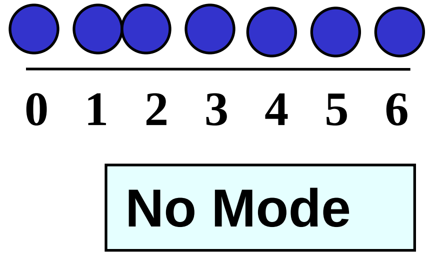


- In an ordered array, the median is the “middle” number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the two middle numbers

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





Weighted Mean

- Used when values are grouped by frequency or relative importance

Example: Sample of 26 Repair Projects

Days to Complete	Frequency
5	4
6	12
7	8
8	2

Weighted Mean Days to Complete:

$$\overline{X}_W = \frac{\sum w_i x_i}{\sum w_i} = \frac{(4\times 5)+(12\times 6)+(8\times 7)+(2\times 8)}{4+12+8+2}$$

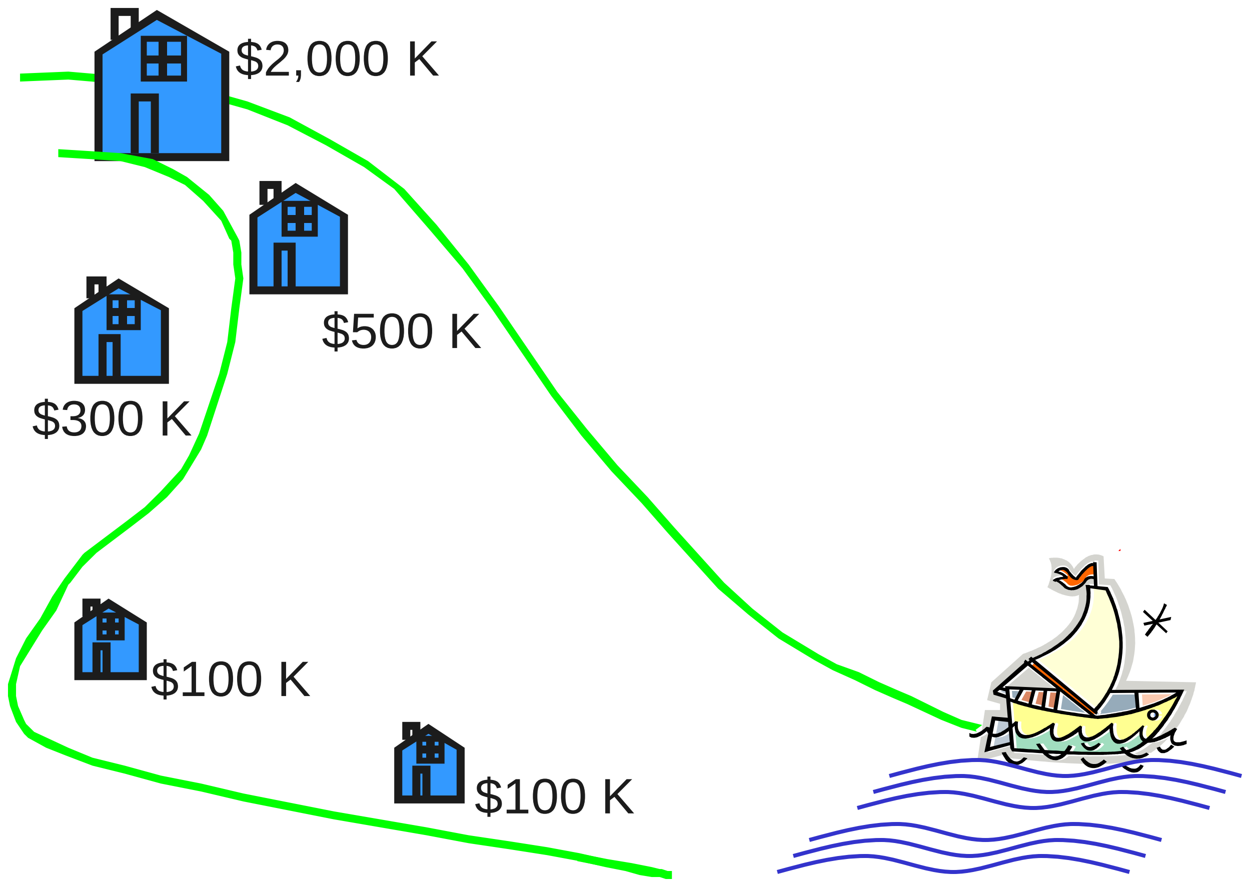
$$= \frac{164}{26} = 6.31 \text{ days}$$

Review Example

- Five houses on a hill by the beach

House Prices:

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



Summary Statistics

House Prices:

$2,000,000
500,000
300,000
100,000
100,000
Sum **3,000,000**

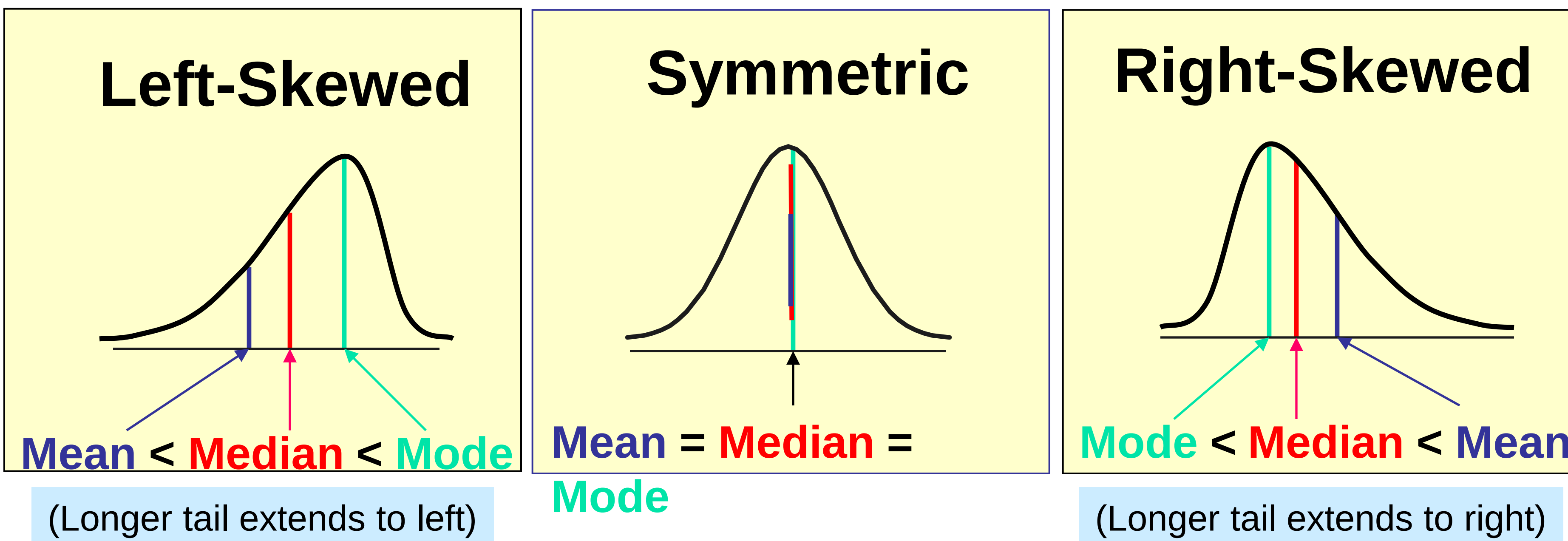
- **Mean:** ($3,000,000/5)
  = **$600,000**
- **Median:** middle value of ranked data
  = **$300,000**
- **Mode:** most frequent value
  = **$100,000**

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 is distributed
- Symmetric or skewed



Other Location Measures

Other Measures of Location

Percentiles | **Quartiles**

The p^{th} percentile in a data array:

- p% are less than or equal to this value
- (100 – p)% are greater than or equal to this value

(where $0 \le p \le 100$)

- 1^{st} quartile = 25^{th} percentile
- 2^{nd} quartile = 50^{th} percentile = median
- 3^{rd} quartile = 75^{th} percentile

Percentiles

- The p^{th} percentile in an ordered array of n values is the value in i^{th} position, where

$$i = \frac{p}{100}(n+1)$$

- Example: The 60^{th} percentile in an ordered array of 19 values is the value in 12^{th} position:

$$i = \frac{p}{100}(n+1) = \frac{60}{100}(19+1) = 12$$

Quartiles

- Quartiles split the ranked data into 4 equal groups

25%	**25%**	**25%**	**25%**

Q1 Q2 Q3

- Example: Find the first quartile

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q1 = 25th percentile, so find the $\frac{25}{100}(9+1) = 2.5$ position

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

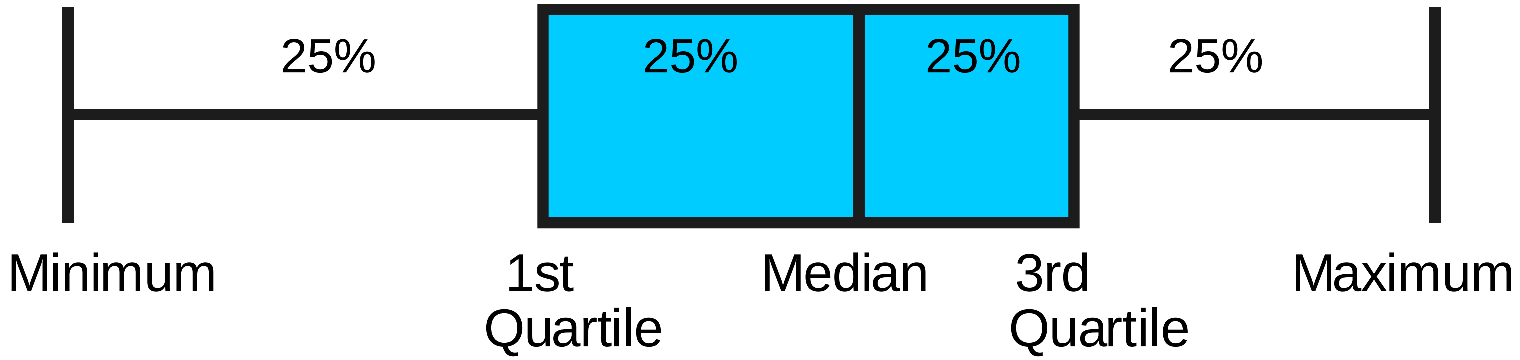
so **Q1 = 12.5**

Box and Whisker Plot

- A Graphical display of data using 5-number summary:

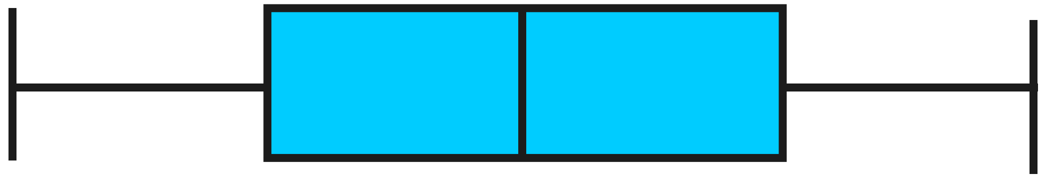
Minimum -- Q1 -- Median -- Q3 -- Maximum

Example:



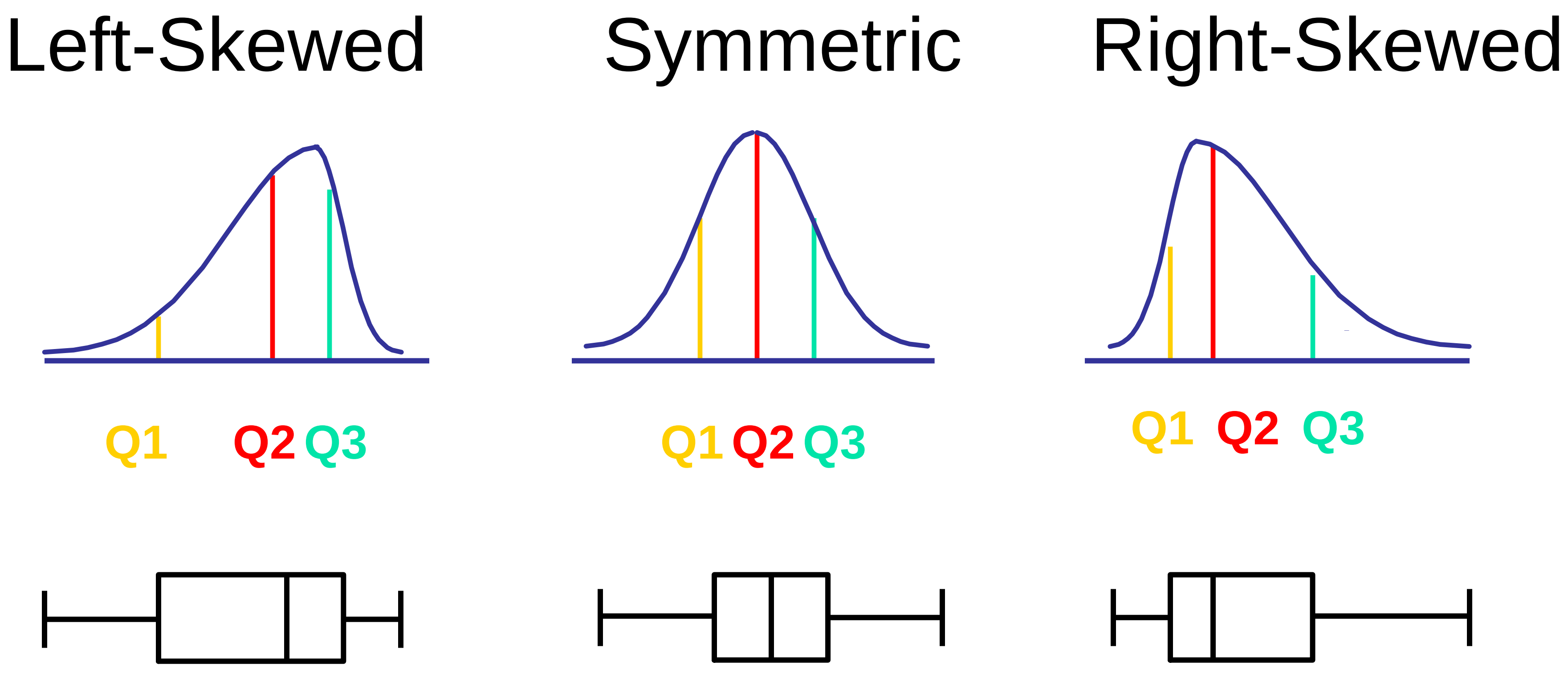
Shape of Box and Whisker Plots

- The Box and central line are centered between the endpoints if data is symmetric around the median



- A Box and Whisker plot can be shown in either vertical or horizontal format

Distribution Shape and Box and Whisker Plot



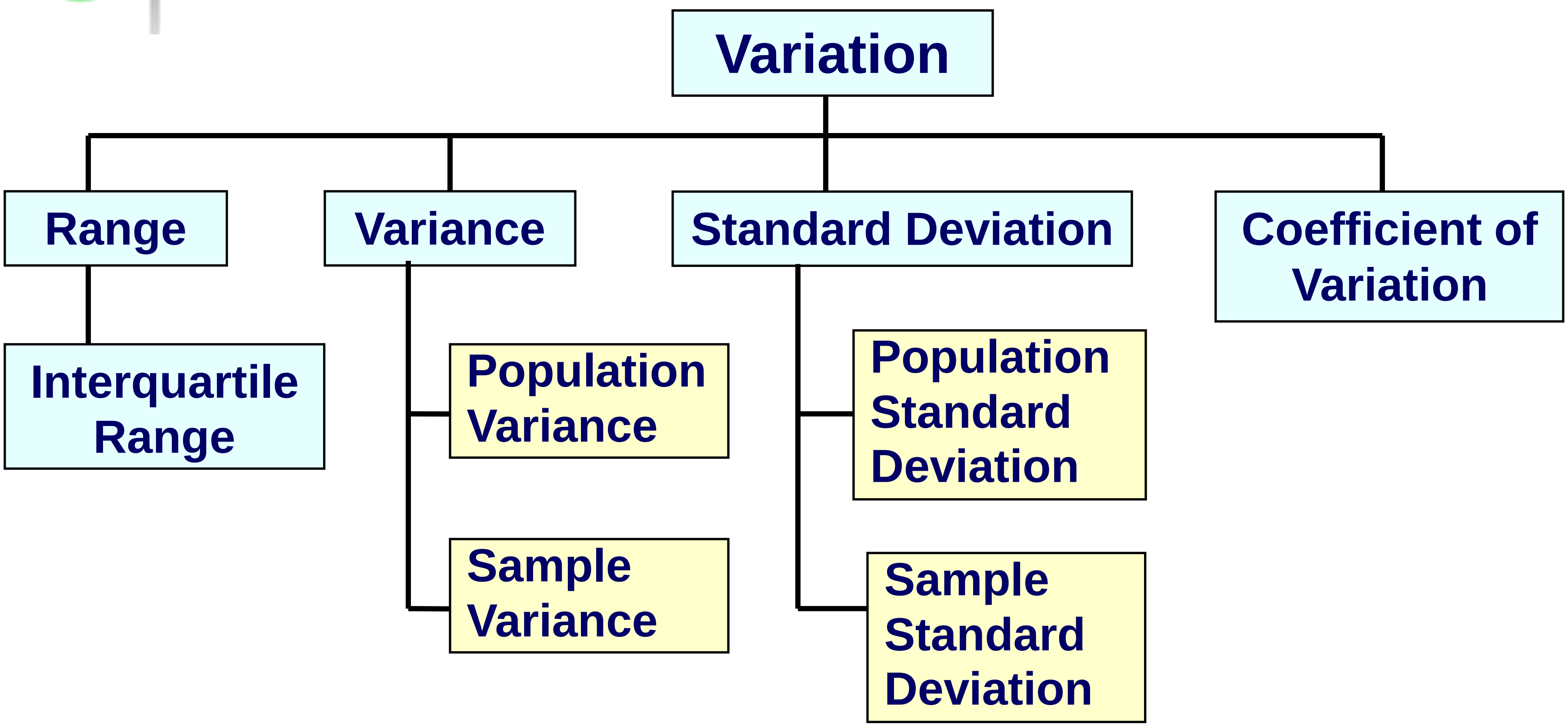
Box-and-Whisker Plot Example

- Below is a Box-and-Whisker plot for the following data:

Min		Q1			Q2			Q3		Max
0	2	2	2	3	3	4	5	5	10	27

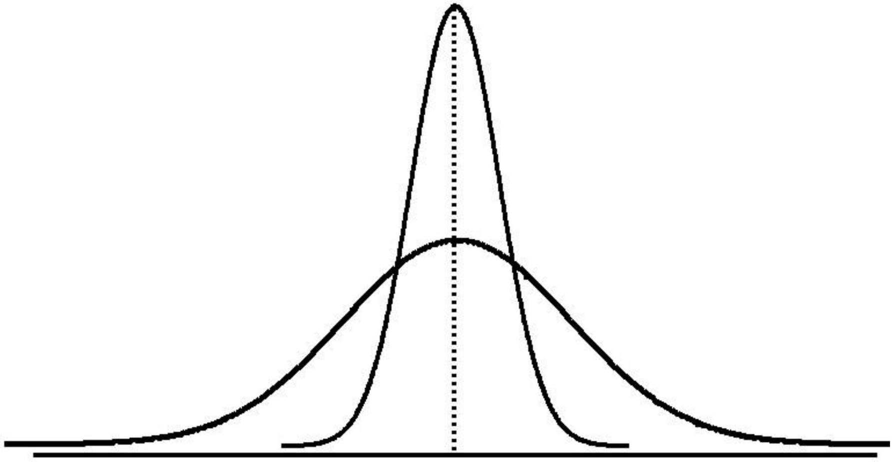
- This data is very right skewed, as the plot depicts

Measures of Variation



Variation

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



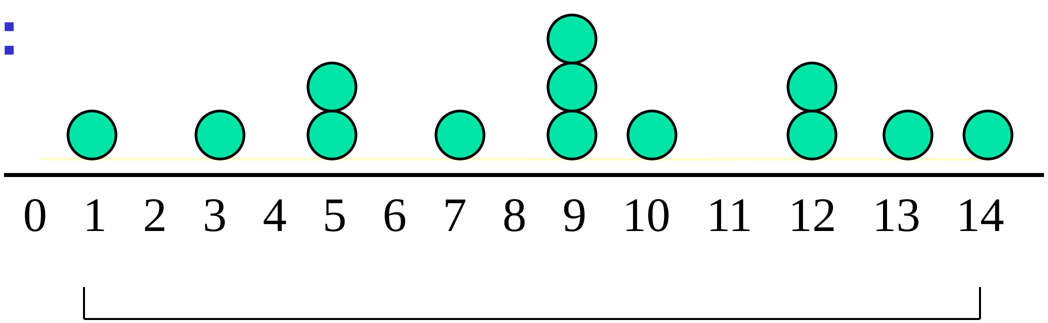
Same center, different variation

Range

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

$$\text{Range} = x_{maximum} - x_{minimum}$$

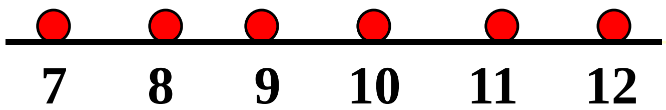
Example:



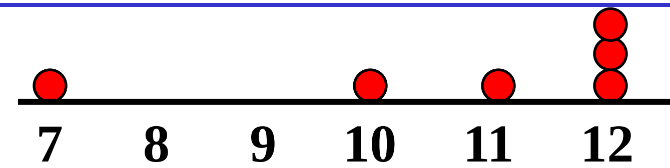
Range = 14 - 1 = 13

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

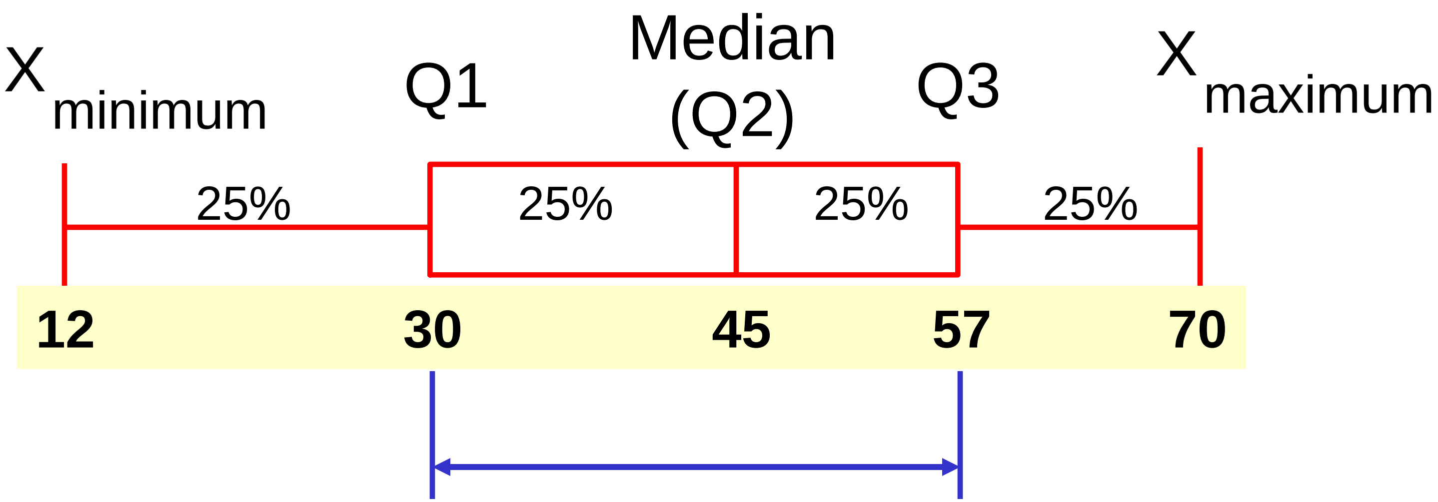
Range = 120 - 1 = 119

Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high-and low-valued observations and calculate the range from the remaining values.
- Interquartile range = 3rd quartile – 1st quartile

Interquartile Range

Example:



Interquartile range
= 57 – 30 = 27

Variance

- Average of squared deviations of values from the mean
  - **Sample** variance:

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

  - **Population** variance:

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

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data
  - **Sample** standard deviation:

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

  - **Population** standard deviation:

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

Calculation Example: Sample Standard Deviation

Sample Data (X_i) :

10	12	14	15	17	18	18	24

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

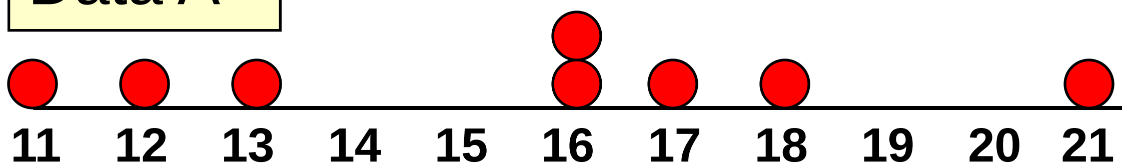
$$s = \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

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

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

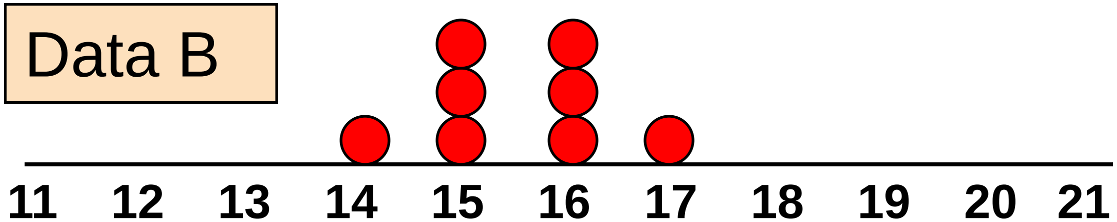
Comparing Standard Deviations

Data A



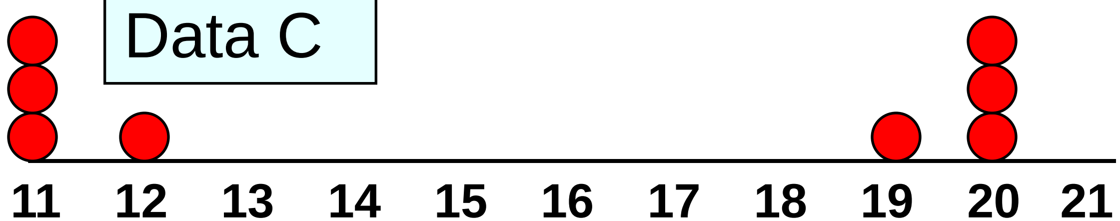
Mean = 15.5
s = 3.338

Data B



Mean = 15.5
s = .9258

Data C



Mean = 15.5
s = 4.57

Coefficient of Variation

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

Population

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

Sample

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

Comparing Coefficient of Variation

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

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

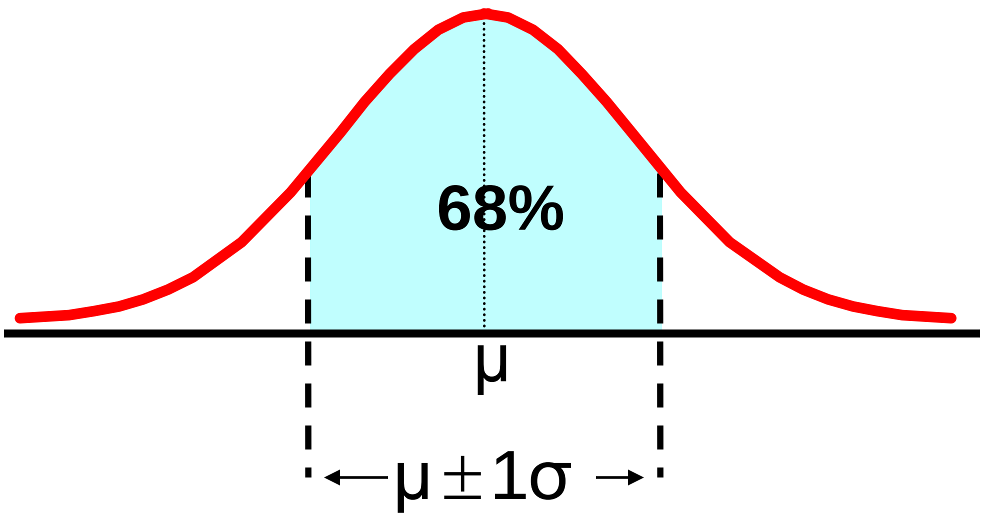
- Stock B:
  - Average price last year = \$100
  - Standard deviation = \$5

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

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

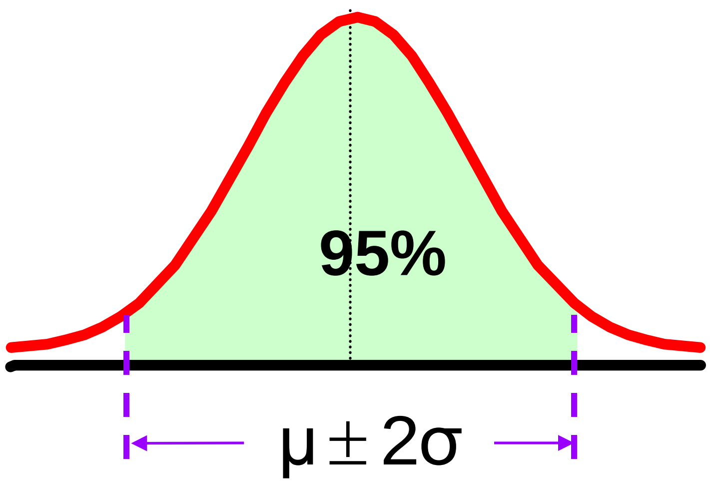
The Empirical Rule

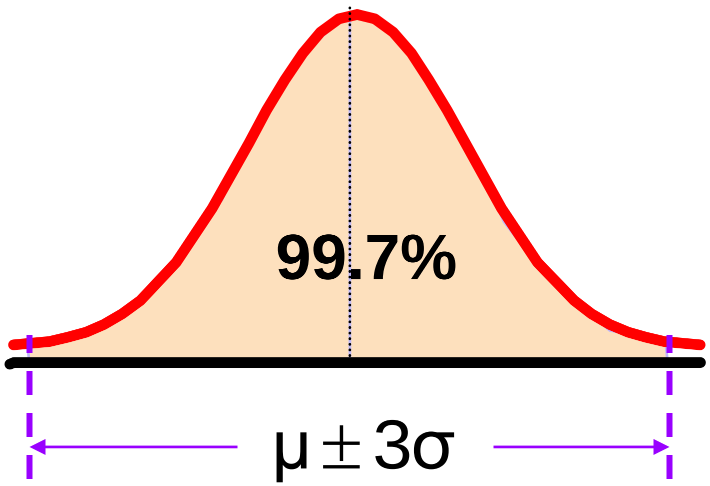
- If the data distribution is bell-shaped, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample



The Empirical Rule

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





Tchebysheff's Theorem

- 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

  - Examples:

At least		within
$(1 - 1/1^2) =$ 0%	....	$k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) =$ 75%	....	$k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) =$ 89%	....	$k=3$ $(\mu \pm 3\sigma)$

Standardized Data Values

- A standardized data value refers to the number of standard deviations a value is from the mean

- Standardized data values are sometimes referred to as z-scores

Standardized Population Values

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

where:

- x = original data value
- μ = population mean
- σ = population standard deviation
- z = standard score

  (number of standard deviations x is from μ)

Standardized Sample Values

$$z = \frac{x - \bar{x}}{s}$$

where:

- x = original data value
- $\bar{x}$ = sample mean
- s = sample standard deviation
- z = standard score

(number of standard deviations x is from μ)

Using Microsoft Excel

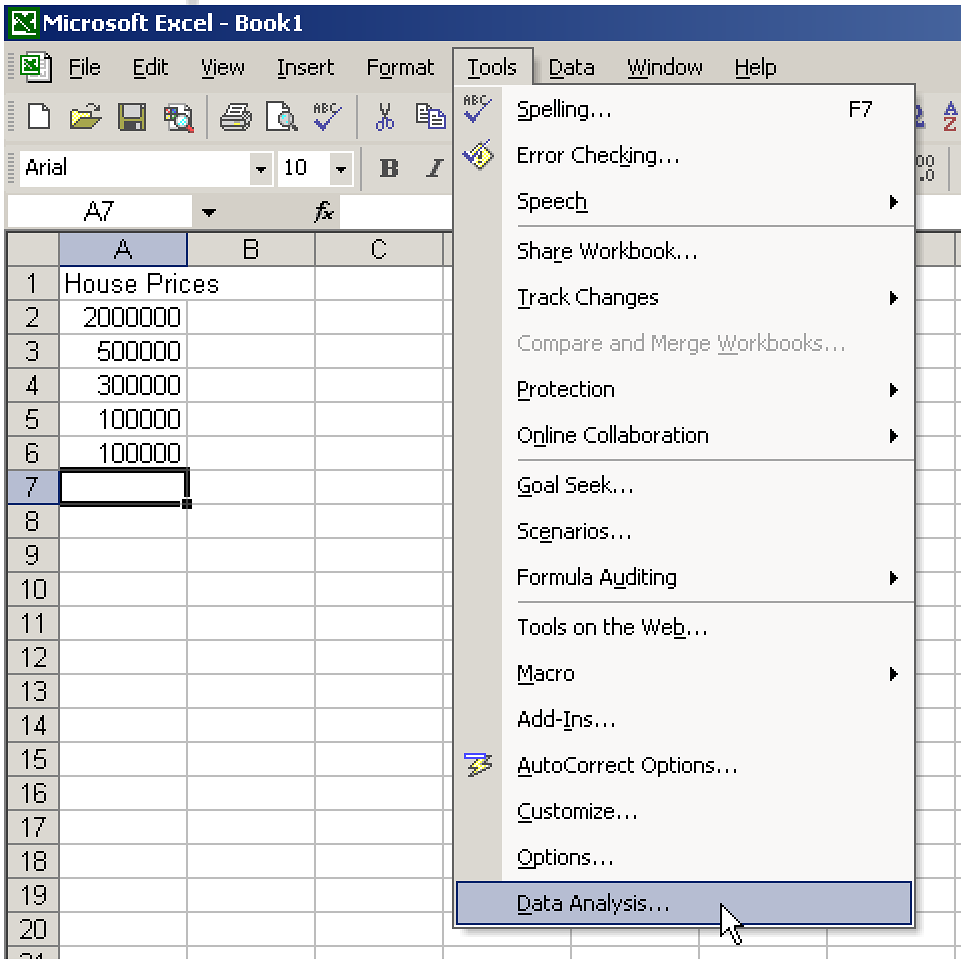
- Descriptive Statistics are easy to obtain from Microsoft Excel
  - Use menu choice:

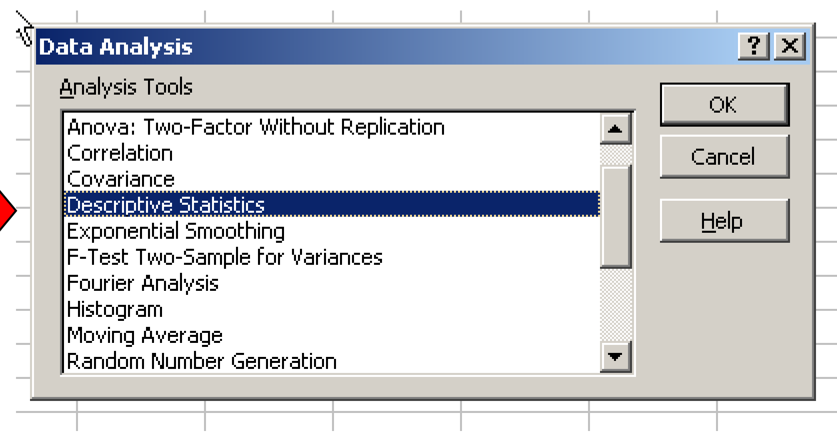
    tools / data analysis / descriptive statistics
  - Enter details in dialog box

Using Excel

- Use menu choice:

tools / data analysis / descriptive statistics

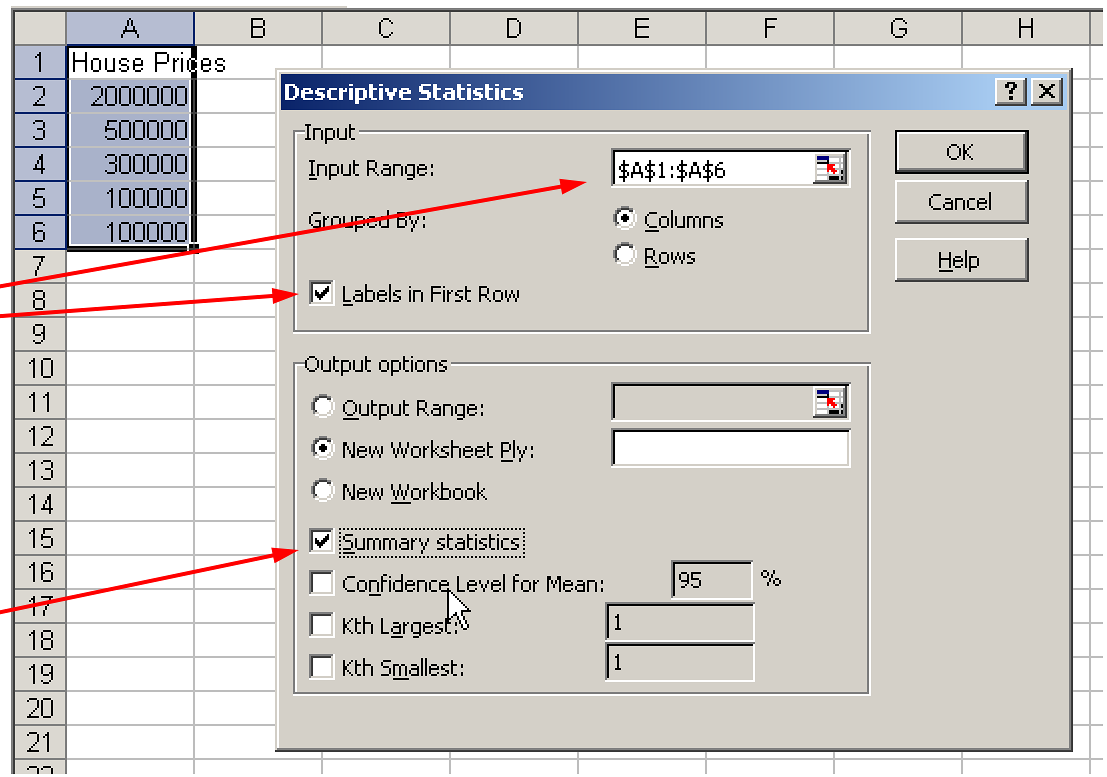




Using Excel

(continued)

- Enter dialog box 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		
17		

Chapter Summary

- Described measures of center and location
  - Mean, median, mode, geometric mean, midrange
- Discussed percentiles and quartiles
- Described measure of variation
  - Range, interquartile range, variance, standard deviation, coefficient of variation
- Created Box and Whisker Plots

Chapter Summary

(continued)

- Illustrated distribution shapes
  - Symmetric, skewed
- Discussed Tchebysheff's Theorem
- Calculated standardized data values