Statistics for Managers Using Microsoft® Excel 4th Edition

Chapter 3

Numerical Descriptive Measures

Chapter Goals

After completing this chapter, you should be able to:

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

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
- Five number summary and box-and-whisker plots
- Coefficient of correlation
- Ethical considerations in numerical descriptive measures

Summary Measures

- **Describing Data Numerically**
  - **Central Tendency**
    - Arithmetic Mean
    - Median
    - Mode
    - Geometric Mean
  - **Quartiles**
  - **Variation**
    - Range
    - Interquartile Range
    - Variance
    - Standard Deviation
    - Coefficient of Variation
  - **Shape**
    - Skewness

Measures of Central Tendency

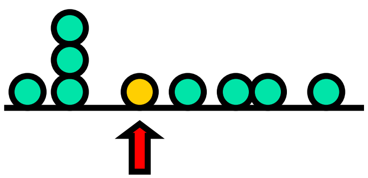
Overview

Central Tendency

- **Arithmetic Mean**

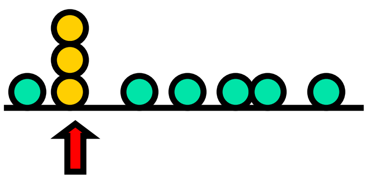
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Median**



Midpoint of ranked values

- **Mode**



Most frequently observed value

- **Geometric Mean**

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency

  - For a sample of size n:

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

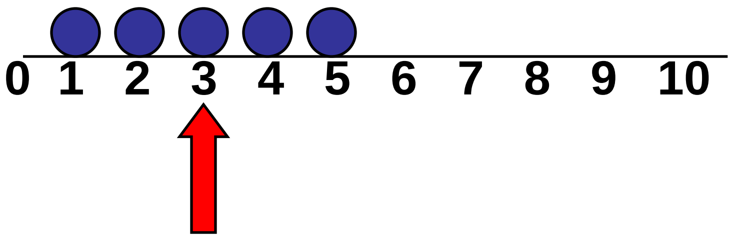
Sample size

Observed values

Arithmetic Mean

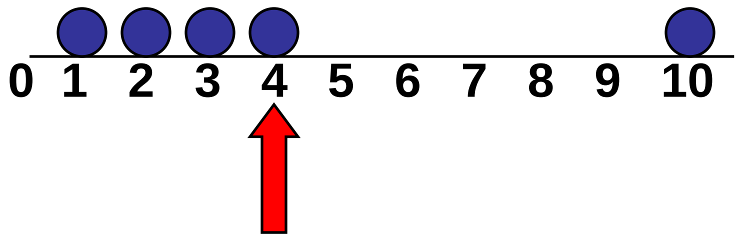
(continued)

- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



Mean = 3

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

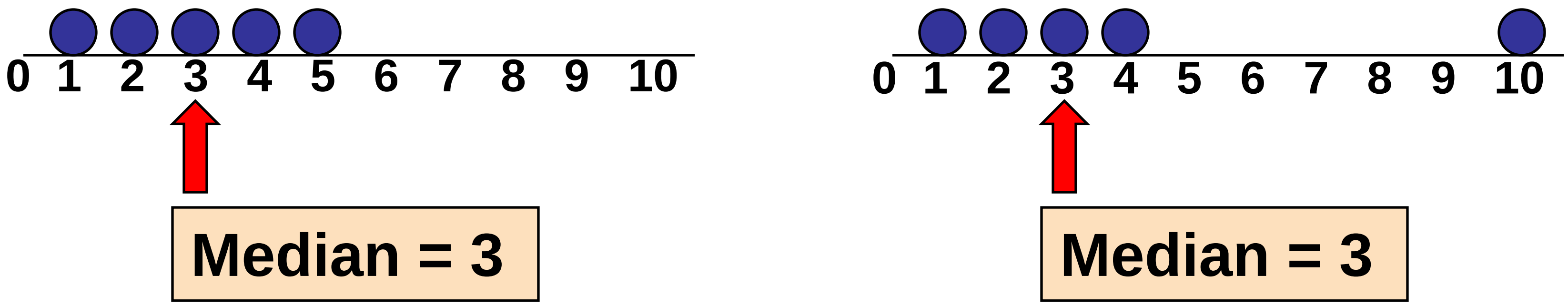


Mean = 4

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

Median

- Not affected by extreme values



- In an ordered array, the median is the "middle" number (50% above, 50% below)

Finding the Median

- The location of the median:

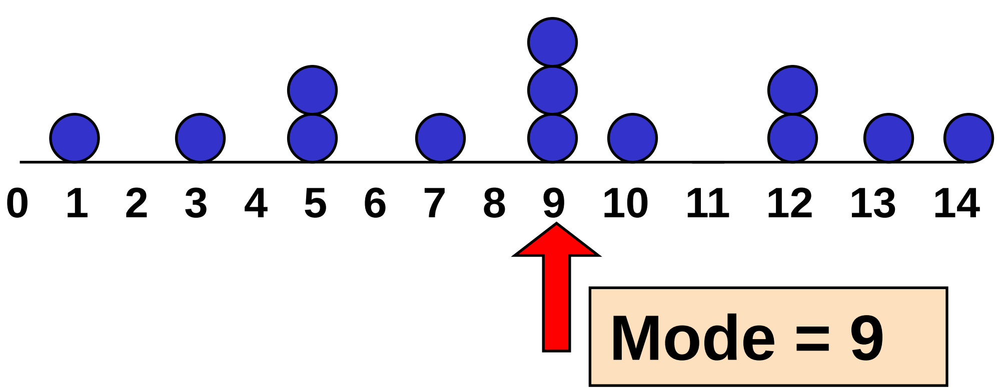
$$Median\ position = \frac{n+1}{2}\ position\ in\ the\ ordered\ array$$

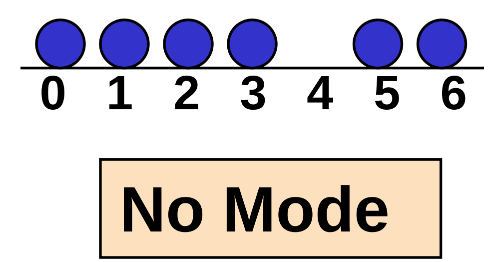
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- 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
- Mainly used for grouped numerical data or categorical data
- There may may be no mode
- There may be several modes



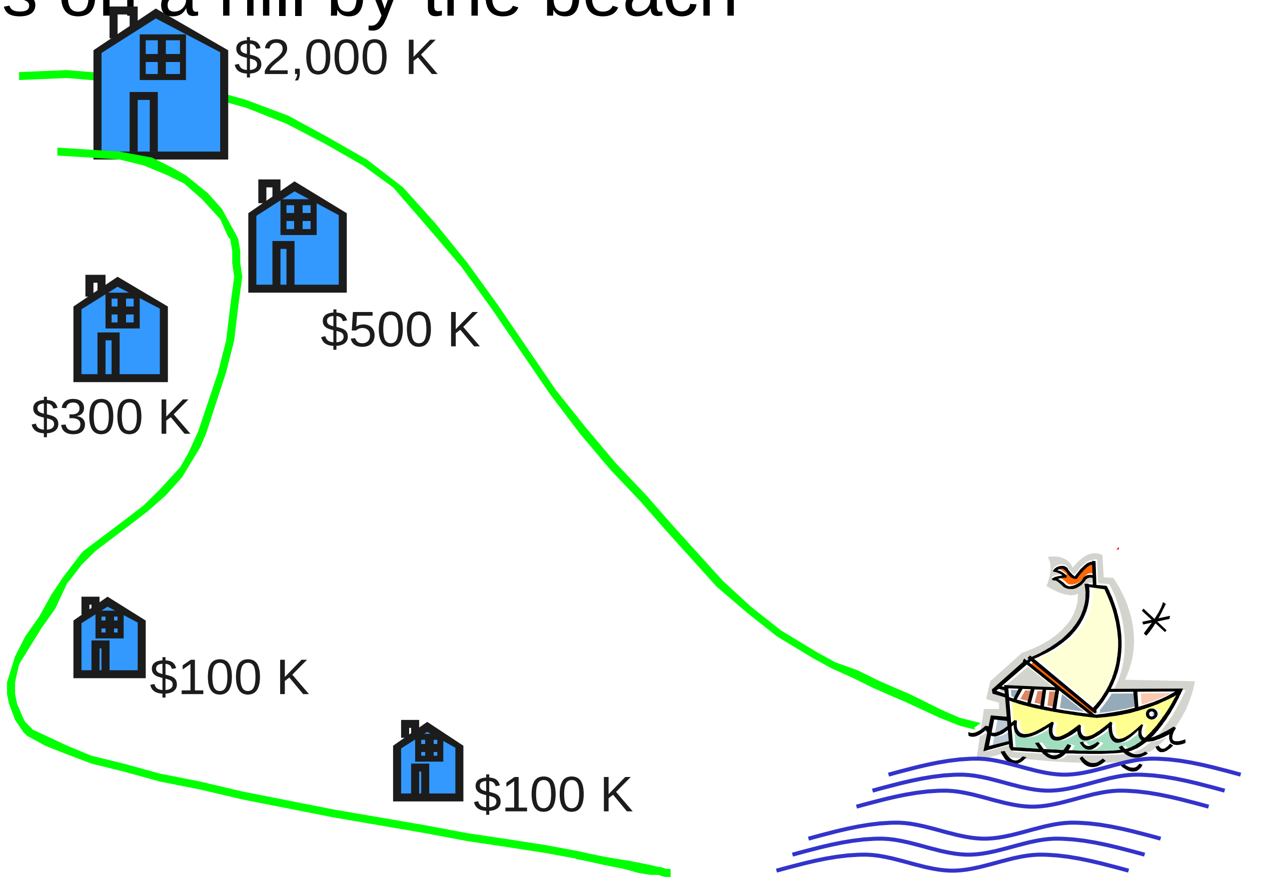


Review Example

- 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

Geometric Mean

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

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

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

$$\bar{R}_G = [(1+R_1) \times (1+R_2) \times \cdots \times (1+R_n)]^{1/n} - 1$$

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

Geometric Mean Example

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$$\underbrace{X_1 = \$100{,}000 \quad X_2 = \$50{,}000}_{\text{50\% decrease}} \qquad \underbrace{\quad X_3 = \$100{,}000}_{\text{100\% increase}}$$

The overall two-year return is zero, since it started and ended at the same level.

Geometric Mean Example

(continued)

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

Arithmetic mean rate of return:

$$\bar{X} = \frac{(-50\%) + (100\%)}{2} = 25\%$$

Misleading result

Geometric mean rate of return:

$$\begin{aligned}\bar{R}_G &= [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1 \\ &= [(1+(-50\%))\times(1+(100\%))]^{1/2} - 1 \\ &= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More accurate result

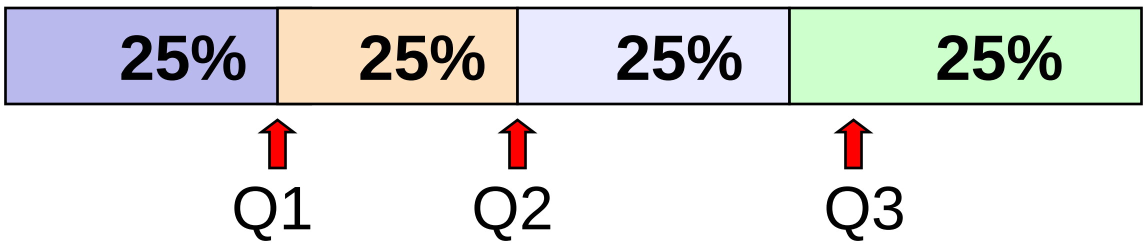
Geometric Mean Example

An investment of $100,000 declined to $50,000 at the end of year one and rebounded to $100,000 at end of year two:

1. **Returns as percents: –50% and 100% are converted to decimals -.5 and 1.00**
2. **Add 1 to each decimal yields .5 and 2**
3. **Find the geometric mean using the geomean function**
4. **Subtract 1 from the answer to get a rate of return of 0**

Quartiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- 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: $Q_1 = (n+1)/4$

Second quartile position: $Q_2 = (n+1)/2$ (the median position)

Third quartile position: $Q_3 = 3(n+1)/4$

where **n** is the number of observed values

Quartiles

- Example: Find the first quartile

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

(n = 9)

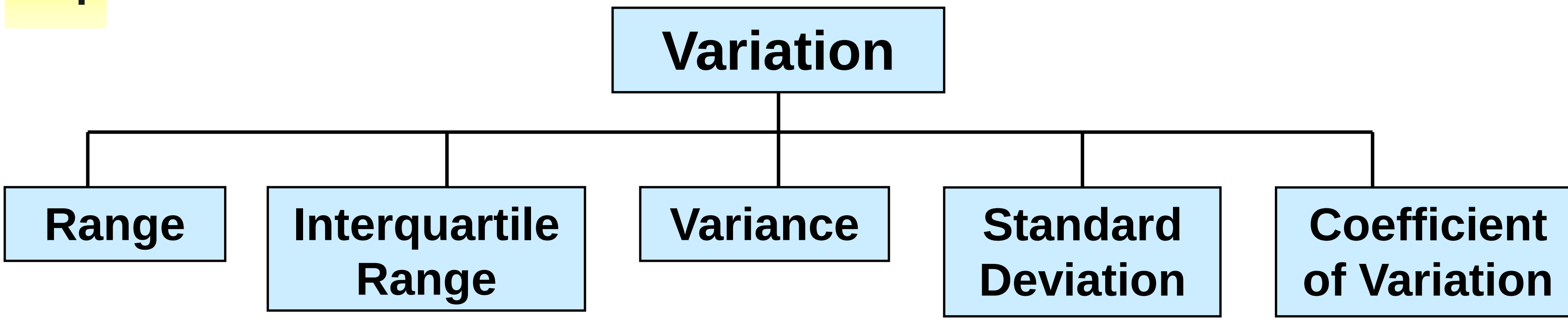
Q_1 = is in the (9+1)/4 = 2.5 position of the ranked data

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

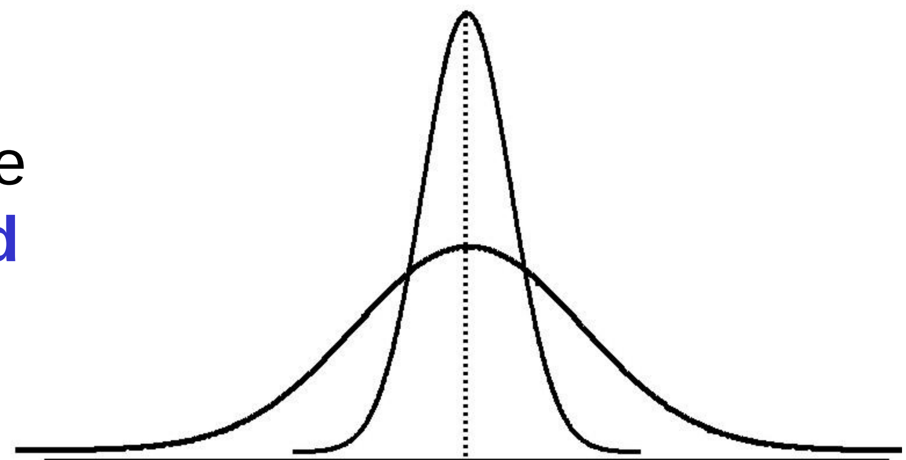
Q_1 = 12.5

Q_1 and Q_3 are measures of noncentral location
Q_2 = median, a measure of central tendency

Measures of Variation



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



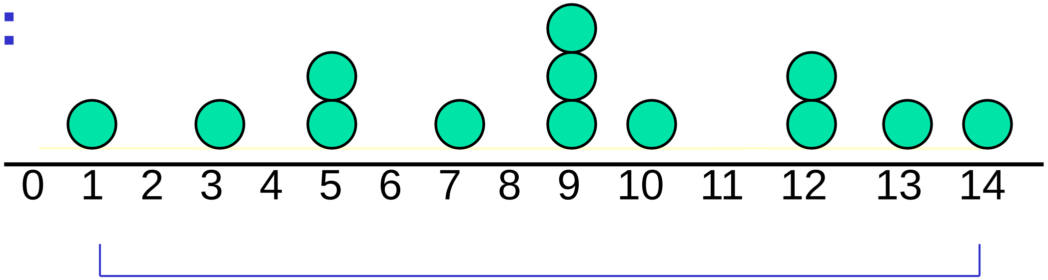
Same center, different variation

Range

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

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

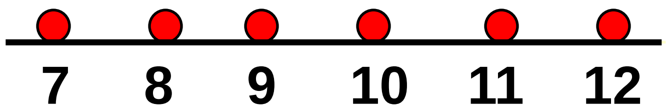
Example:



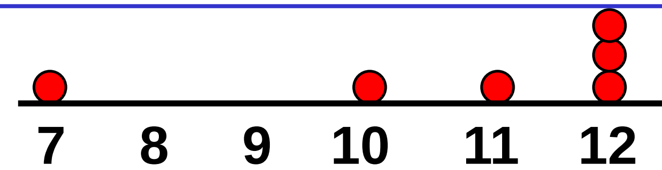
Range = 14 - 1 = 13

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

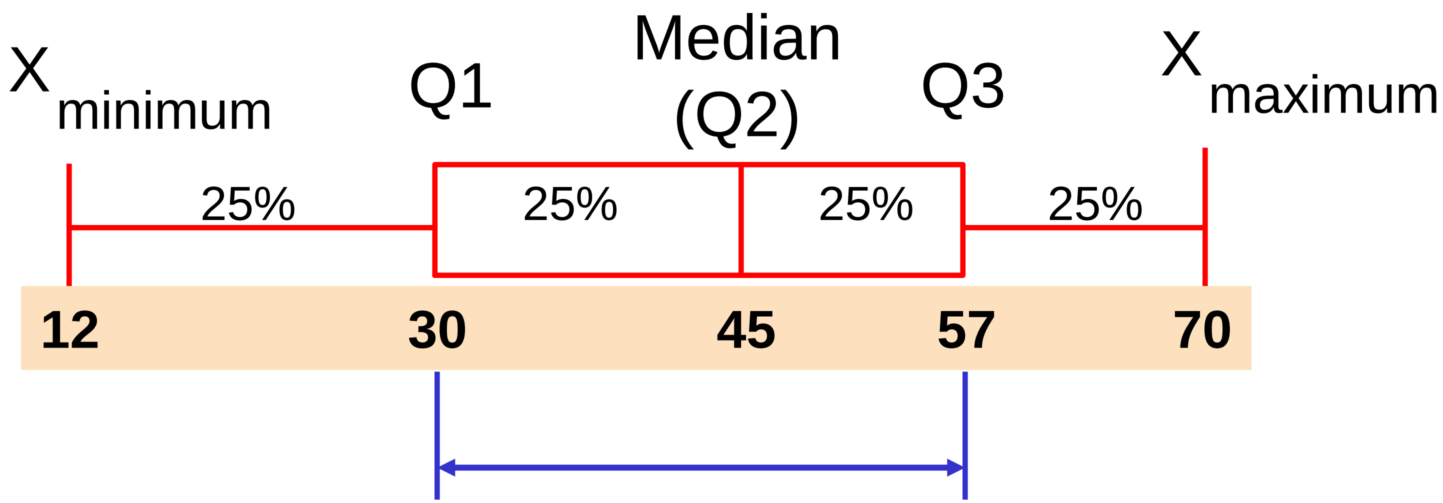
Range = 120 - 1 = 119

Interquartile Range

- You can eliminate some outlier problems by using the **interquartile range**
- Difference between the first and third quartiles
- Interquartile range = 3rd quartile – 1st quartile
  $= Q_3 – Q_1$

Interquartile Range

Example:



Interquartile range
= 57 – 30 = 27

Variance

- Average of squared deviations of each value from the mean
  - Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Where $\overline{X}$ = arithmetic mean

n = sample size

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

Standard Deviation

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

  - Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Calculation Example: Sample Standard Deviation

Sample Data (X_i) :

10	12	14	15	17	18	18	24

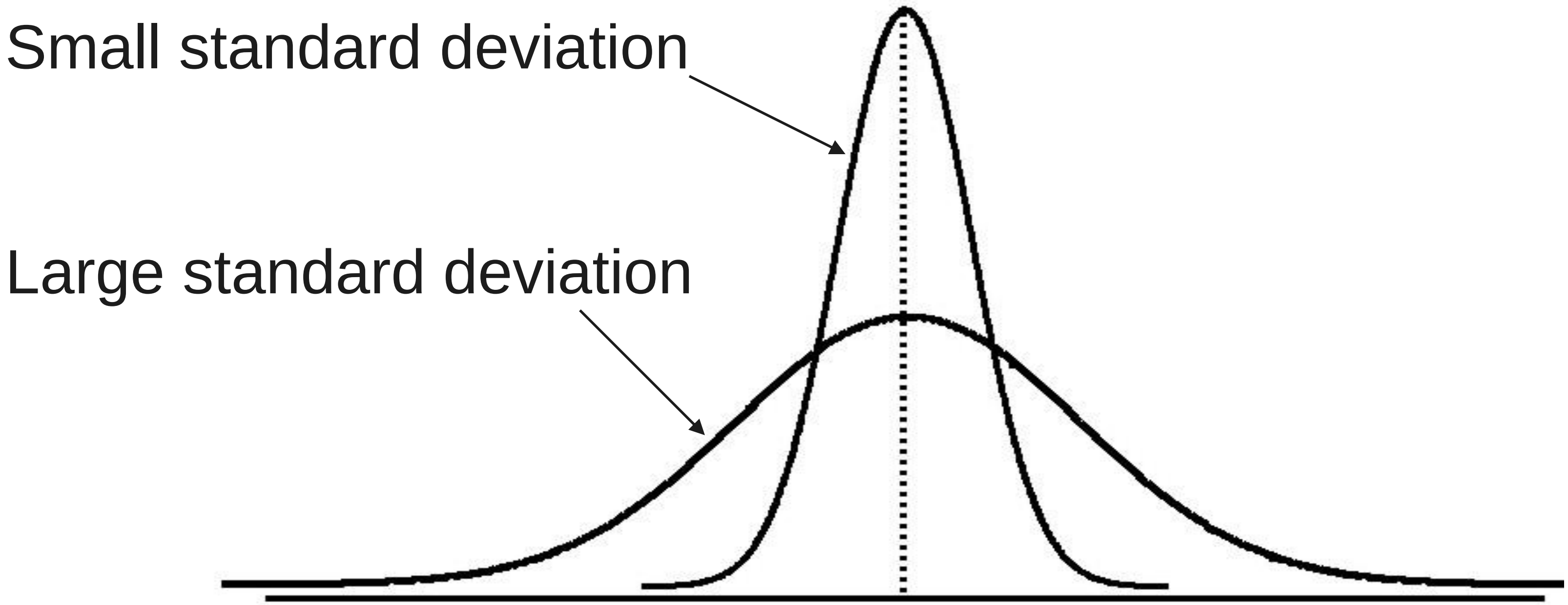
n = 8 Mean = $\overline{X}$ = 16

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

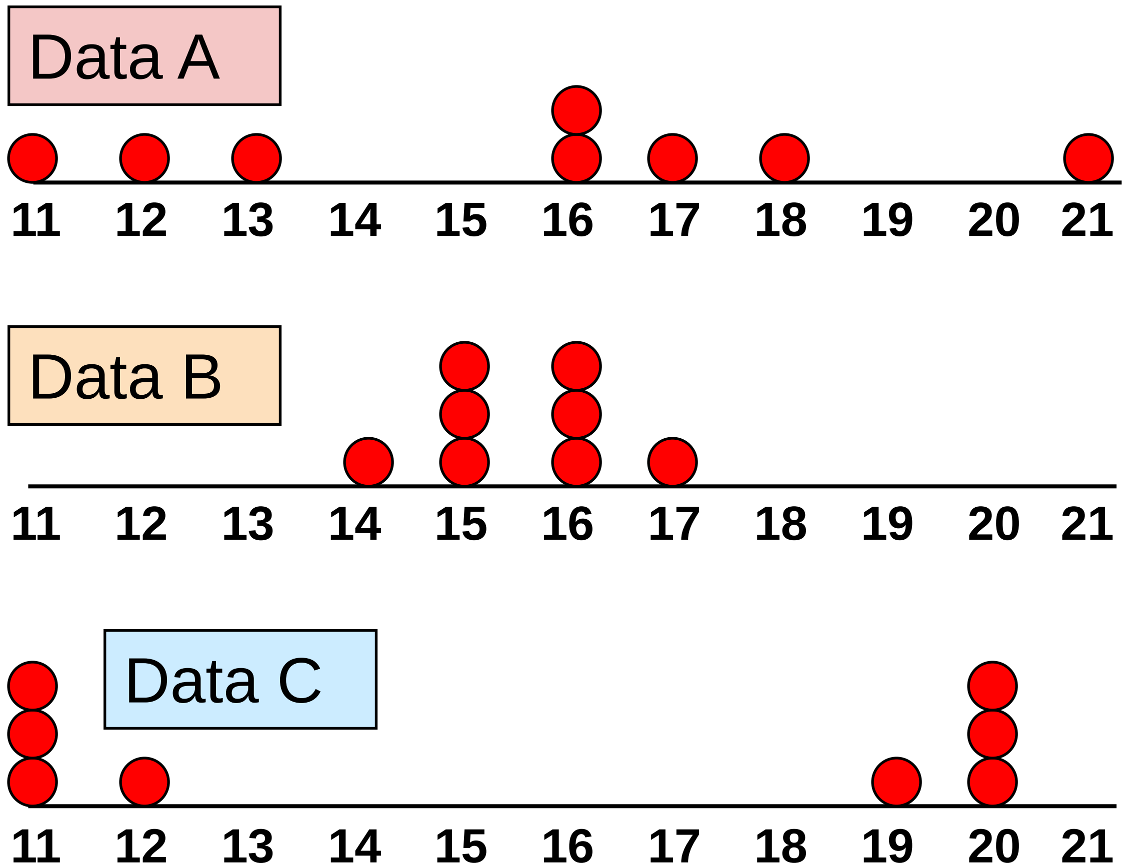
$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

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

Measuring variation



Comparing Standard Deviations



Mean = 15.5
S = 3.338

Mean = 15.5
S = .9258

Mean = 15.5
S = 4.57

Coefficient of Variation

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

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

Comparing Coefficients of Variation

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

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

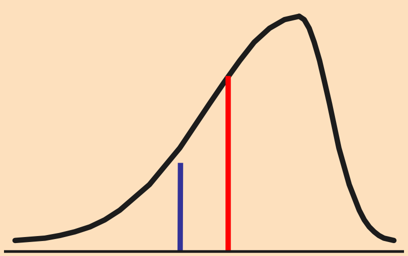
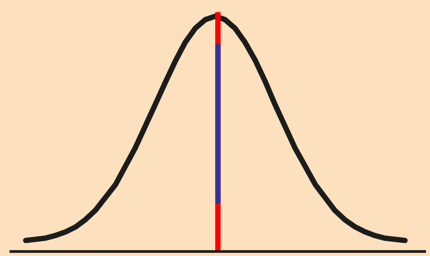
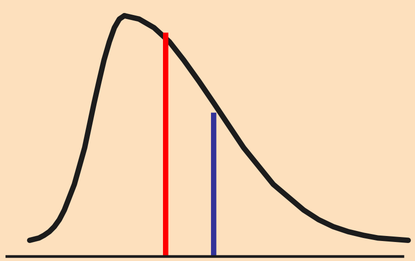
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

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

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

Shape of a Distribution

- Describes how data is distributed
- Shape - Symmetric or skewed

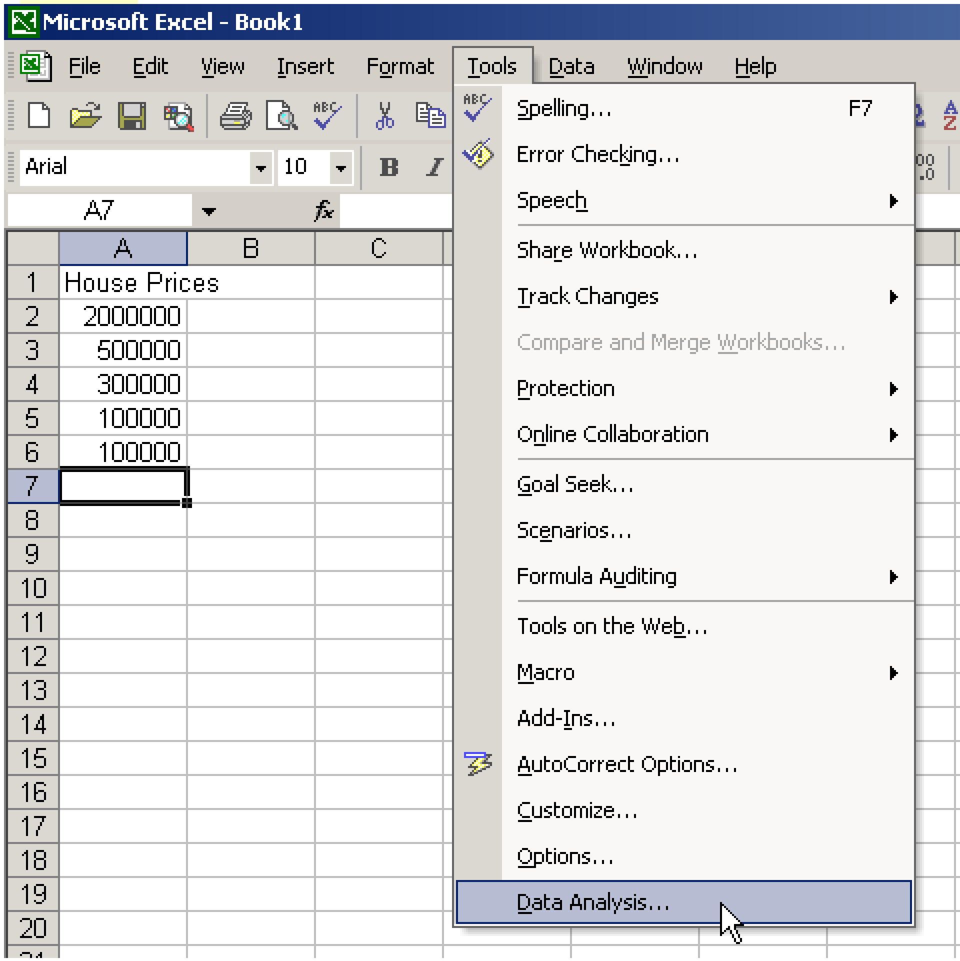
Left-Skewed	Symmetric	Right-Skewed
Mean < Median	Mean = Median	Median < Mean
		

Microsoft Excel

- Descriptive Statistics can be obtained from Microsoft® Excel
  - Use menu choice:

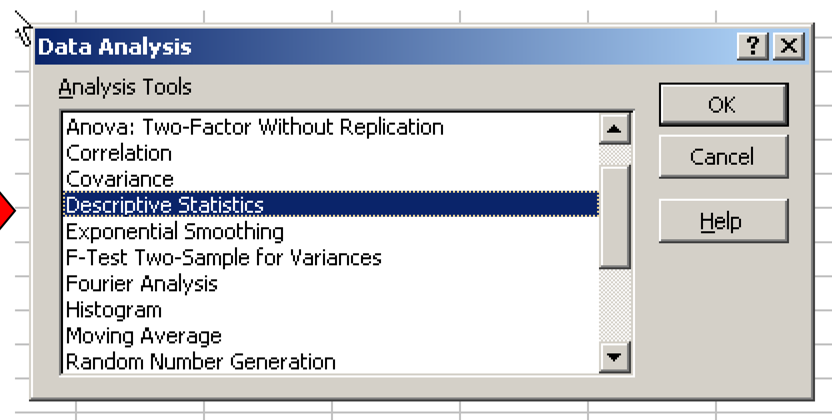
    tools / data analysis / descriptive statistics
  - Enter details in dialog box

Using Excel



- Use menu choice:

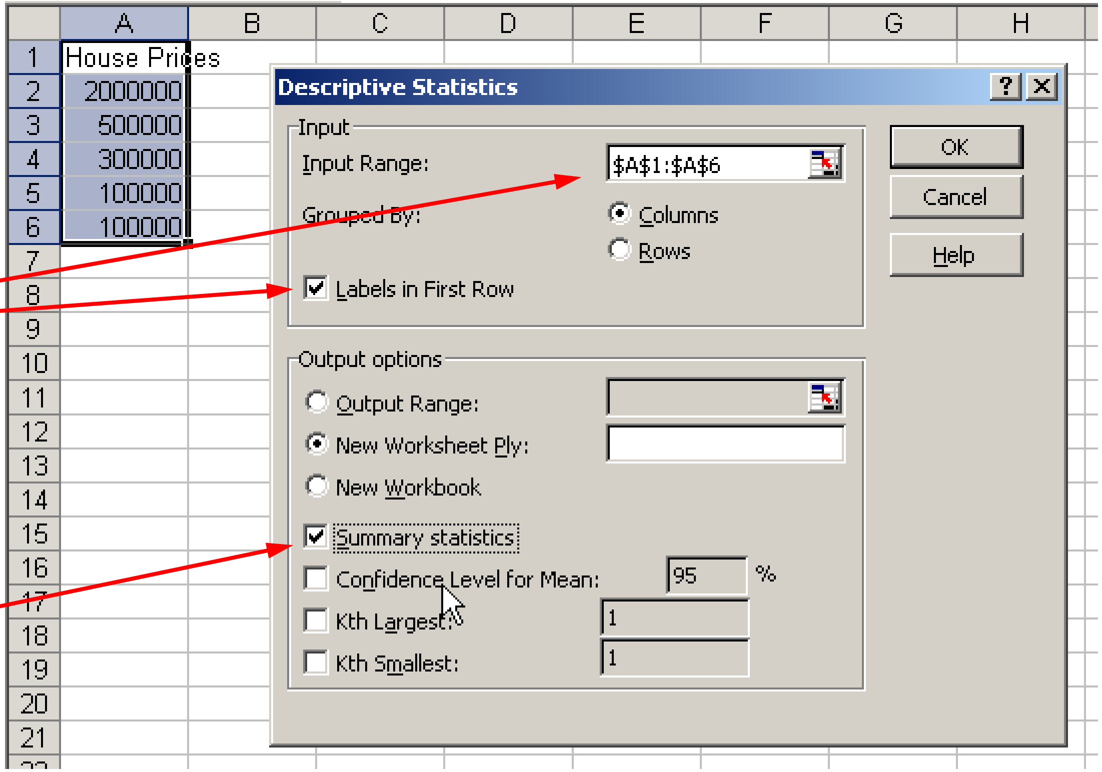
  tools / data analysis / descriptive statistics



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		

Population Summary Measures

- The population mean is the sum of the values in the population divided by the population size, N

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Where μ = population mean

N = population size

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

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

Population Standard Deviation

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

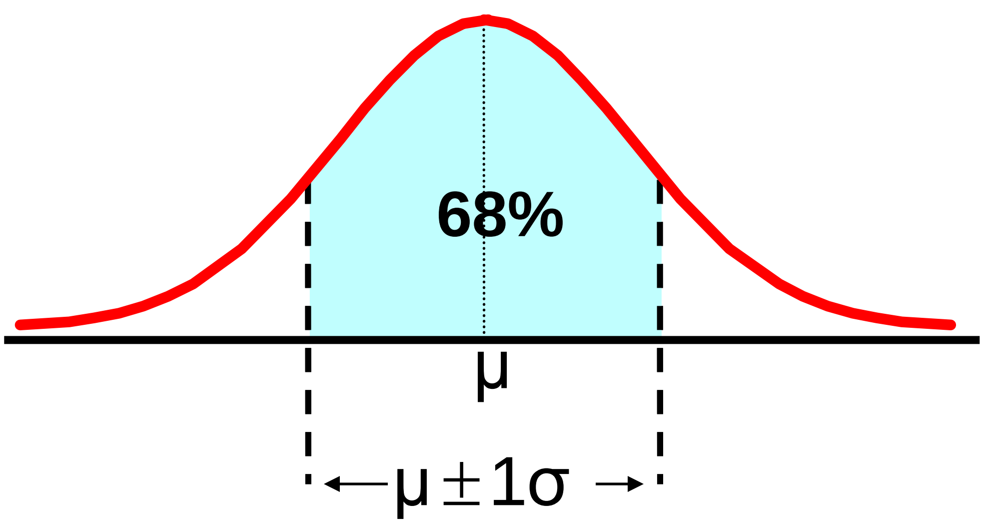
  - Population standard deviation:

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

The Empirical Rule

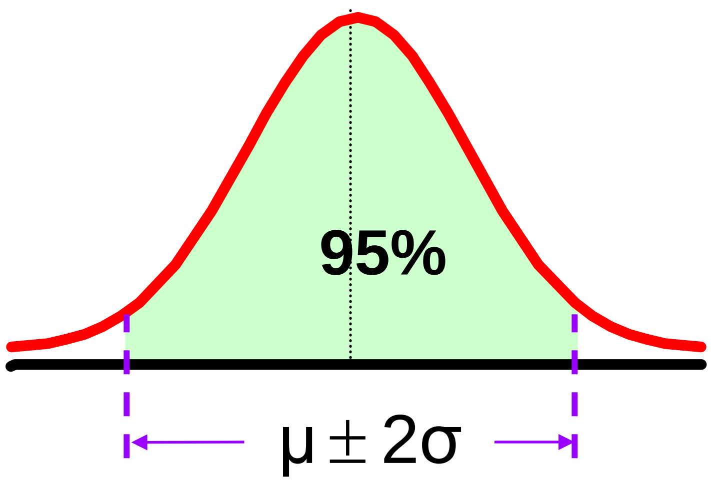
- If the data distribution is bell-shaped, then the interval: $\mu \pm 1\sigma$

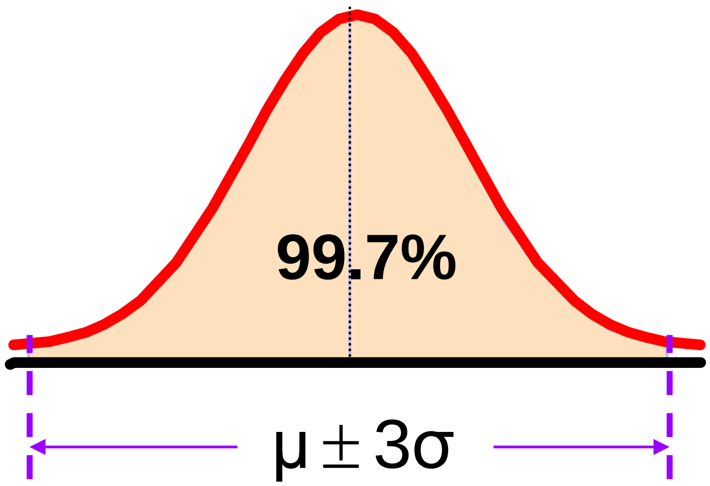
  contains about 68% of the values in the population or the sample



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



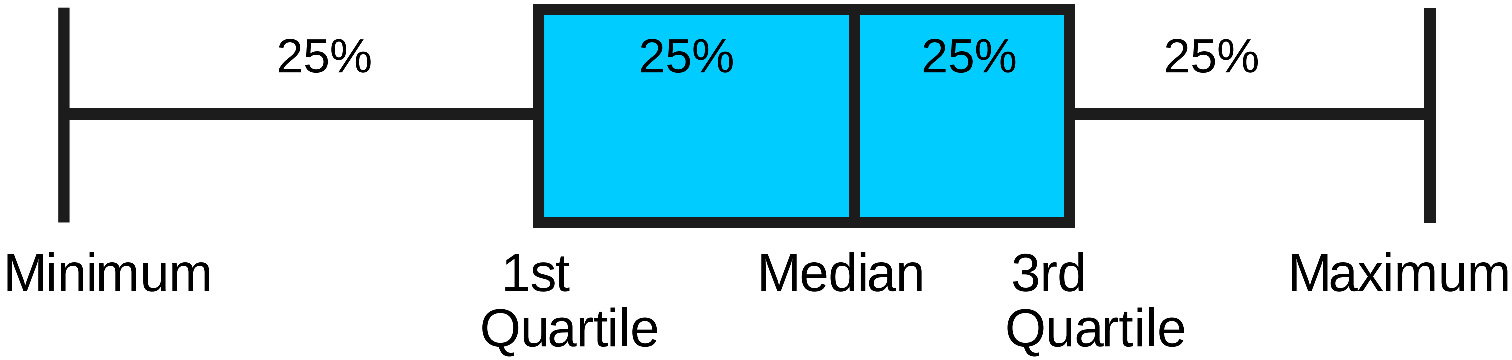


Exploratory Data Analysis

- Box-and-Whisker Plot: A Graphical display of data using 5-number summary:

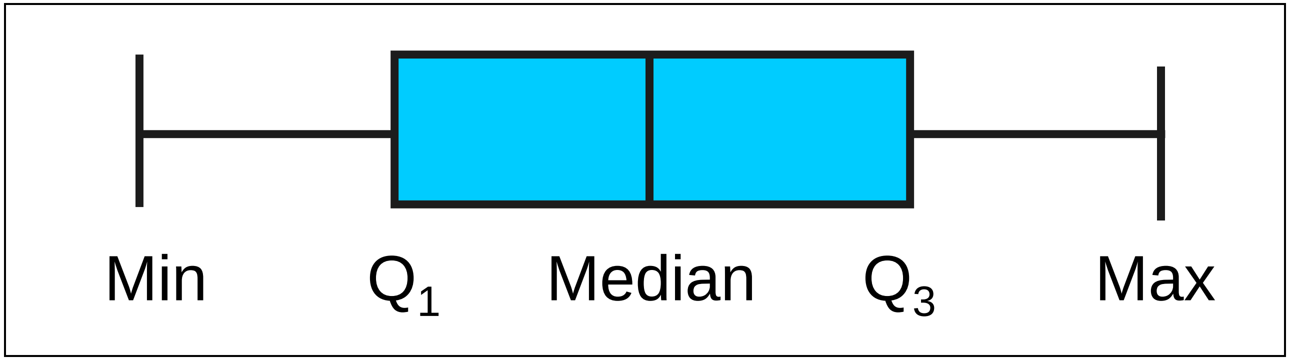
Minimum -- Q1 -- Median -- Q3 -- Maximum

Example:



Shape of Box and Whisker Plots

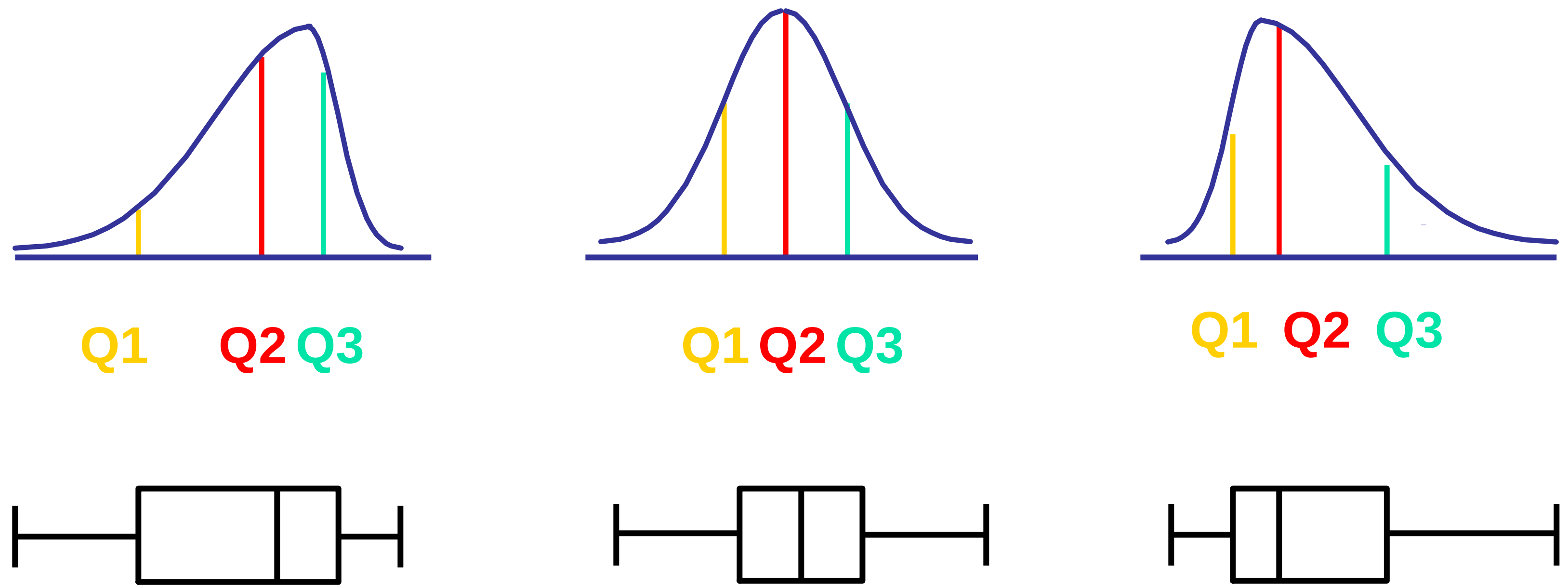
- The Box and central line are centered between the endpoints if data is symmetric around the median



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

Distribution Shape and Box and Whisker Plot

Left-Skewed | Symmetric | Right-Skewed



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 right skewed, as the plot depicts

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two variables

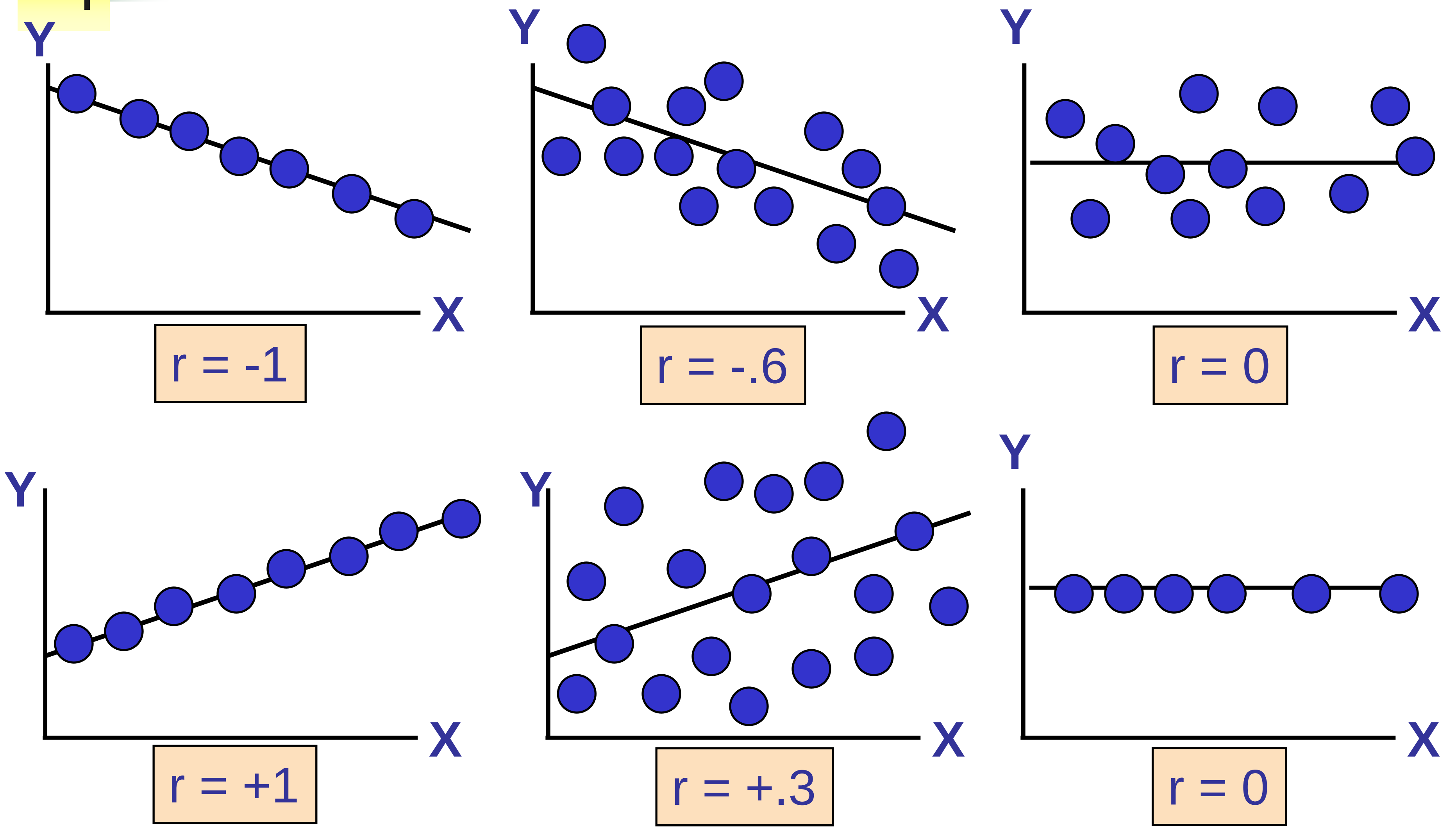
- Sample coefficient of correlation:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\ \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

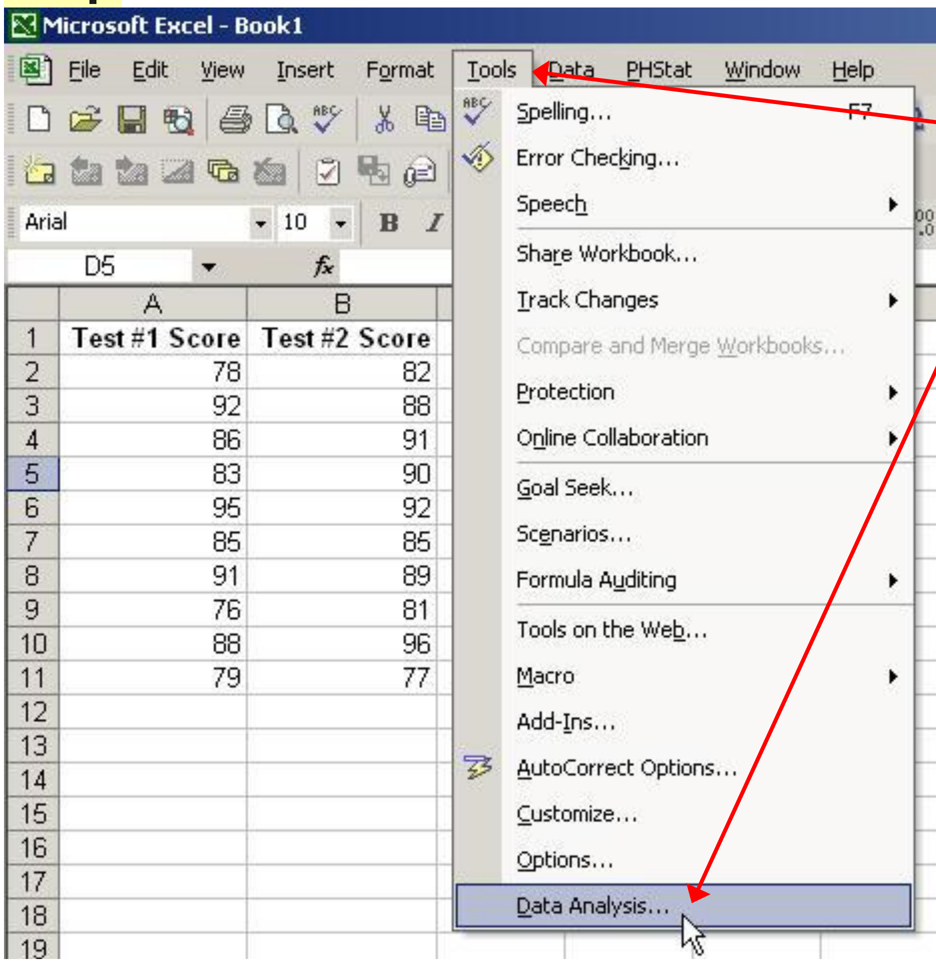
Features of Correlation Coefficient, r

- 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 linear relationship

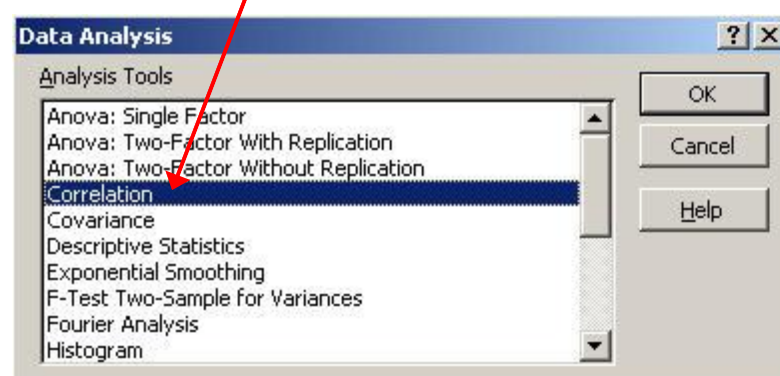
Scatter Plots of Data with Various Correlation Coefficients



Using Excel to Find the Correlation Coefficient

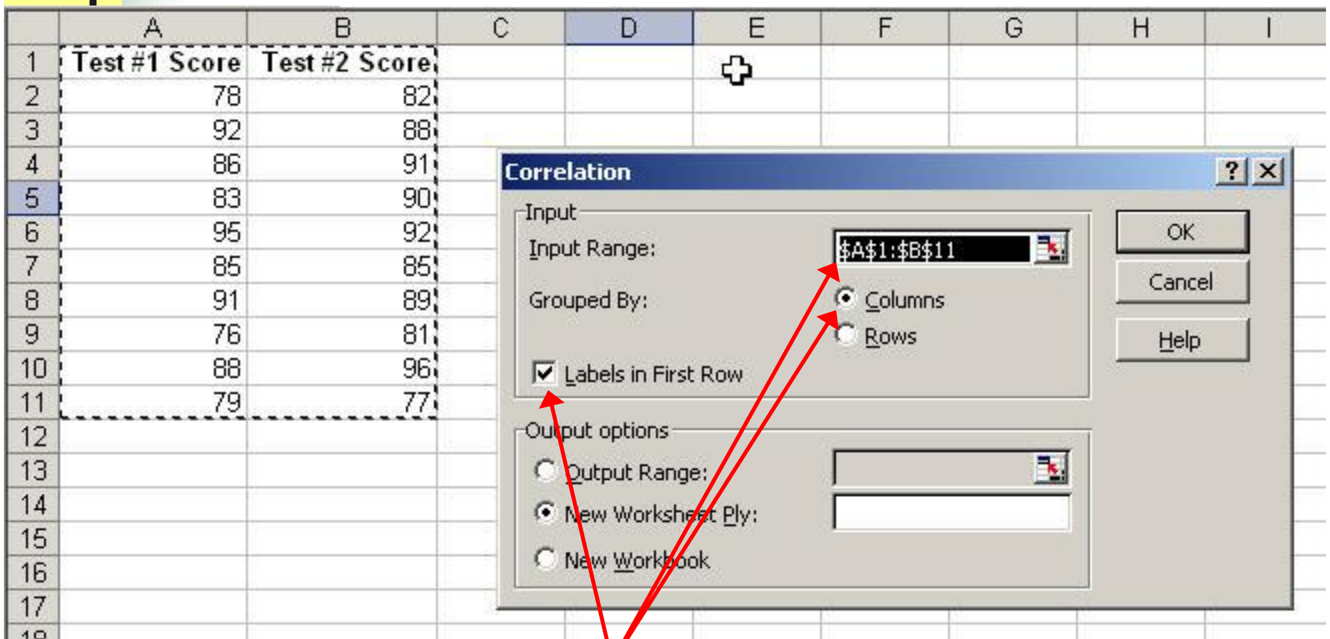


- Select Tools/Data Analysis
- Choose Correlation from the selection menu
- Click OK . . .

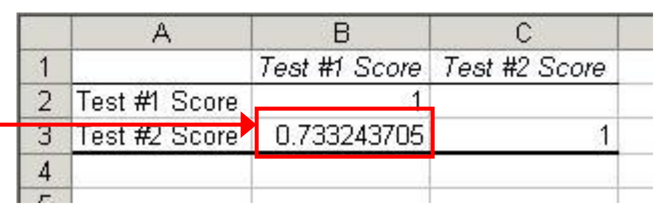


Using Excel to Find the Correlation Coefficient

(continued)

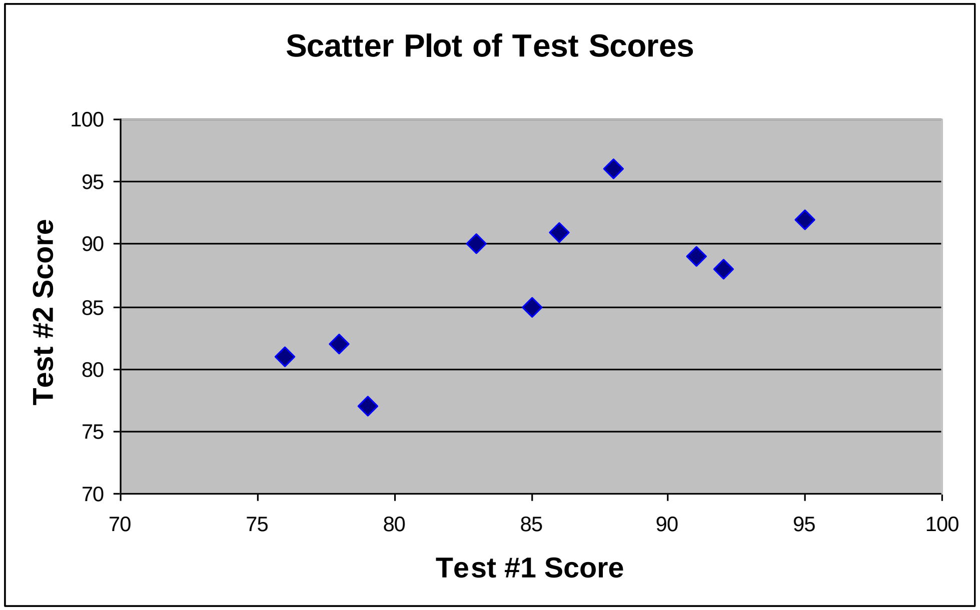


- Input data range and select appropriate options
- Click OK to get output



Interpreting the Result

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



Ethical Considerations

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts

Chapter Summary

- Described measures of central tendency
  - Mean, median, mode, geometric mean
- Discussed quartiles
- Described measures of variation
  - Range, interquartile range, variance and standard deviation, coefficient of variation
- Illustrated shape of distribution
  - Symmetric, skewed, box-and-whisker plots
- Discussed correlation coefficient
- Addressed pitfalls in numerical descriptive measures and ethical considerations