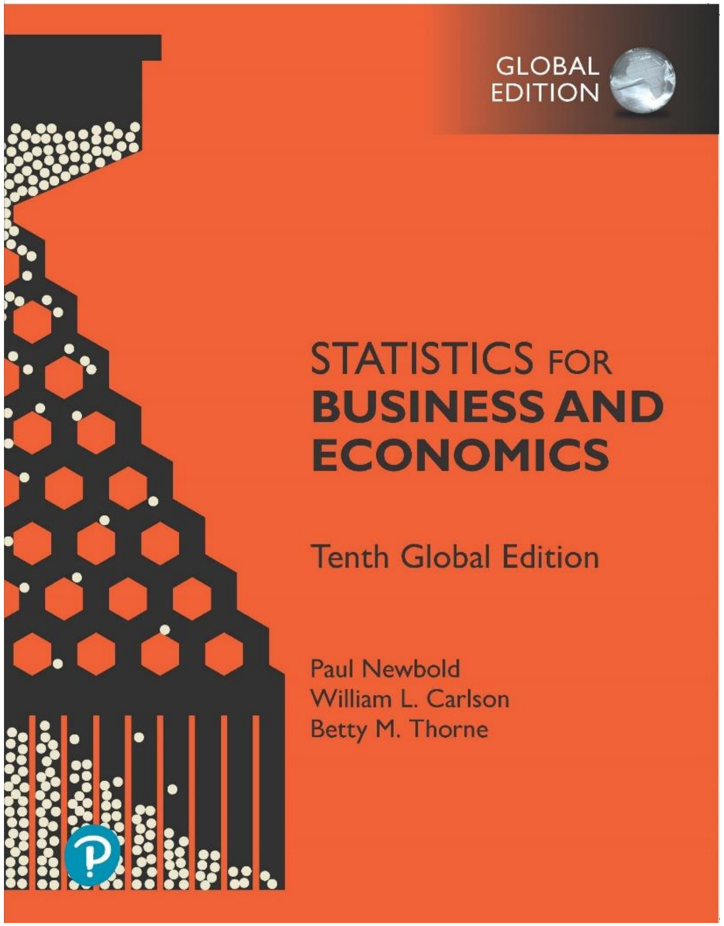
Statistics for Business and Economics

Tenth Edition, Global Edition



Chapter 1

Describing Data: Graphical

Chapter Goals (1 of 3)

After completing this chapter, you should be able to:

- Explain how decisions are often based on incomplete information
- Explain key definitions:
  - Population vs. Sample
  - Parameter vs. Statistic
  - Descriptive vs. Inferential Statistics
- Describe random sampling and systematic sampling
- Explain the difference between Descriptive and Inferential statistics

Chapter Goals (2 of 3)

After completing this chapter, you should be able to:

- Identify types of data and levels of measurement
- Create and interpret graphs to describe categorical variables:
  - frequency distribution, bar chart, pie chart, Pareto diagram
- Create a line chart to describe time-series data
- Create and interpret graphs to describe numerical variables:
  - frequency distribution, histogram, ogive, stem-and-leaf display

Chapter Goals (3 of 3)

After completing this chapter, you should be able to:

- Construct and interpret graphs to describe relationships between variables:
  - Scatter plot, cross table
- Describe appropriate and inappropriate ways to display data graphically

Section 1.1 Decision Making in an Uncertain Environment (1 of 2)

Everyday decisions are based on incomplete information

Examples:

- Will the job market be strong when I graduate?
- Will the price of Yahoo stock be higher in six months than it is now?
- Will interest rates remain low for the rest of the year if the federal budget deficit is as high as predicted?

Section 1.1 Decision Making in an Uncertain Environment (2 of 2)

Data are used to assist decision making

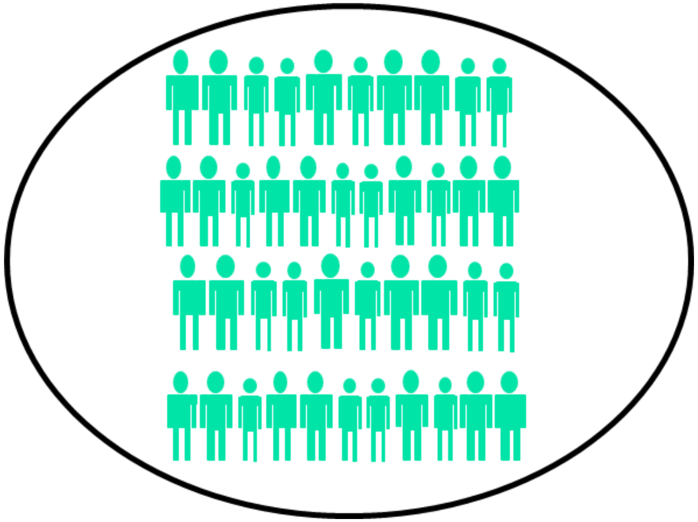
- Statistics is a tool to help process, summarize, analyze, and interpret data

Key Definitions

- A population is the collection of all items of interest or under investigation
  - N represents the population size
- A sample is an observed subset of the population
  - n represents the sample size
- A parameter is a specific characteristic of a population
- A statistic is a specific characteristic of a sample

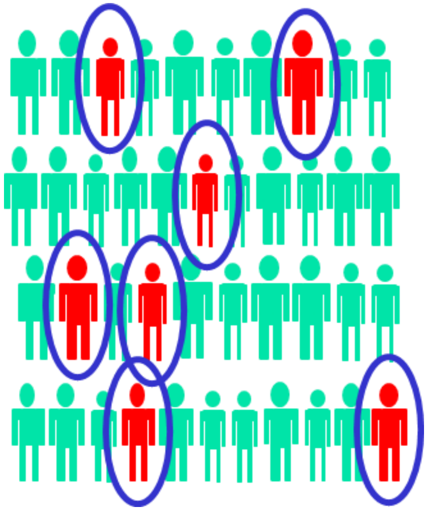
Population vs. Sample

Population



Values calculated using population data are called parameters

Sample



Values computed from sample data are called statistics

Examples of Populations

- Names of all registered voters in the United States
- Incomes of all families living in Daytona Beach
- Annual returns of all stocks traded on the New York Stock Exchange
- Grade point averages of all the students in your university

Random Sampling

Simple random sampling is a procedure in which

- each member of the population is chosen strictly by chance,
- each member of the population is equally likely to be chosen,
- every possible sample of n objects is equally likely to be chosen

The resulting sample is called a random sample

Systematic Sampling (1 of 2)

For systematic sampling,

- Assure that the population is arranged in a way that is not related to the subject of interest
- Select every j^{th} item from the population…
- …where j is the ratio of the population size to the sample size, $j = \frac{N}{n}$
- Randomly select a number from 1 to j for the first item selected

The resulting sample is called a systematic sample

Systematic Sampling (2 of 2)

Example:

Suppose you wish to sample n = 9 items from a population of N = 72.

$$j = \frac{N}{n} = \frac{72}{9} = 8$$

Randomly select a number from 1 to 8 for the first item to include in the sample; suppose this is item number 3.

Then select every 8^{th} item thereafter

$$(\text{items } 3, 11, 19, 27, 35, 43, 51, 59, 67)$$

Descriptive and Inferential Statistics

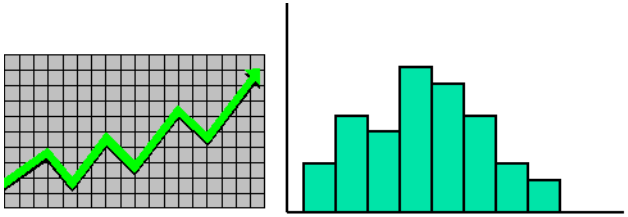
Two branches of statistics:

- Descriptive statistics
  - Graphical and numerical procedures to summarize and process data
- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs



- Summarize data
  - e.g., Sample mean $= \frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 140 pounds

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Section 1.2 Classification of Variables

Data

Categorical

Numerical

Discrete

Continuous

Examples:

- Marital Status
- Are you registered to vote?
- Eye Color

(Defined categories or groups)

Examples:

- Number of Children
- Defects per hour

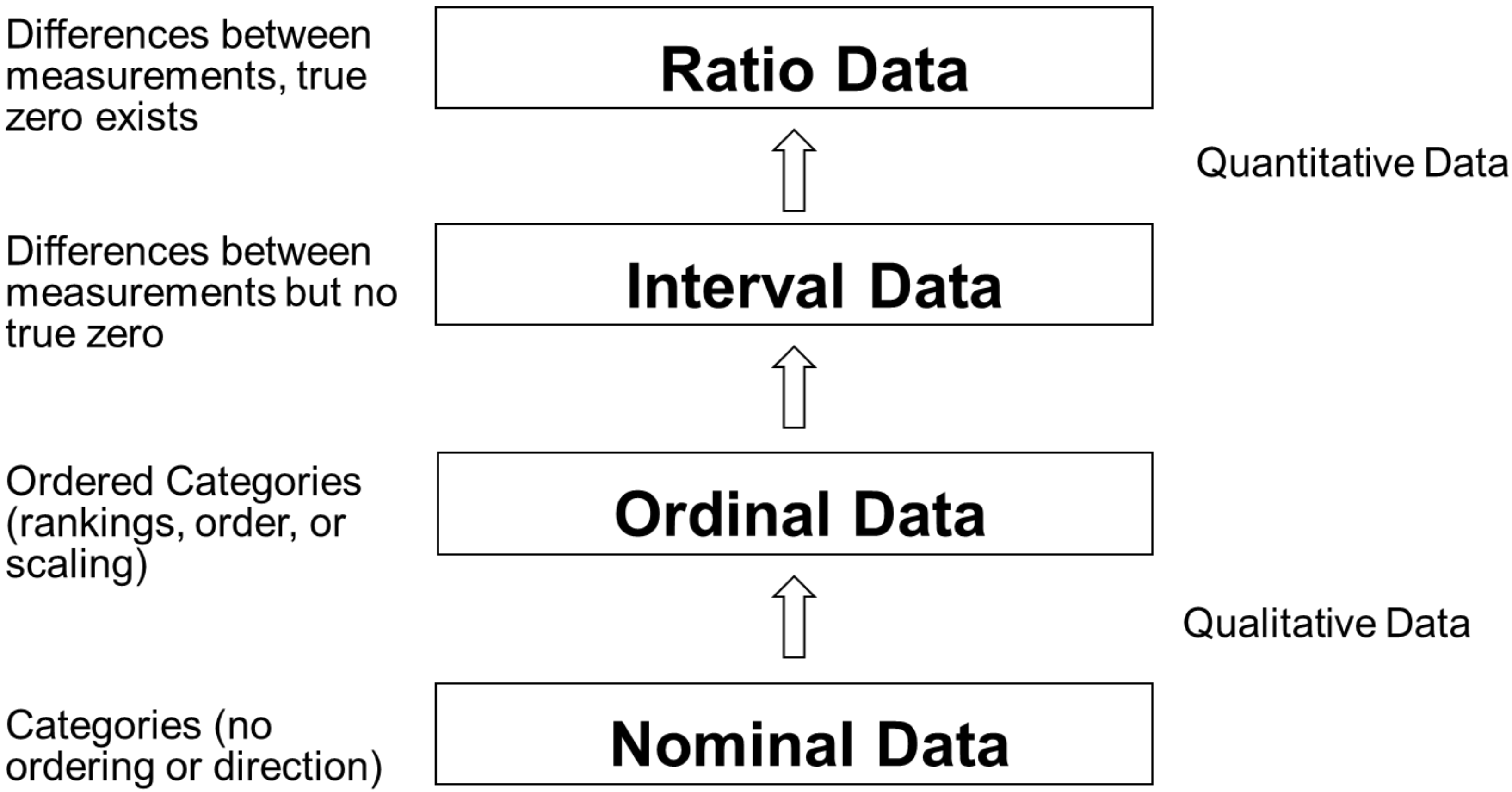
(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Measurement Levels



Section 1.3-1.5 Graphical Presentation of Data (1 of 2)

- Data in raw form are usually not easy to use for decision making
- Some type of organization is needed
  - Table
  - Graph
- The type of graph to use depends on the variable being summarized

Section 1.3-1.5 Graphical Presentation of Data (2 of 2)

- Techniques reviewed in this chapter:

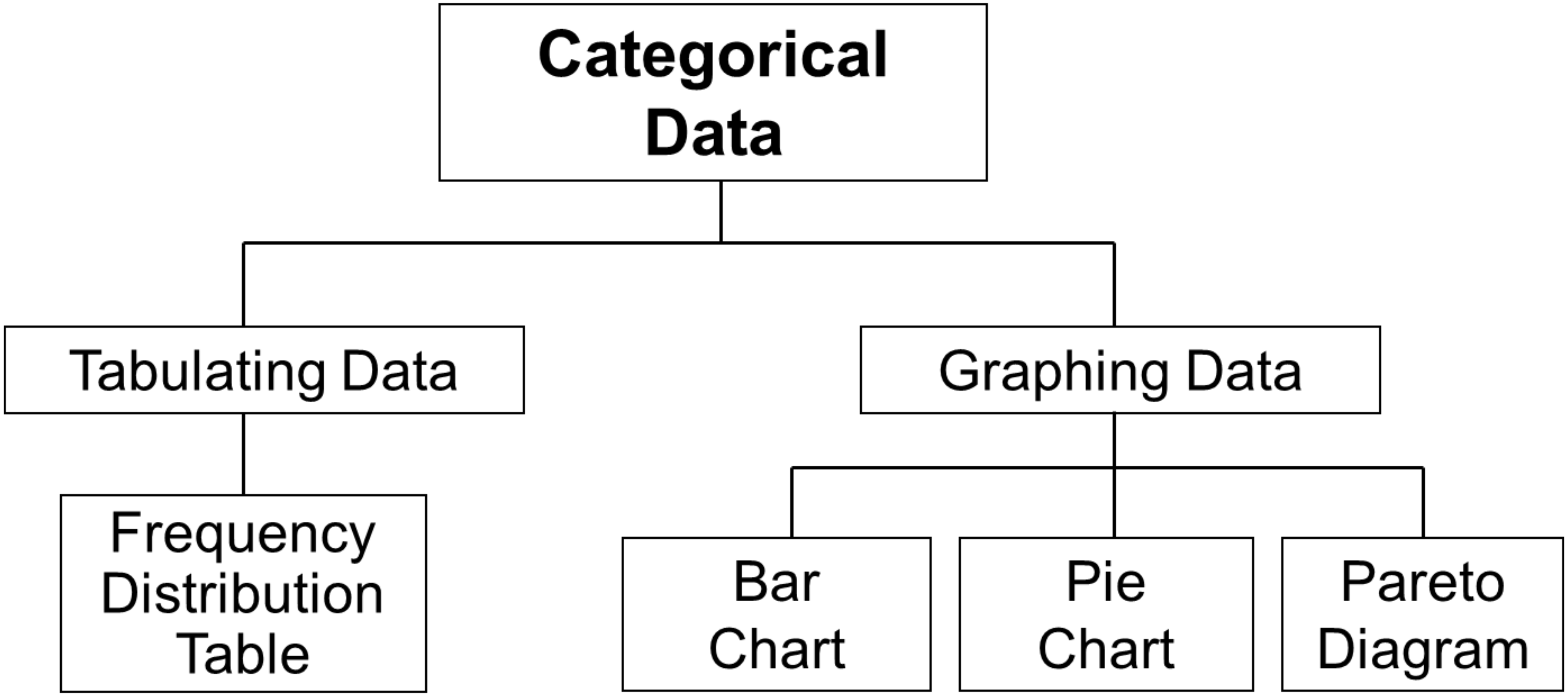
Categorical Variables

- Frequency distribution
- Cross table
- Bar chart
- Pie chart
- Pareto diagram

Numerical Variables

- Line chart
- Frequency distribution
- Histogram and ogive
- Stem-and-leaf display
- Scatter plot

Section 1.3 Tables and Graphs for Categorical Variables



The Frequency Distribution Table

Summarize data by category

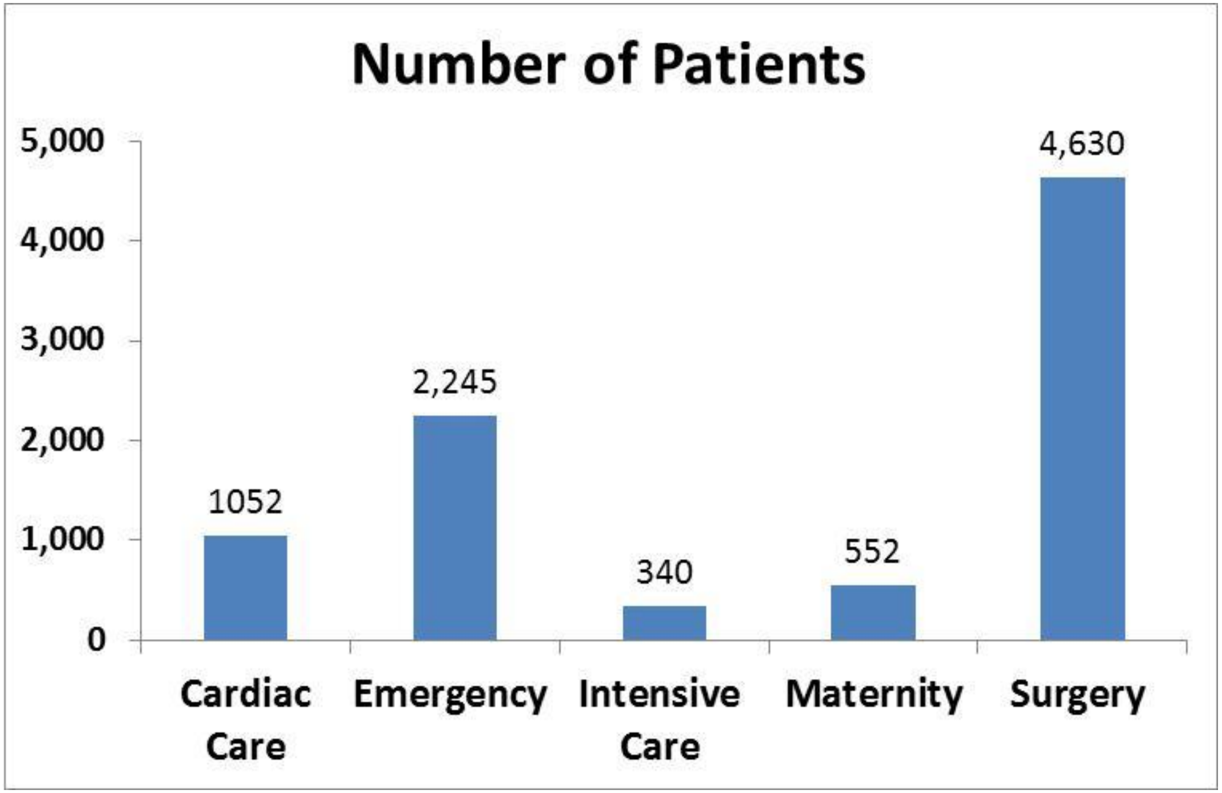
Example: Hospital Patients by Unit

Hospital Unit	Number of Patients	Percent (rounded)
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50
Total:	8,819	100.0

(Variables are categorical)

Graph of Frequency Distribution

- Bar chart of patient data



Cross Tables

- Cross Tables (or contingency tables) list the number of observations for every combination of values for two categorical or ordinal variables
- If there are *r* categories for the first variable (rows) and *c* categories for the second variable (columns), the table is called an $r \times c$ cross table

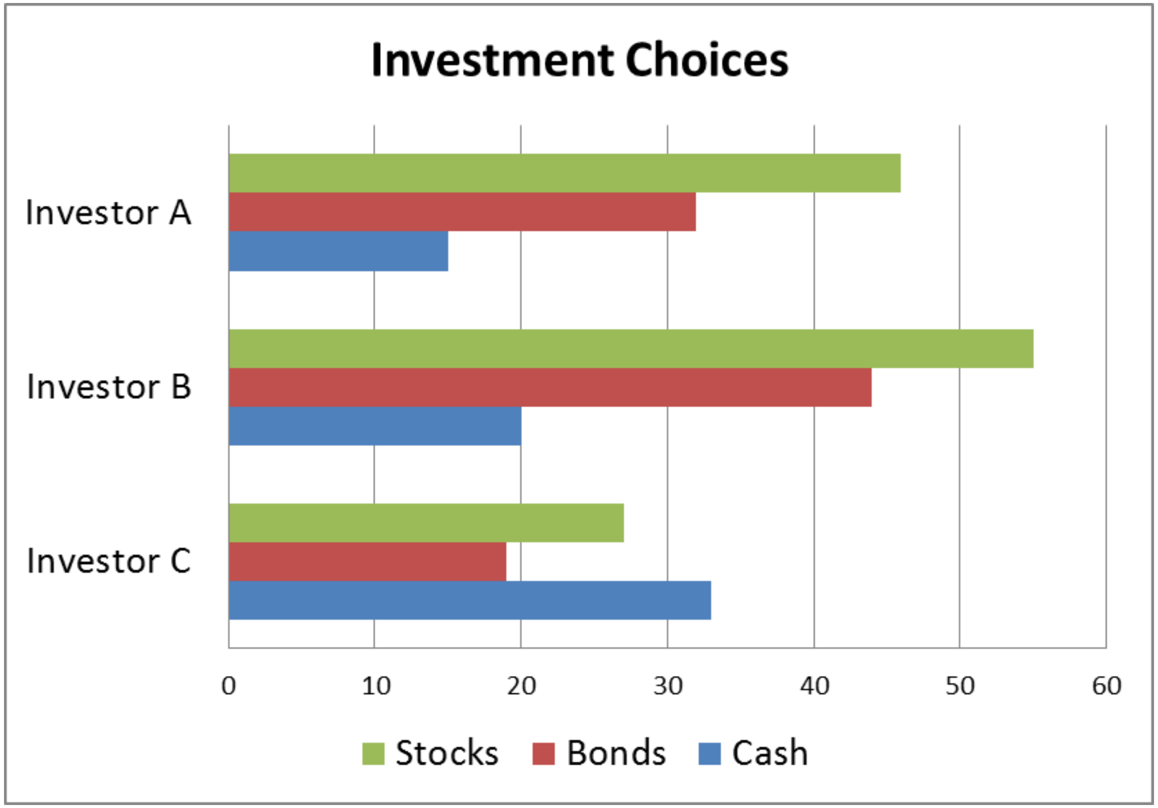
Cross Table Example

- 3×3 Cross Table for Investment Choices by Investor (values in $1000’s)

Investment Category	**Investor A**	**Investor B**	**Investor C**	**Total**
Stocks	46	55	27	**128**
Bonds	32	44	19	**95**
Cash	15	20	33	**68**
Total	**93**	**119**	**79**	**291**

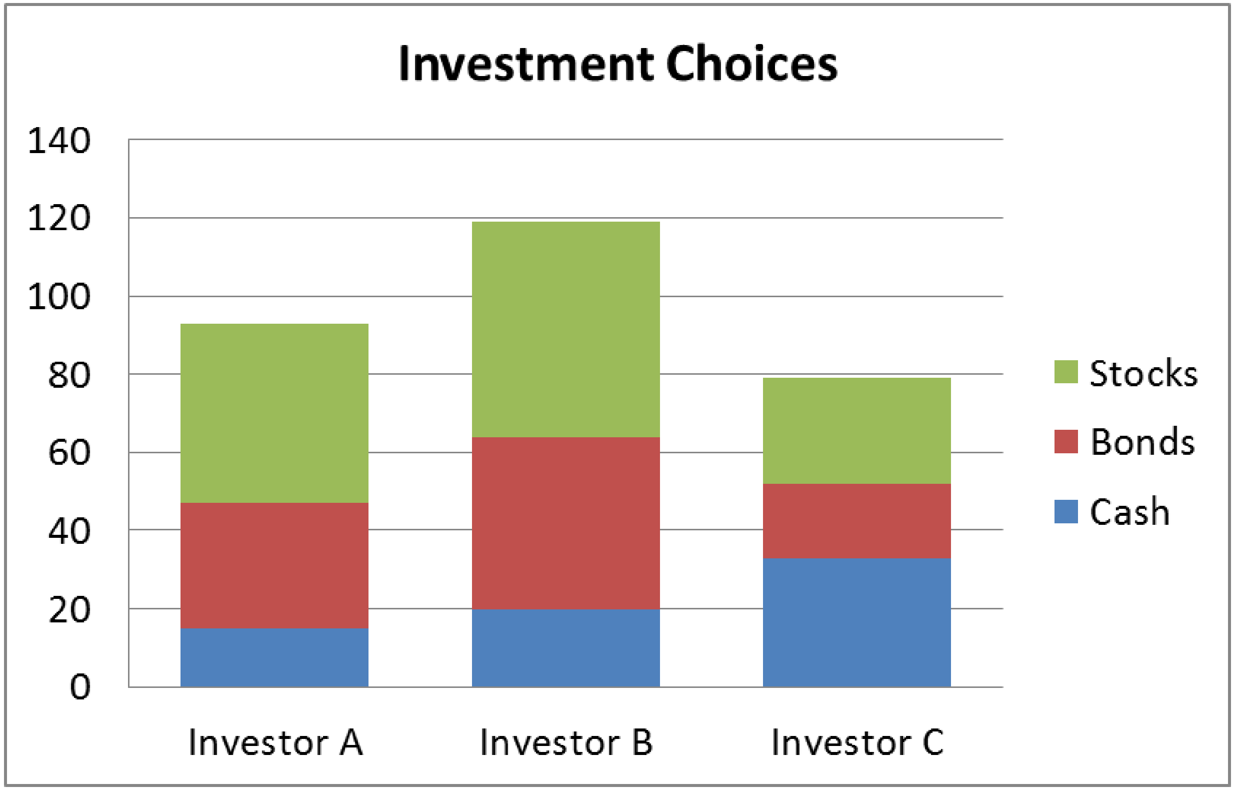
Graphing Multivariate Categorical Data (1 of 2)

- Side by side horizontal bar chart



Graphing Multivariate Categorical Data (2 of 2)

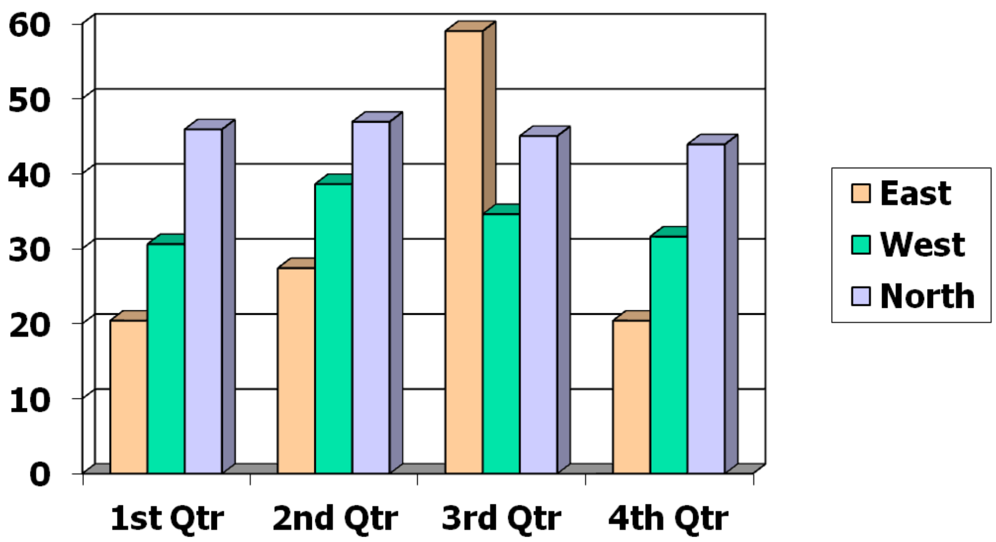
- Stacked bar chart



Vertical Side-by-Side Chart Example

- Sales by quarter for three sales territories:

	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
East	20.4	27.4	59	20.4
West	30.6	38.6	34.6	31.6
North	45.9	46.9	45	43.9

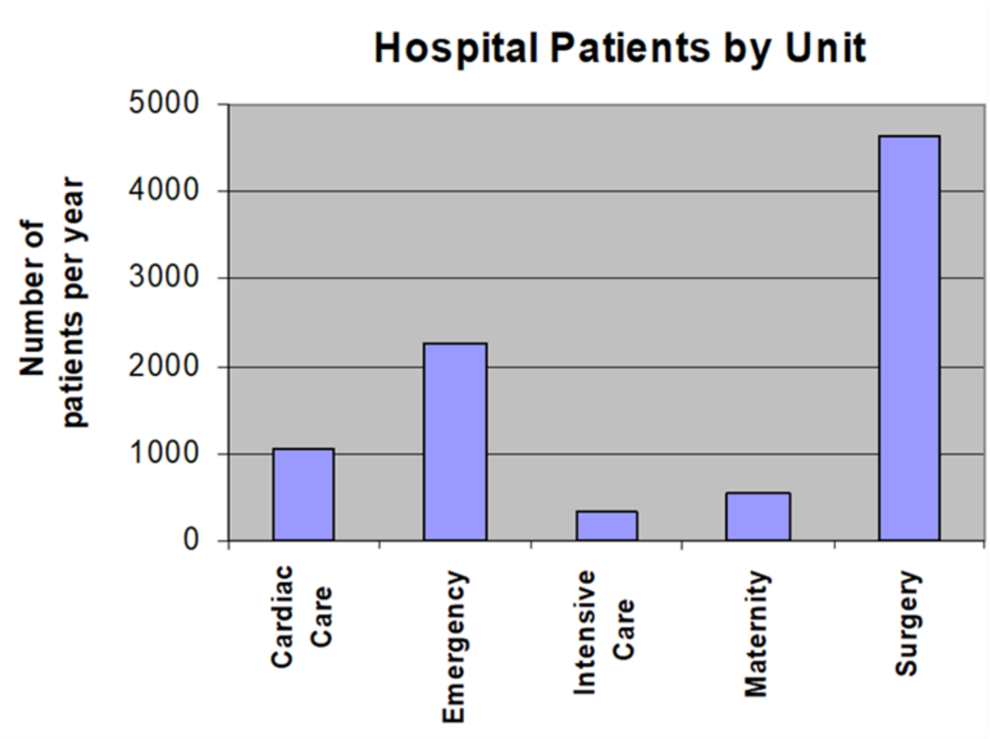


Bar and Pie Charts

- Bar charts and Pie charts are often used for qualitative (categorical) data
- Height of bar or size of pie slice shows the frequency or percentage for each category

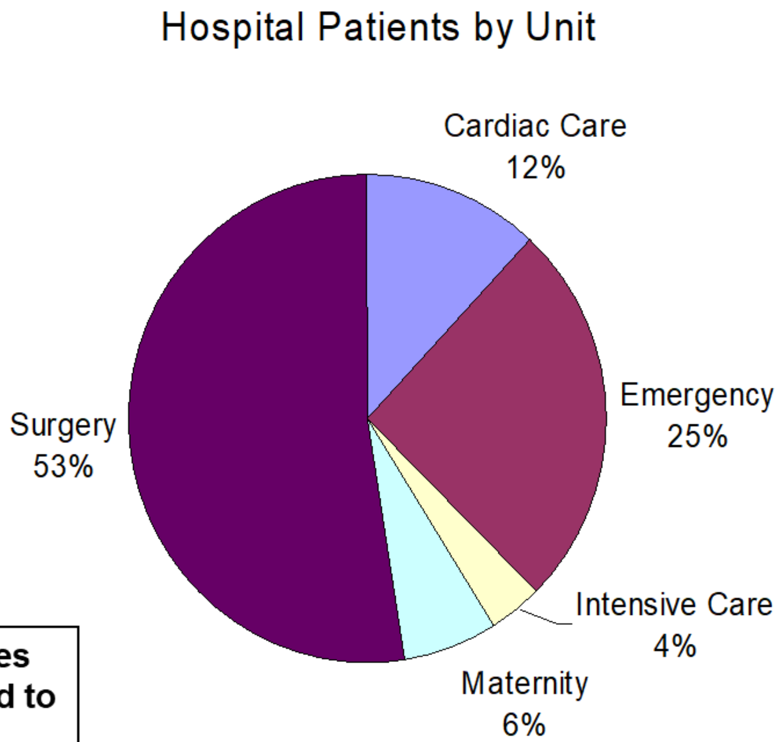
Bar Chart Example

Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630



Pie Chart Example

Hospital Unit	Number of Patients	% of Total
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50



(Percentages are rounded to the nearest percent)

Pareto Diagram

- Used to portray categorical data
- A bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is often shown in the same graph
- Used to separate the “vital few” from the “trivial many”

Pareto Diagram Example (1 of 3)

Example: 400 defective items are examined for cause of defect:

Source of Manufacturing Error	Number of defects
Bad Weld	34
Poor Alignment	223
Missing Part	25
Paint Flaw	78
Electrical Short	19
Cracked case	21
Total	**400**

Pareto Diagram Example (2 of 3)

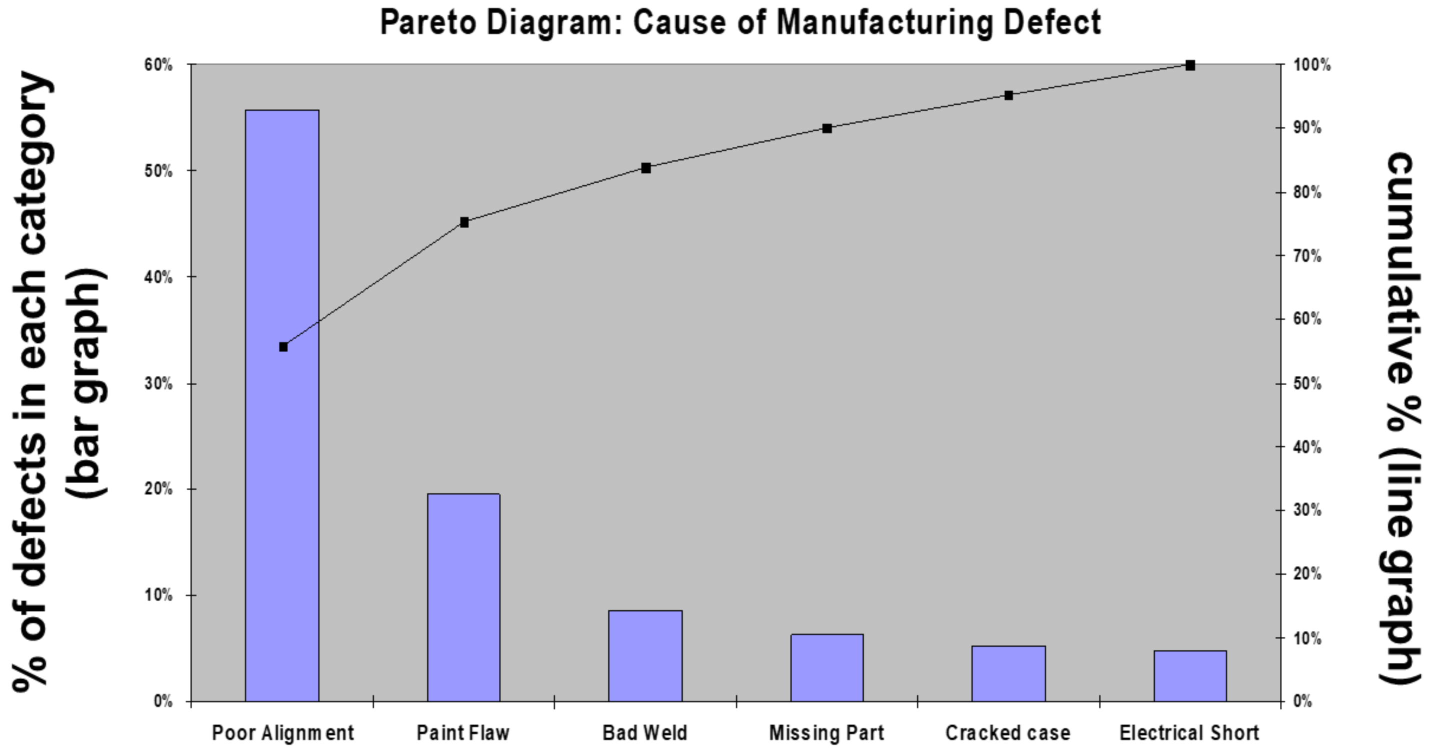
Step 1: Sort by defect cause, in descending order

Step 2: Determine % in each category

Source of Manufacturing Error	Number of defects	% of Total Defects
Poor Alignment	223	55.75
Paint Flaw	78	19.50
Bad Weld	34	8.50
Missing Part	25	6.25
Cracked case	21	5.25
Electrical Short	19	4.75
Total	**400**	**100%**

Pareto Diagram Example (3 of 3)

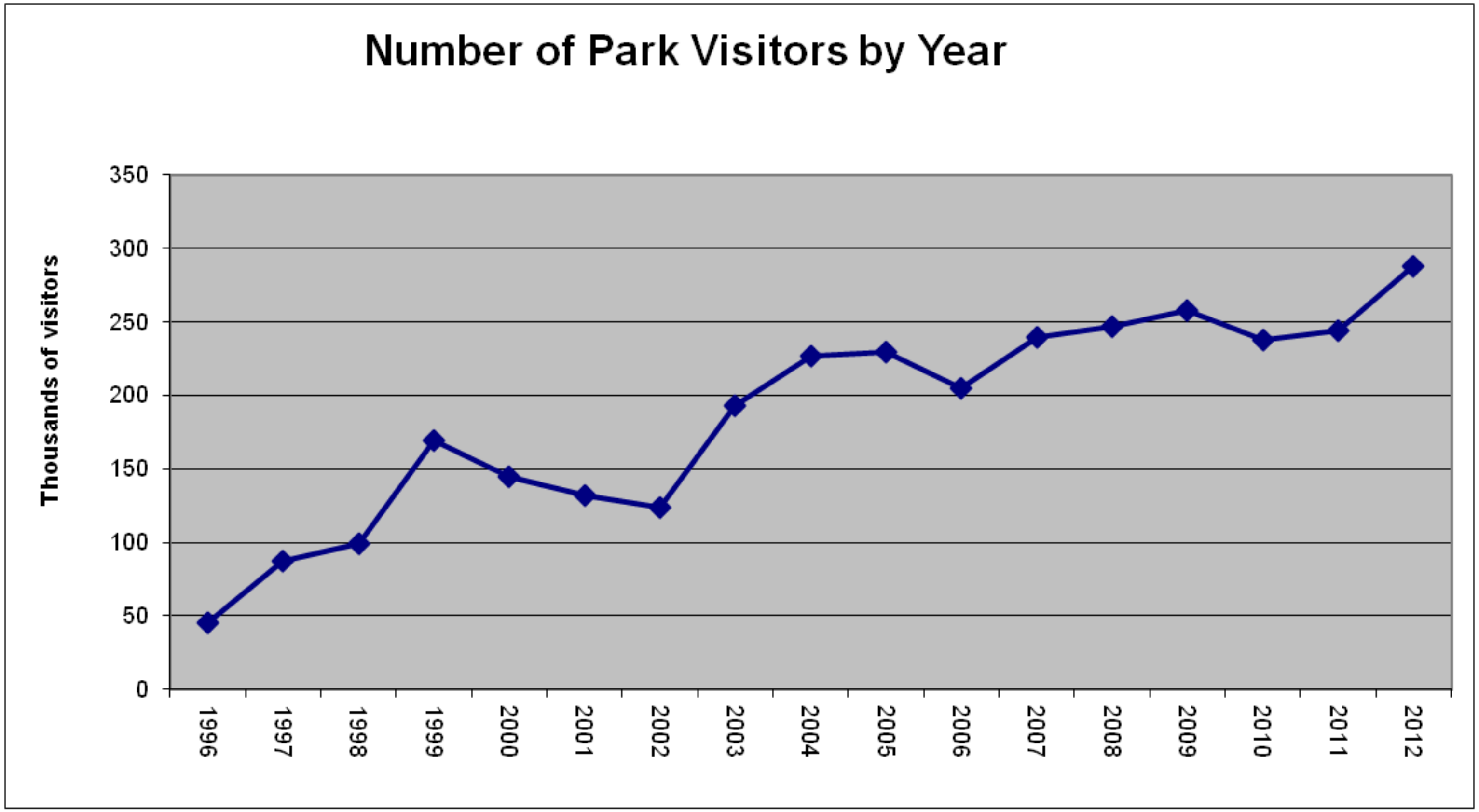
Step 3: Show results graphically



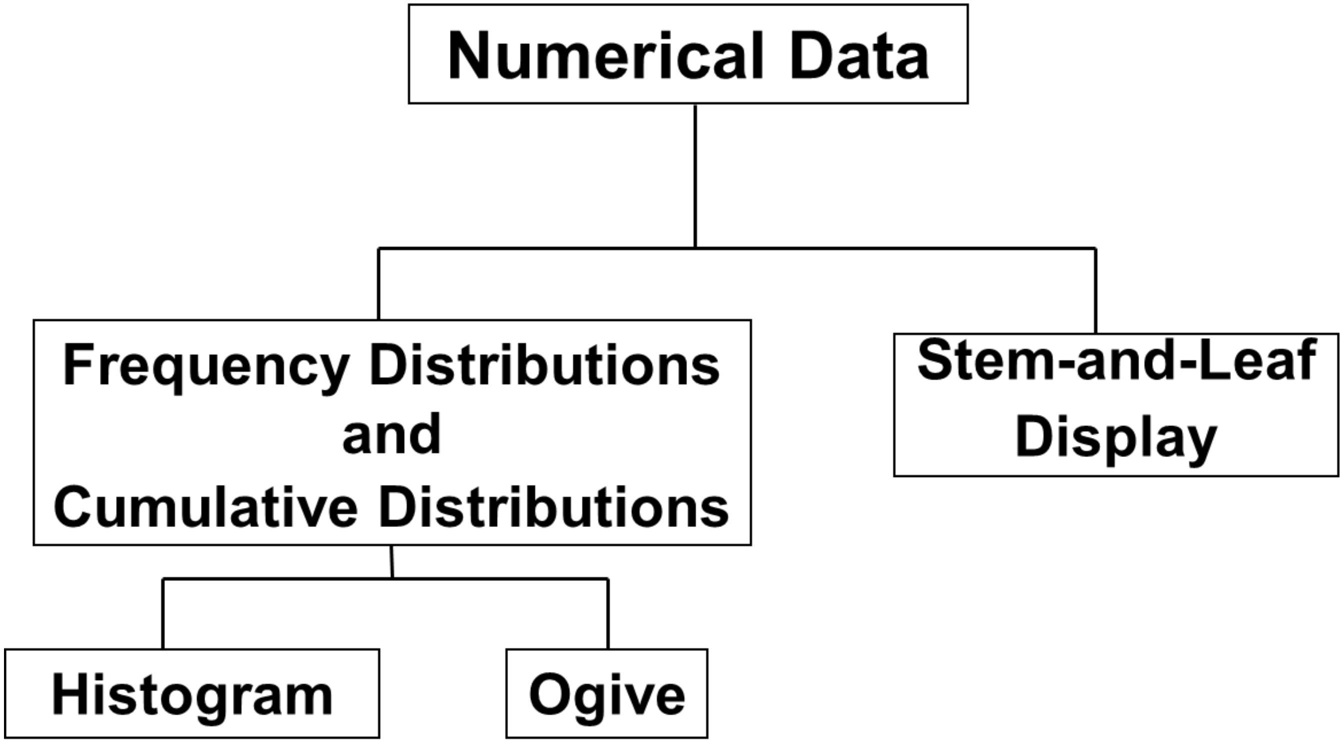
Section 1.4 Graphs to Describe Time-Series Data

- A line chart (time-series plot) is used to show the values of a variable over time
- Time is measured on the horizontal axis
- The variable of interest is measured on the vertical axis

Line Chart Example



Section 1.5 Graphs to Describe Numerical Variables



Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a list or a table…
- containing class groupings (categories or ranges within which the data fall)...
- and the corresponding frequencies with which data fall within each class or category

Why Use Frequency Distributions?

- A frequency distribution is a way to summarize data
- The distribution condenses the raw data into a more useful form...
- and allows for a quick visual interpretation of the data

Class Intervals and Class Boundaries

- Each class grouping has the same width
- Determine the width of each interval by

$$w = \text{interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of desired intervals}}$$

- Use at least 5 but no more than 15-20 intervals
- Intervals never overlap
- Round up the interval width to get desirable interval endpoints

Frequency Distribution Example (1 of 3)

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

data:

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,

32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Frequency Distribution Example (2 of 3)

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: $\mathbf{58 - 12 = 46}$
- Select number of classes: **5 (usually between 5 and 15)**
- Compute interval width: $\mathbf{10}\left(\dfrac{\mathbf{46}}{\mathbf{5}}\ \textbf{then round up}\right)$
- Determine interval boundaries: **10 but less than 20, 20 but less than $\mathbf{30,\ldots,60}$ but less than 70**
- Count observations & assign to classes

Frequency Distribution Example (3 of 3)

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

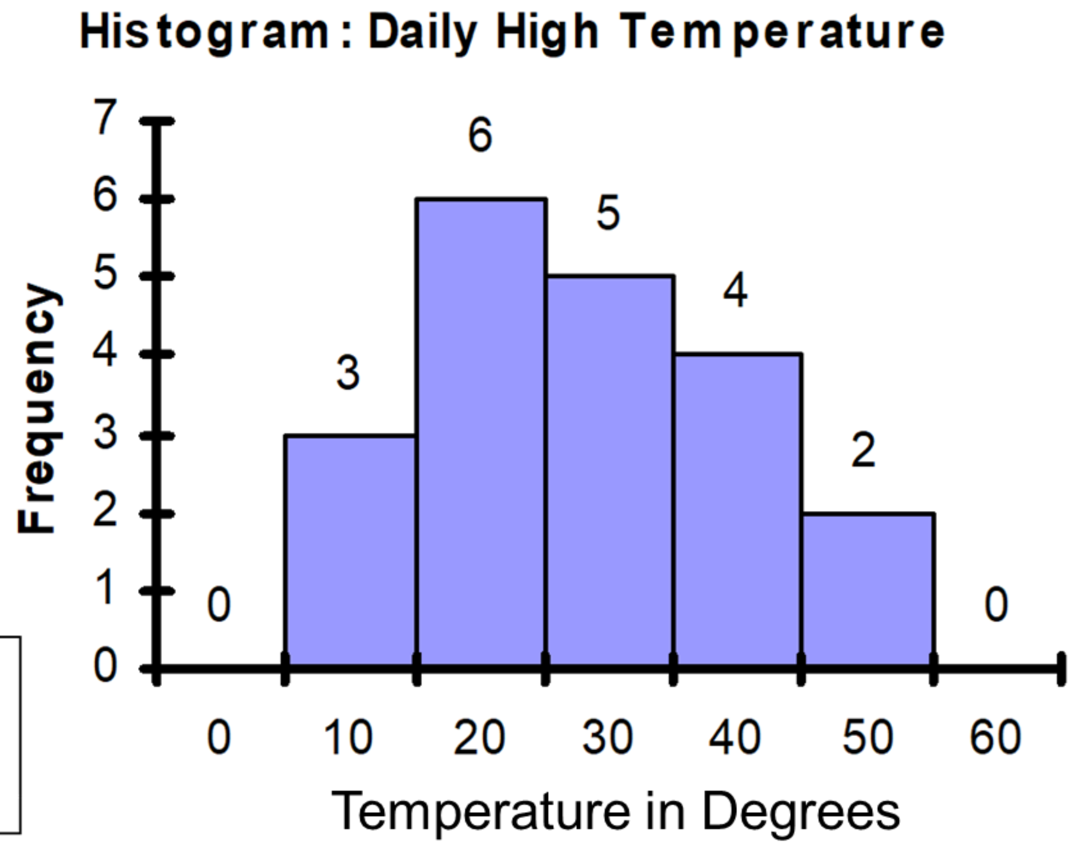
Interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Histogram

- A graph of the data in a frequency distribution is called a **histogram**
- The **interval endpoints** are shown on the horizontal axis
- the vertical axis is either **frequency, relative frequency,** or **percentage**
- Bars of the appropriate heights are used to represent the number of observations within each class

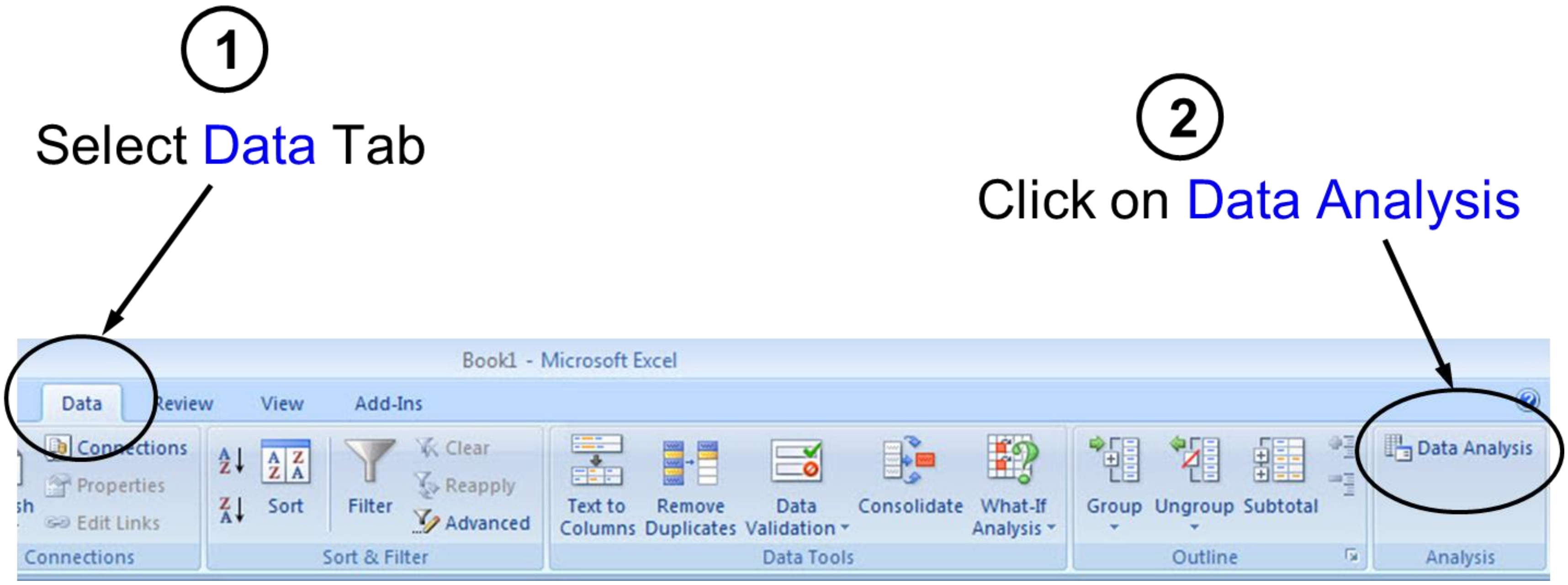
Histogram Example

Interval	Frequency
10 but less than 20	3
20 but less than 30	6
30 but less than 40	5
40 but less than 50	4
50 but less than 60	2

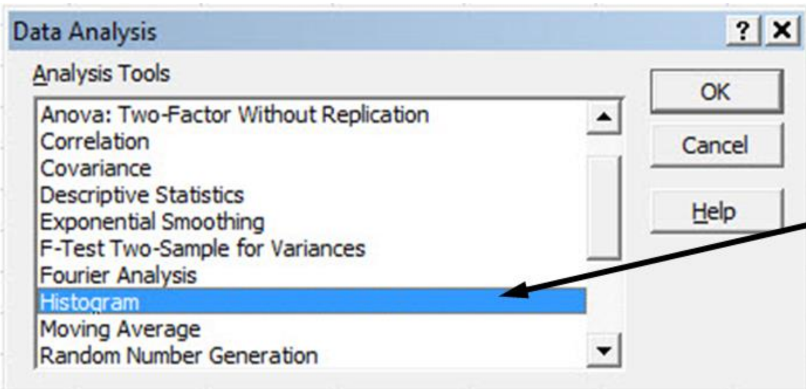


(No gaps between bars)

Histograms in Excel (1 of 2)

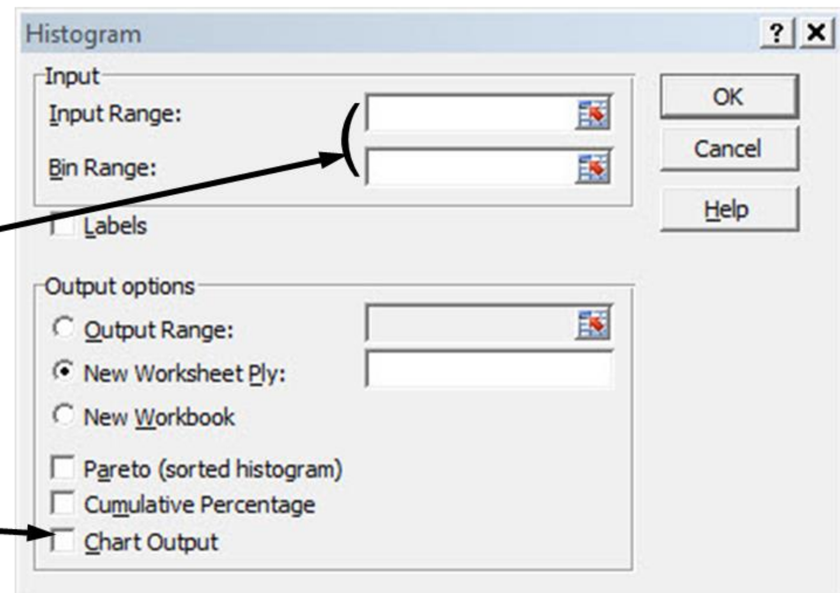


Histograms in Excel (2 of 2)



3

Choose Histogram



4

Input data range and bin range (bin range is a cell range containing the upper interval endpoints for each class grouping)

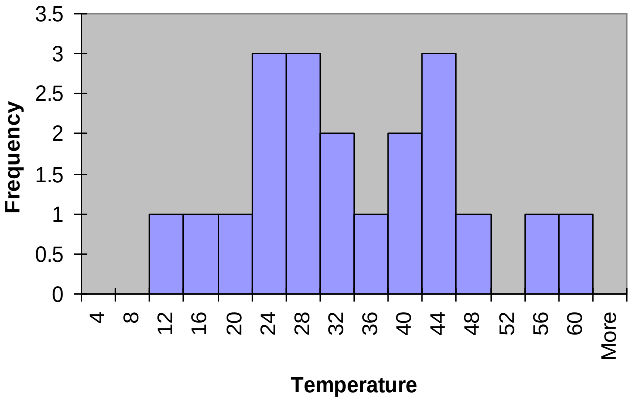
Select Chart Output and click "OK"

Questions for Grouping Data into Intervals

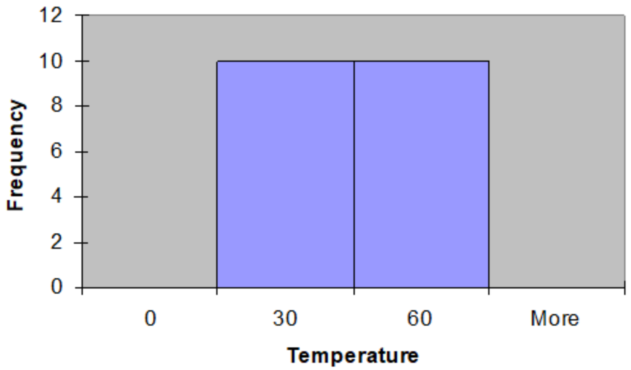
- How wide should each interval be?
  (How many classes should be used?)
- How should the endpoints of the intervals be determined?
  - Often answered by trial and error, subject to user judgment
  - The goal is to create a distribution that is neither too "jagged" nor too "blocky”
  - Goal is to appropriately show the pattern of variation in the data

How Many Class Intervals?

- **Many (Narrow class intervals)**
  - may yield a very jagged distribution with gaps from empty classes
  - Can give a poor indication of how frequency varies across classes



- **Few (Wide class intervals)**
  - may compress variation too much and yield a blocky distribution
  - can obscure important patterns of variation.



(*X* axis labels are upper class endpoints)

The Cumulative Frequency Distribution

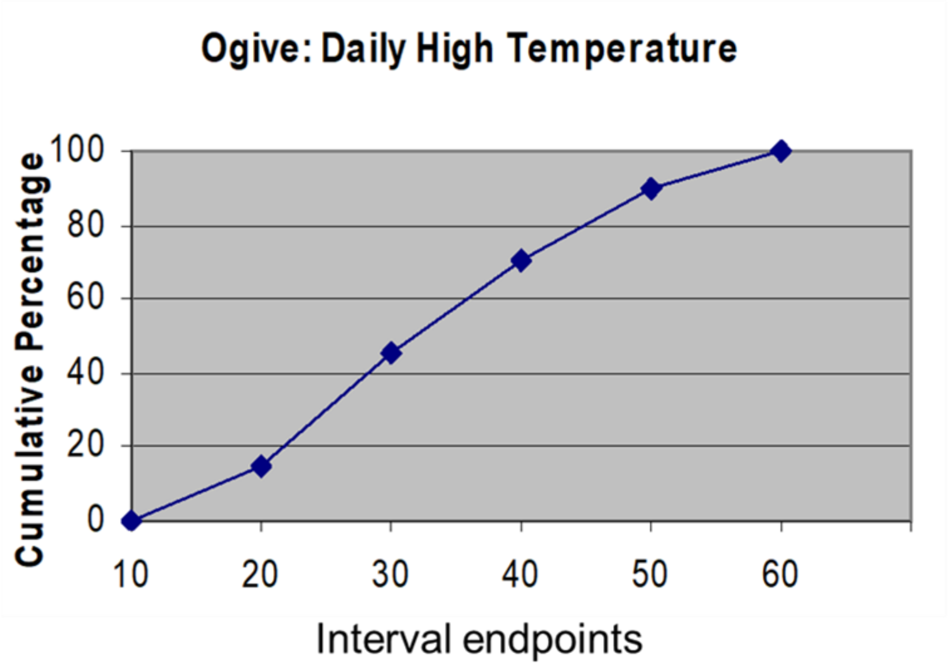
Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

The Ogive Graphing Cumulative Frequencies

Interval	Upper interval endpoint	Cumulative Percentage
Less than 10	10	0
10 but less than 20	20	15
20 but less than 30	30	45
30 but less than 40	40	70
40 but less than 50	50	90
50 but less than 60	60	100



Stem-and-Leaf Diagram

- A simple way to see distribution details in a data set

Method: Separate the sorted data series into leading digits (the **stem**) and the trailing digits (the **leaves**)

Example (1 of 2)

Data in ordered array:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

- Here, use the 10's digit for the stem unit:
  - 21 is shown as → 2 | 1
  - 38 is shown as → 3 | 8

Stem	Leaf
2	1
3	8

Example (2 of 2)

Data in ordered array:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

- Completed stem-and-leaf diagram:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

Using Other Stem Units (1 of 2)

- Using the 100's digit as the stem:
  - Round off the 10's digit to form the leaves

    - 613 would become →
    - 776 would become →
    - . . .
    - 1224 becomes →

Stem	Leaf
6	1
7	8
12	2

Using Other Stem Units (2 of 2)

- Using the 100's digit as the stem:
  - The completed stem-and-leaf display:

Data:

613, 632, 658, 717, 722, 750, 776, 827, 841, 859, 863, 891, 894, 906, 928, 933, 955, 982, 1034, 1047,1056, 1140, 1169, 1224

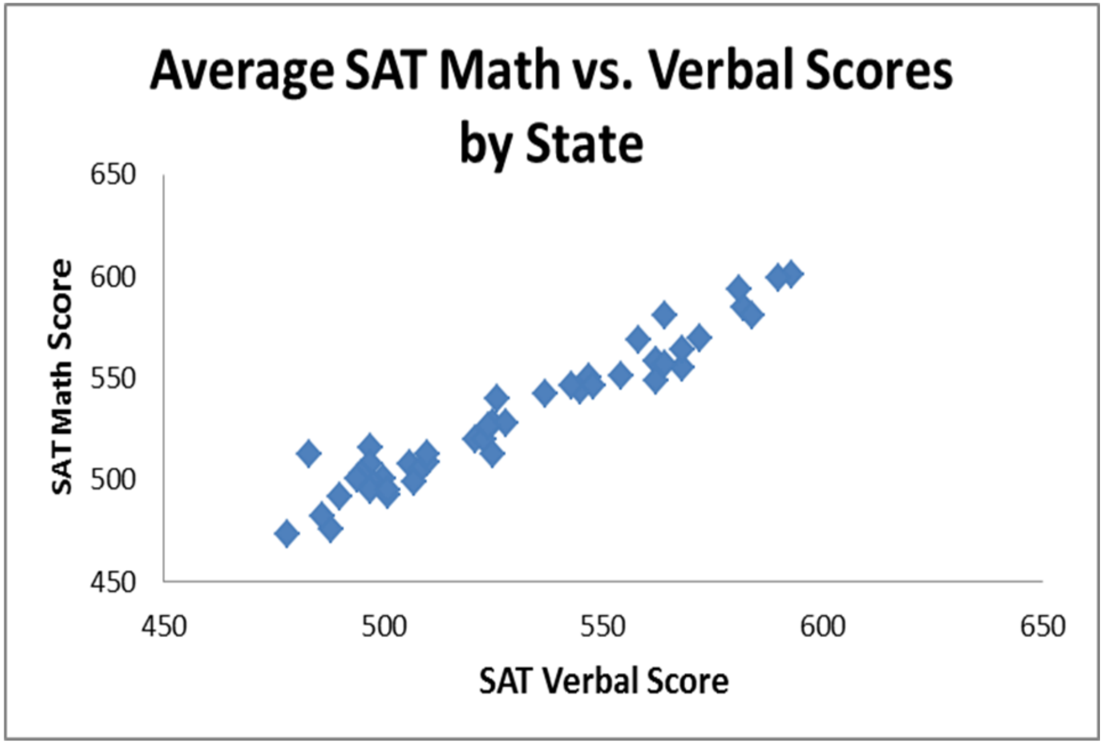
Stem	Leaves
6	1 3 6
7	2 2 5 8
8	3 4 6 6 9 9
9	1 3 3 6 8
10	3 5 6
11	4 7
12	2

Scatter Diagrams

- Scatter Diagrams are used for paired observations taken from two numerical variables
- The Scatter Diagram:
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

Scatter Diagram Example

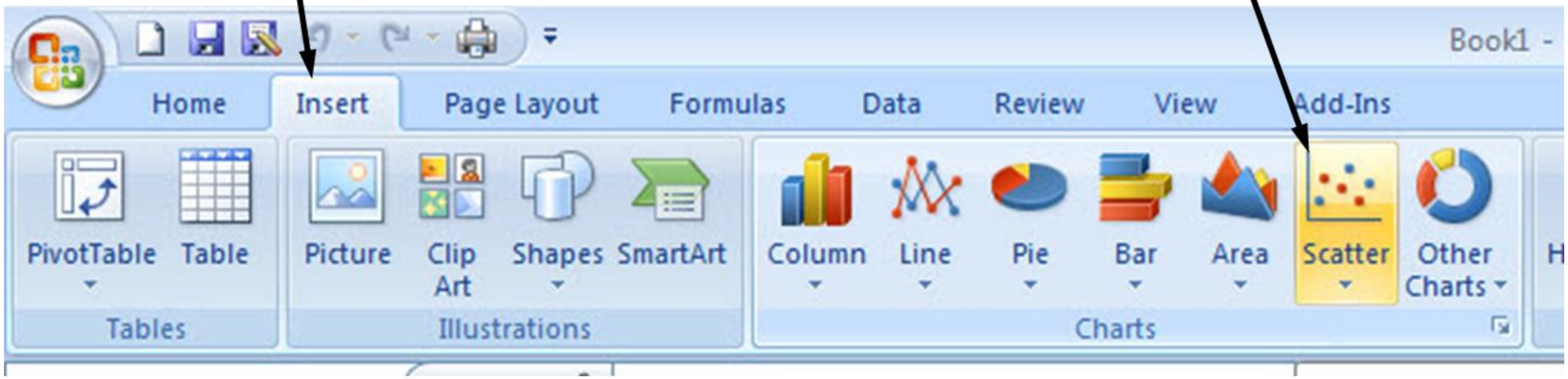
Average SAT scores by state: 1998		
	Verbal	Math
Alabama	562	558
Alaska	521	520
Arizona	525	528
Arkansas	568	555
California	497	516
Colorado	537	542
Connecticut	510	509
Delaware	501	493
D.C.	488	476
Florida	500	501
Georgia	486	482
Hawaii	483	513
. . .		
W.Va.	525	513
Wis.	581	594
Wyo.	548	546



Scatter Diagrams in Excel

1. Select the Insert tab

2. Select Scatter type from the Charts section



3. When prompted, enter the data range, desired legend, and desired destination to complete the scatter diagram

Section 1.6 Data Presentation Errors (1 of 2)

Goals for effective data presentation:

- Present data to display essential information
- Communicate complex ideas clearly and accurately
- Avoid distortion that might convey the wrong message

Section 1.6 Data Presentation Errors (2 of 2)

- Unequal histogram interval widths
- Compressing or distorting the vertical axis
- Providing no zero point on the vertical axis
- Failing to provide a relative basis in comparing data between groups

Chapter Summary (1 of 2)

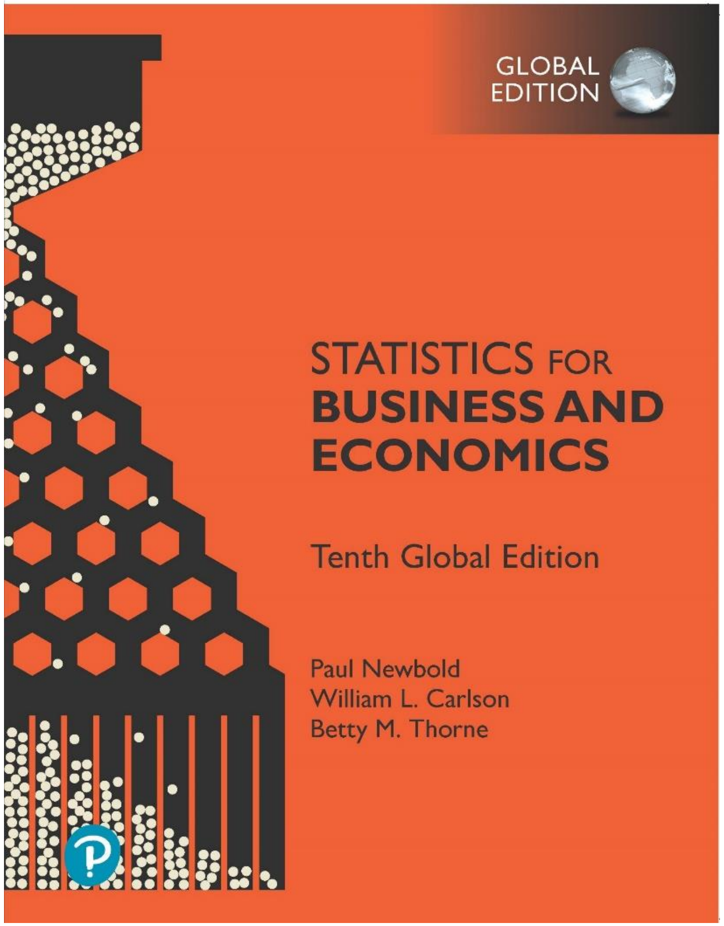
- Reviewed incomplete information in decision making
- Introduced key definitions:
  - Population vs. Sample
  - Parameter vs. Statistic
  - Descriptive vs. Inferential statistics
- Described random sampling
- Examined the decision making process

Chapter Summary (2 of 2)

- Reviewed types of data and measurement levels
- Data in raw form are usually not easy to use for decision making -- Some type of organization is needed:
  - Table
  - Graph
- Techniques reviewed in this chapter:
  - Frequency distribution
  - Cross tables
  - Bar chart
  - Pie chart
  - Pareto diagram
  - Line chart
  - Frequency distribution
  - Histogram and ogive
  - Stem-and-leaf display
  - Scatter plot

Statistics for Business and Economics

Tenth Edition, Global 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

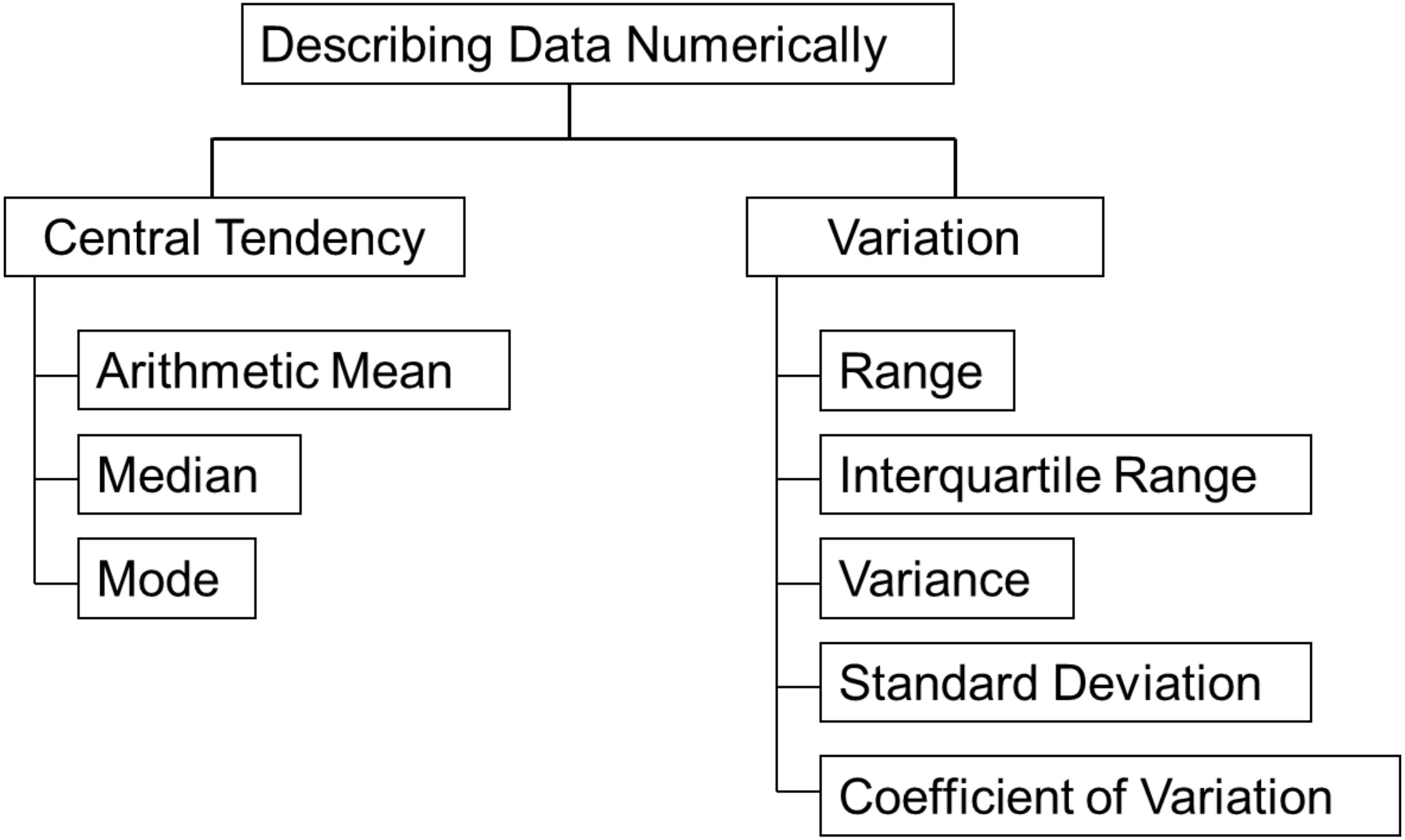
Chapter Topics (1 of 2)

- 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

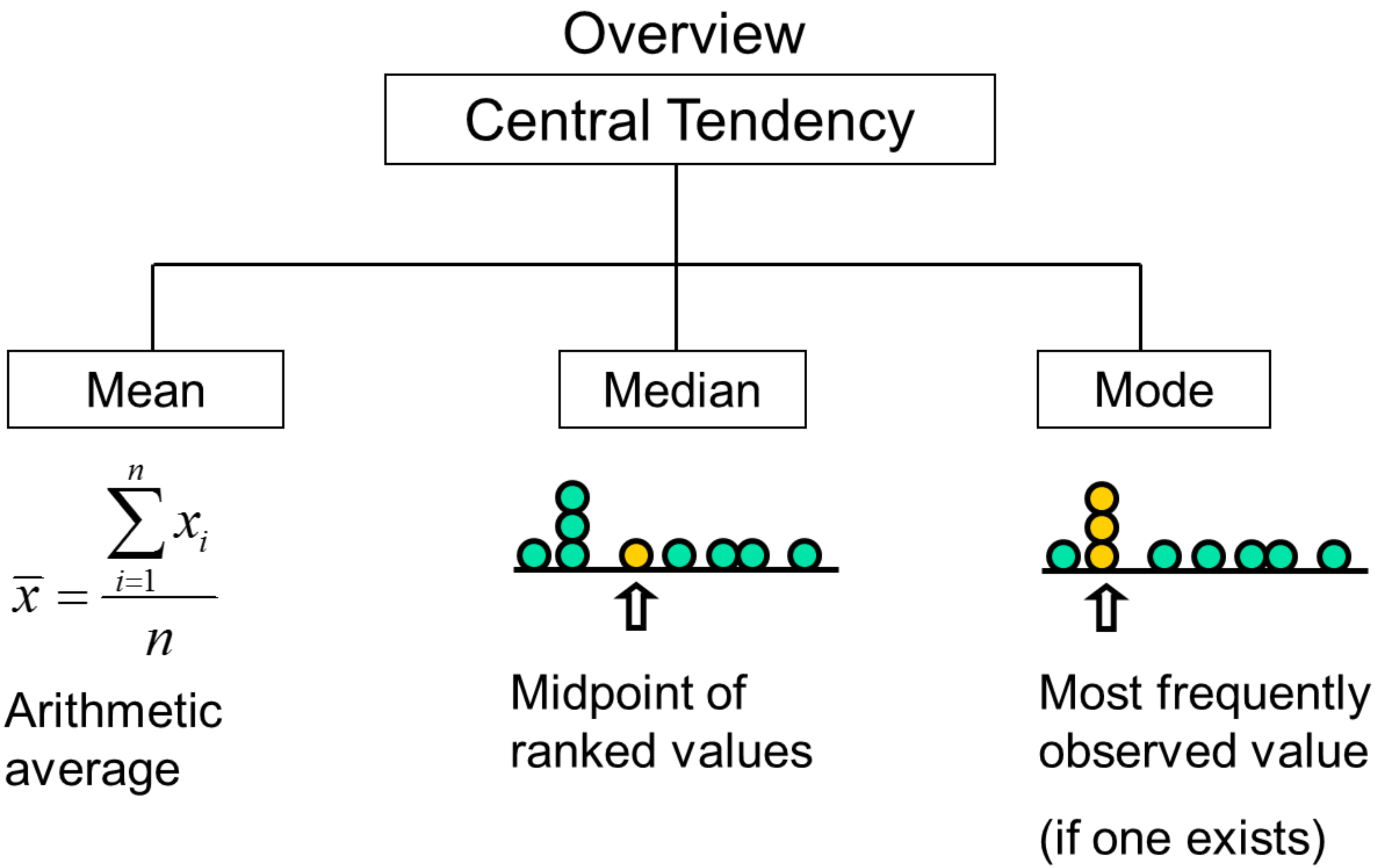
Chapter Topics (2 of 2)

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

Describing Data Numerically



Section 2.1 Measures of Central Tendency



Pearson

Arithmetic Mean (1 of 2)

- 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*:

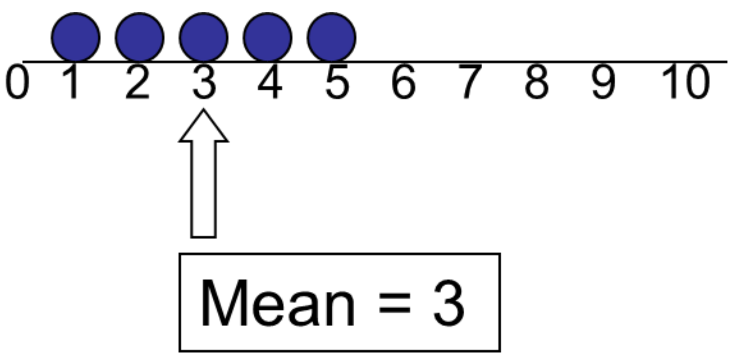
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Observed values

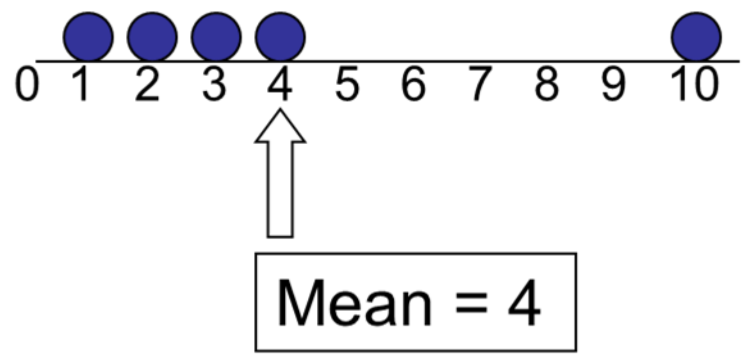
Sample size

Arithmetic Mean (2 of 2)

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



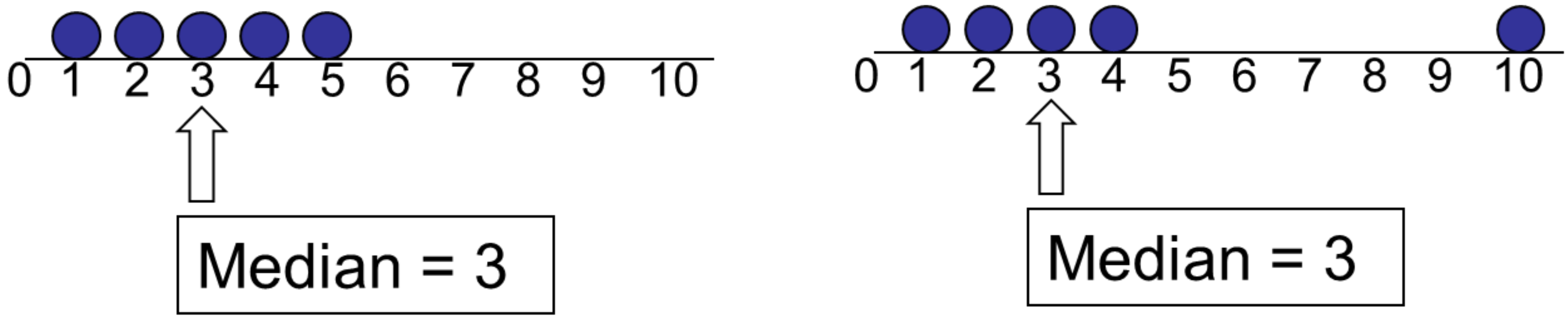
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



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

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:

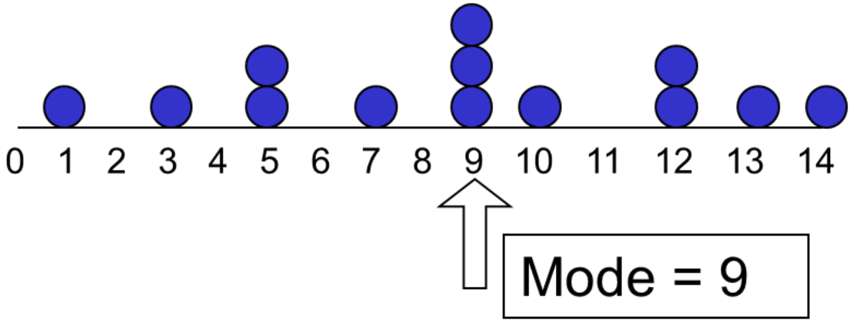
$$\text{Median position} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{position in the ordered data}$$

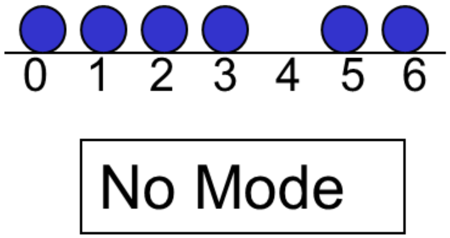
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- 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 be no mode
- There may be several modes



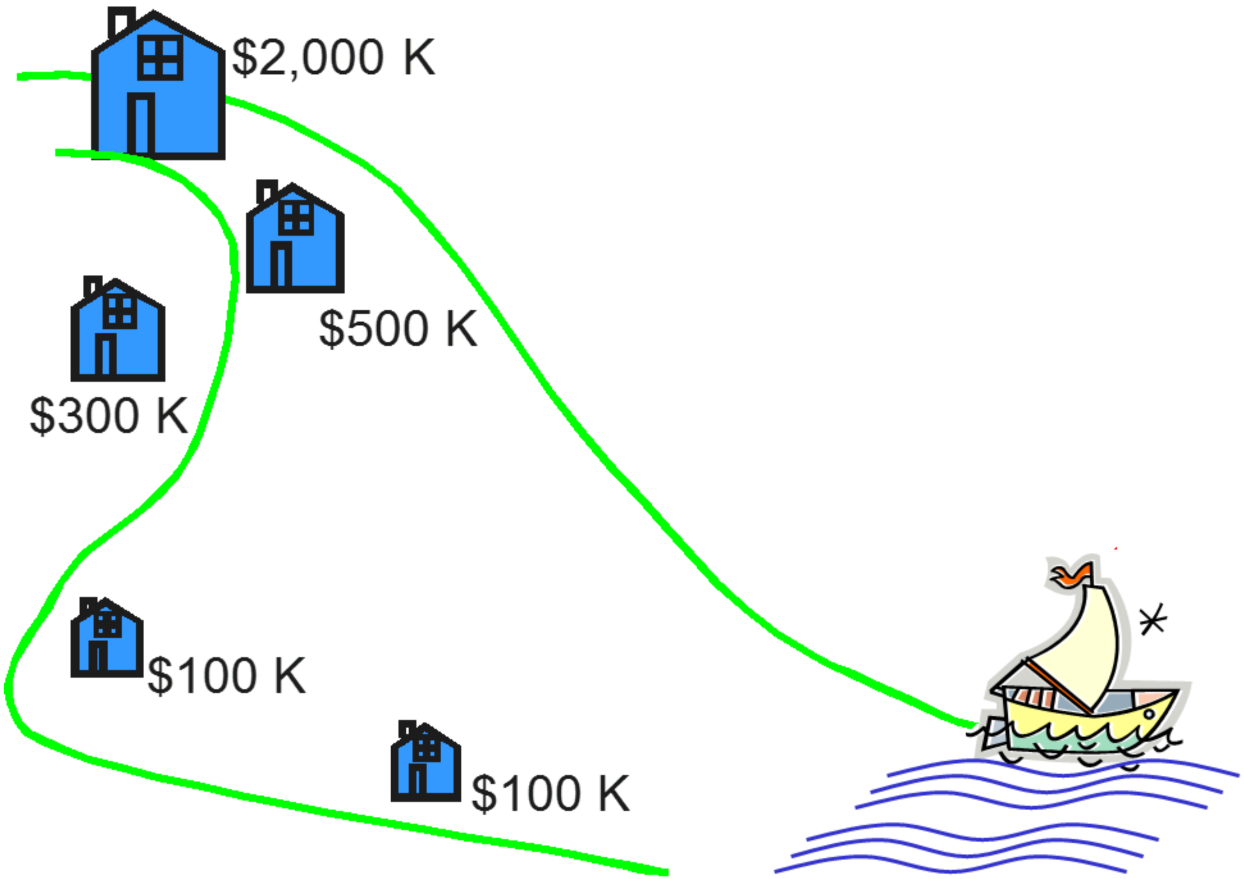


Review Example

- 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:

$$
\begin{array}{r}
\$2,000,000 \\
500,000 \\
300,000 \\
100,000 \\
100,000 \\
\hline
\text{Sum } 3,000,000
\end{array}
$$

- **Mean:** $\left(\frac{\$3,000,000}{5}\right)$

  = **$600,000**

- **Median:** middle value of ranked data

  = **$300,000**

- **Mode:** most frequent value

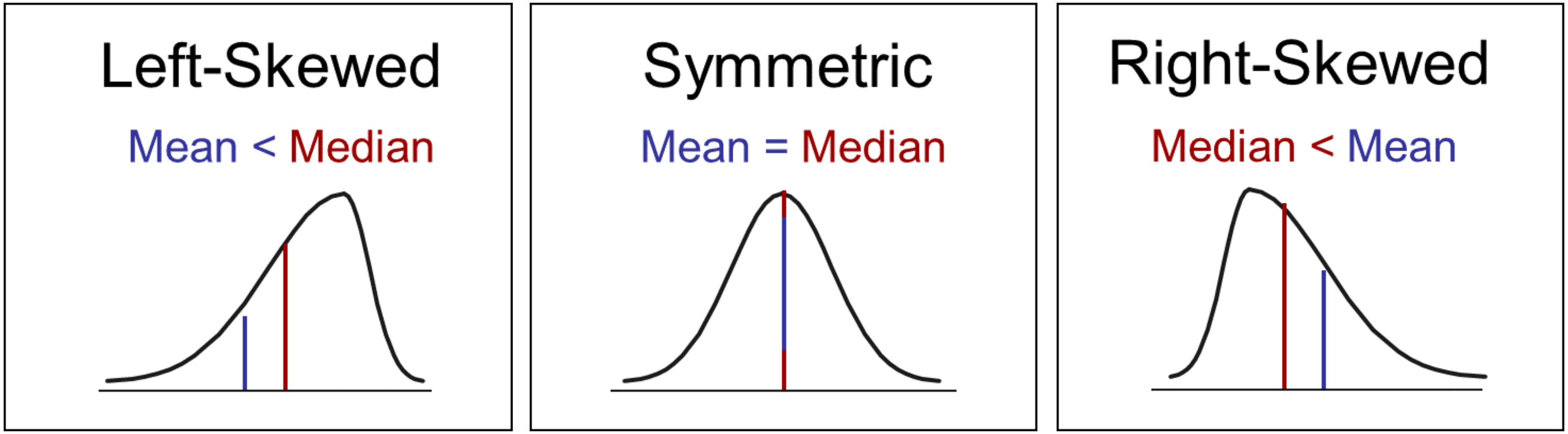
  = **$100,000**

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



Geometric Mean

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

$$\bar{x}_g = \sqrt[n]{(x_1 \times x_2 \times \cdots \times x_n)} = (x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}}$$

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

$$\bar{r}_g = (x_1 \times x_2 \times \ldots \times x_n)^{\frac{1}{n}} - 1$$

  - Where x_i is the rate of return in time period *i*

Example (1 of 2)

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:

$$X_1 = \$100,000 \quad X_2 = \$150,000 \quad X_3 = \$180,000$$

$$\underbrace{X_1 \to X_2}_{\text{50\% increase}} \qquad \underbrace{X_2 \to X_3}_{\text{20\% increase}}$$

What is the mean percentage return over time?

Example (2 of 2)

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

Arithmetic mean rate of return:

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

Misleading result

Geometric mean rate of return:

$$\bar{r}_g = (x_1 \times x_2)^{\frac{1}{n}} - 1$$

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

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

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 90th 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 $\left(\frac{P}{100}\right)(n+1)^{\text{th}}$ ordered position

Quartiles (1 of 2)

- 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: $Q_2 = 0.50(n+1)$
(the median position)

Third quartile position: $Q_3 = 0.75(n+1)$

where *n* is the number of observed values

Quartiles (2 of 2)

- 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 2^{nd} and 3^{rd} values,

so $Q_1 = 12.5$

Five-Number Summary

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

minimum

first quartile

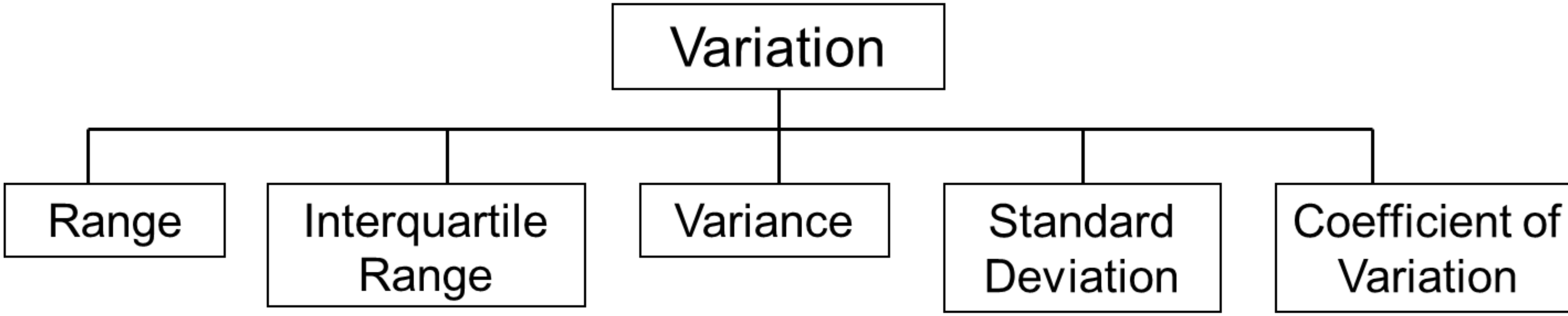
median

third quartile

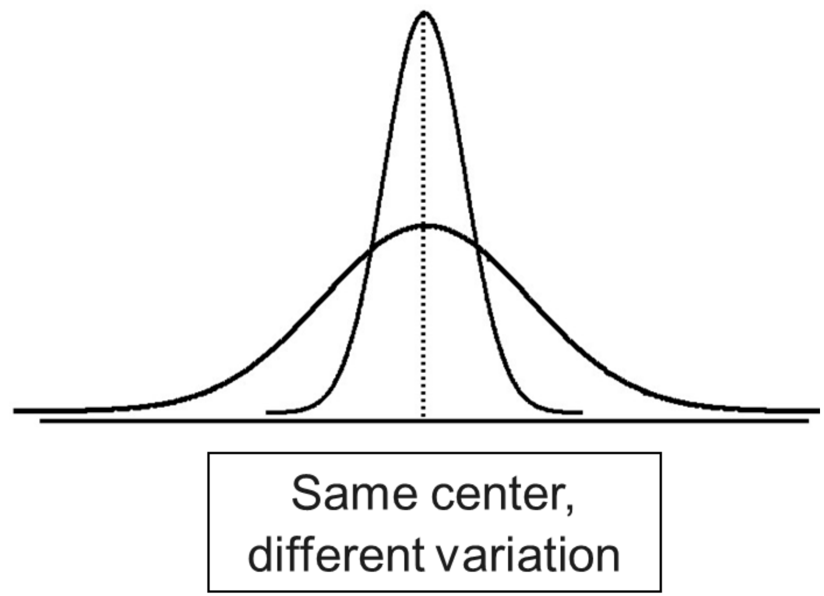
maximum

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

Section 2.2 Measures of Variability



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

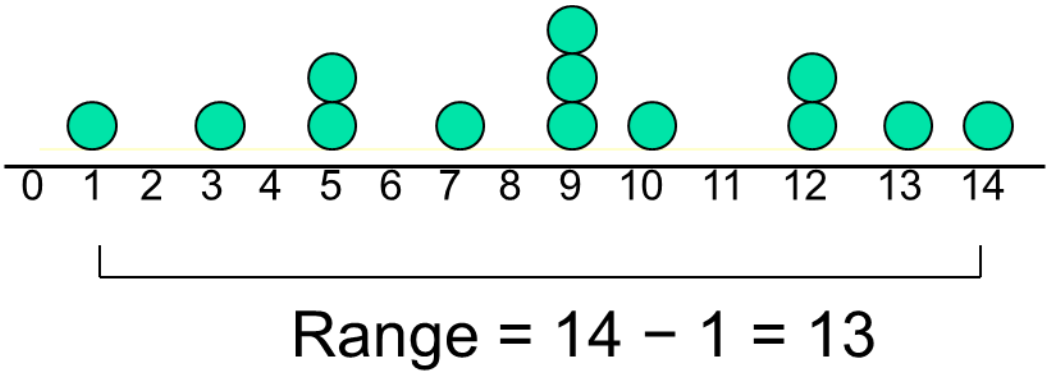


Range

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

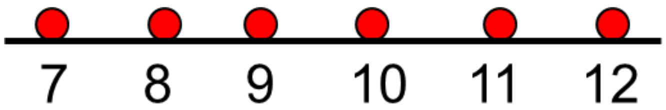
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:

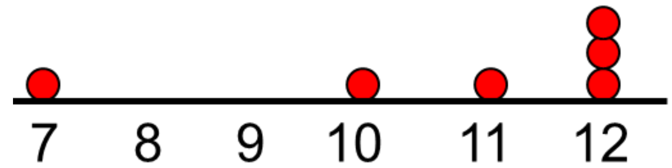


Disadvantages of the Range

- 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 (1 of 2)

- 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

$$\text{IQR} = Q_3 - Q_1$$

Interquartile Range (2 of 2)

- 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

$$\text{IQR} = Q_3 - Q_1$$

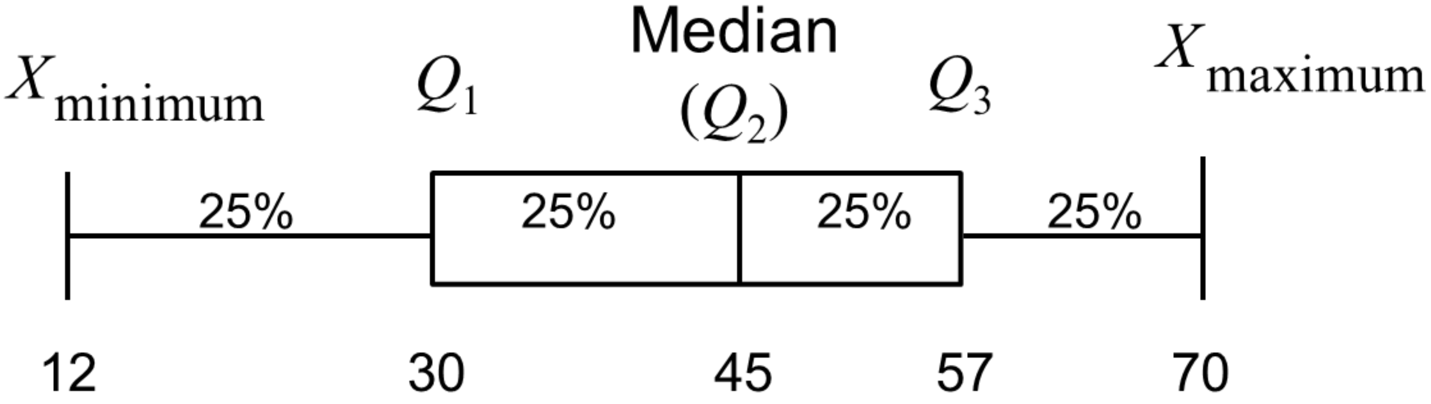
Box-and-Whisker Plot (1 of 2)

- 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 (2 of 2)

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

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

Sample Variance

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

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

Where $\overline{x}$ = arithmetic mean

n = sample size

$x_i = i^{\text{th}}$ 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:

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

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:

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

Calculation Example: Sample Standard Deviation

Sample Data $\left(x_i\right)$:

10	12	14	15	17	18	18	24

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

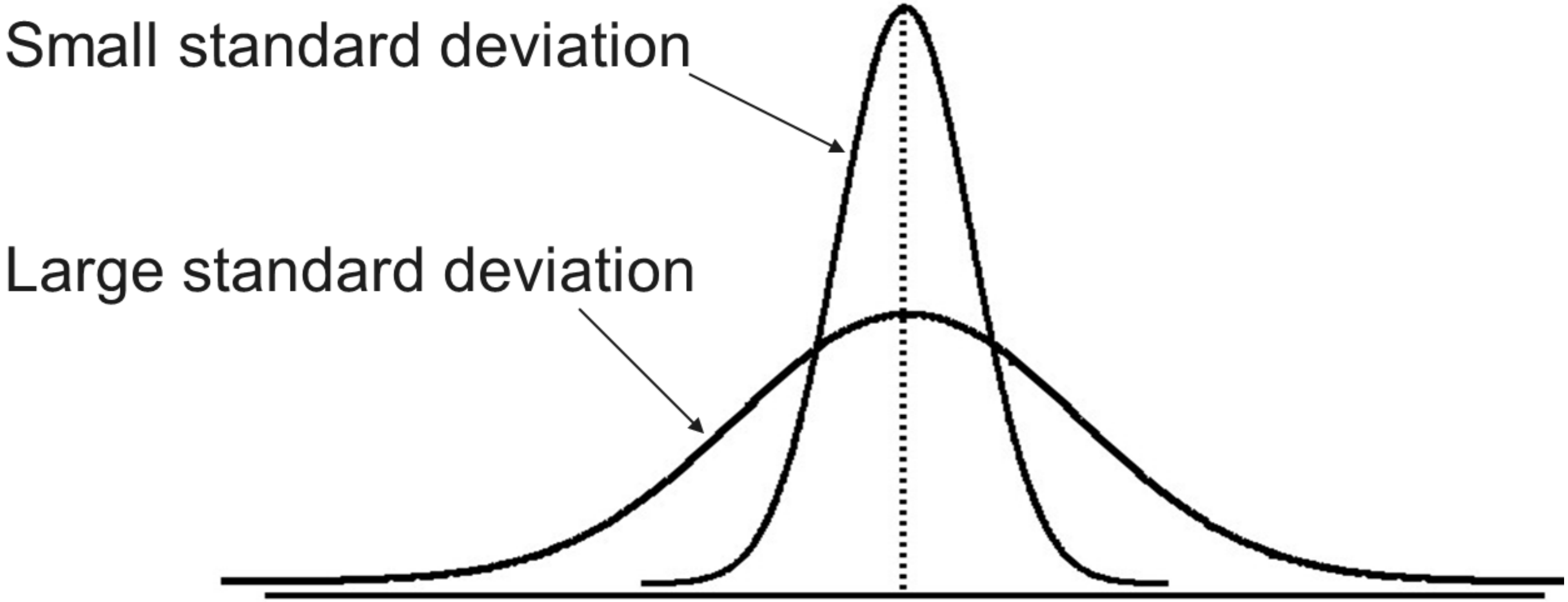
$$s = \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

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

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

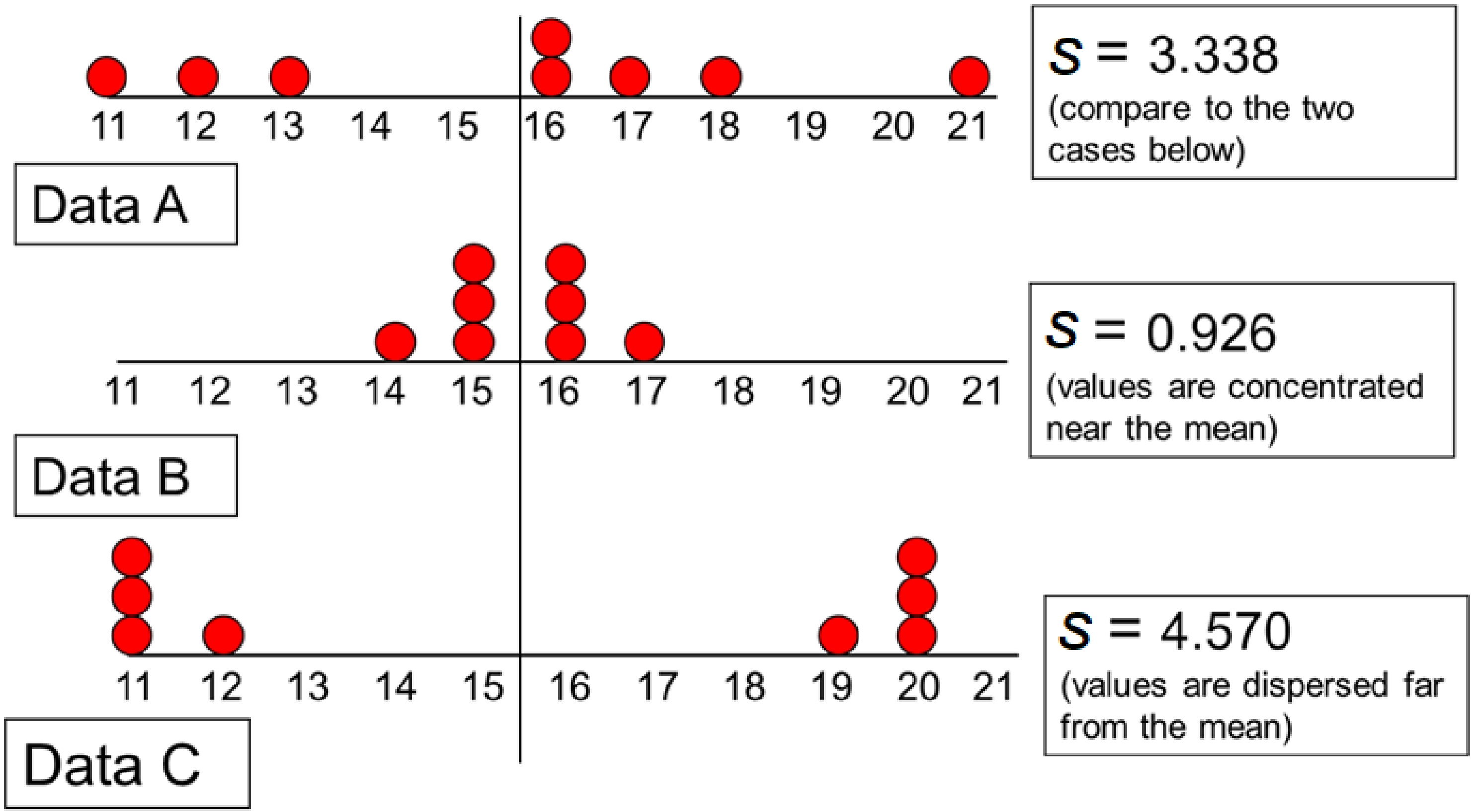
⇨ A measure of the "average" scatter around the mean

Measuring Variation



Comparing Standard Deviations

Mean = 15.5 for each data set



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)

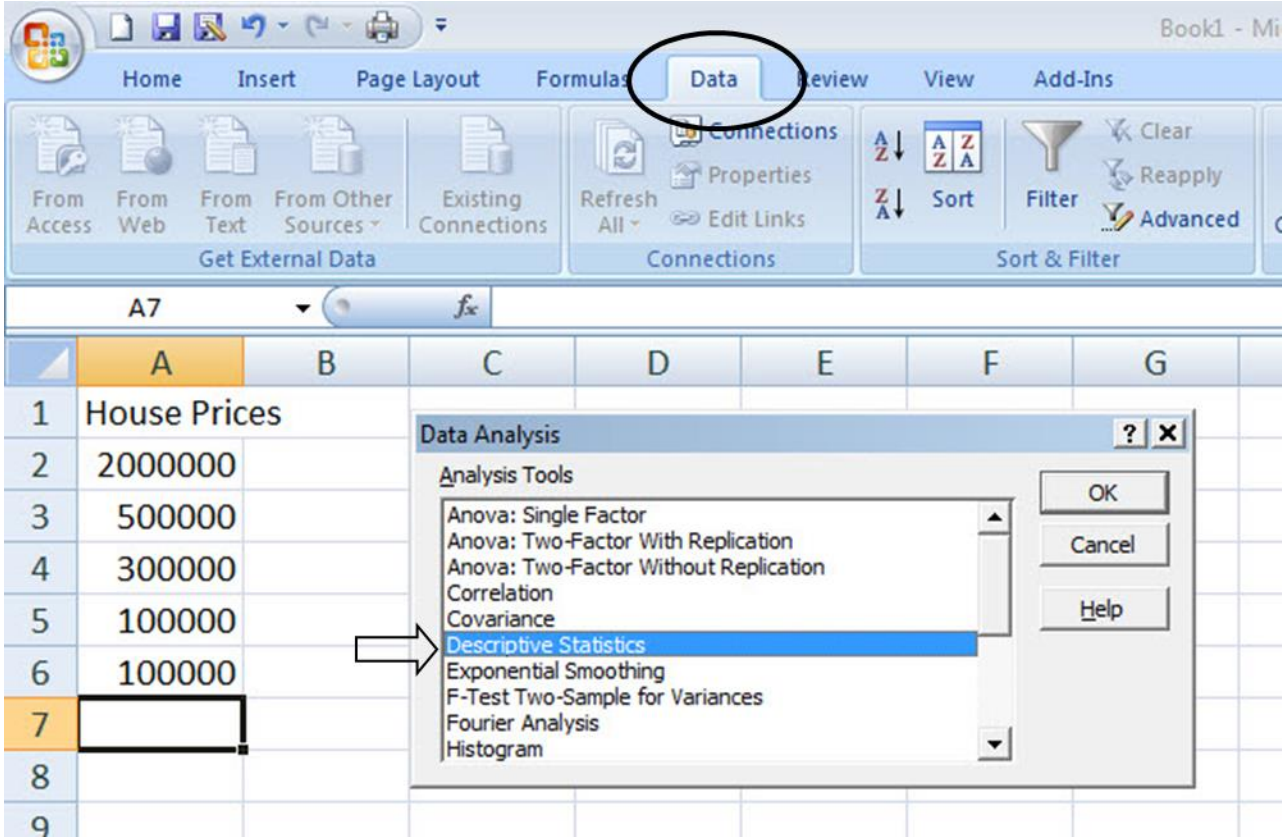
Using Microsoft Excel

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

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

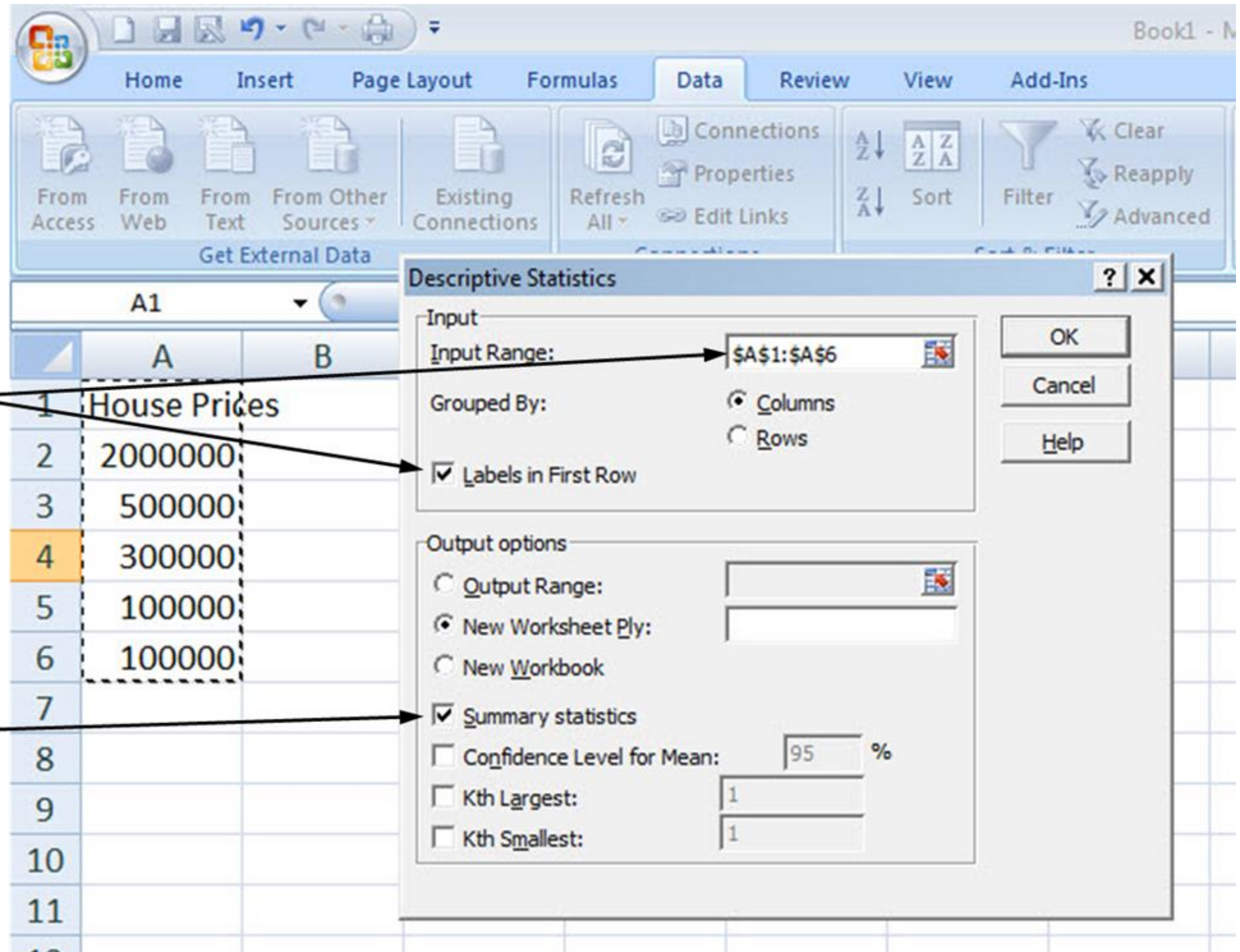
Using Excel (1 of 2)

- Select data/data analysis/descriptive statistics



Using Excel (2 of 2)

- 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:

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

Sample coefficient of variation:

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

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 (1 of 2)

- 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

$$100\left[1-\left(\frac{1}{k^2}\right)\right]\%$$

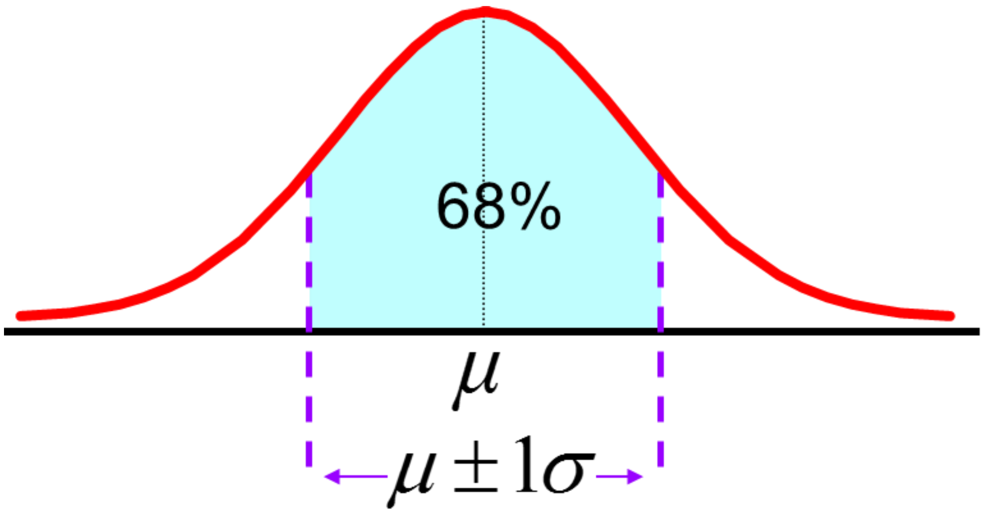
Chebychev's Theorem (2 of 2)

- Regardless of how the data are distributed, at least $\left(1-\frac{1}{k^2}\right)$ of the values will fall within *k* standard deviations of the mean (for *k* > 1)
  - Examples:

At least	within
$\left(1 - \frac{1}{1.5^2}\right) = 55.6\%$	……… $k = 1.5 \; (\mu \pm 1.5\sigma)$
$\left(1 - \frac{1}{2^2}\right) = 75\%$	…………. $k = 2 \quad (\mu \pm 2\sigma)$
$\left(1 - \frac{1}{3^2}\right) = 89\%$	…………. $k = 3 \quad (\mu \pm 3\sigma)$

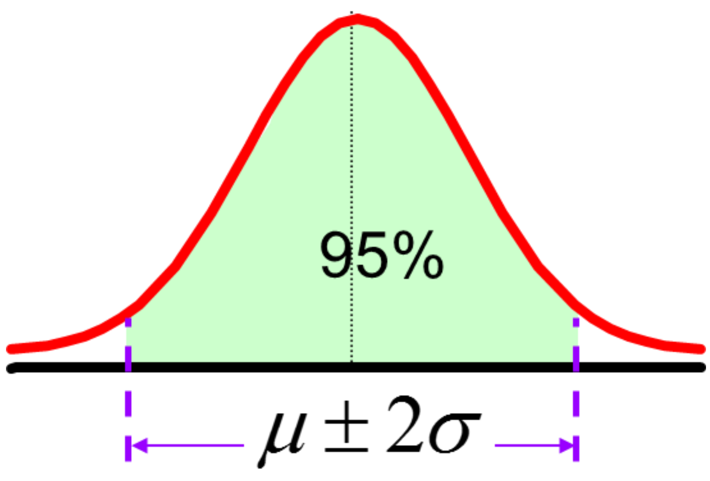
The Empirical Rule (1 of 2)

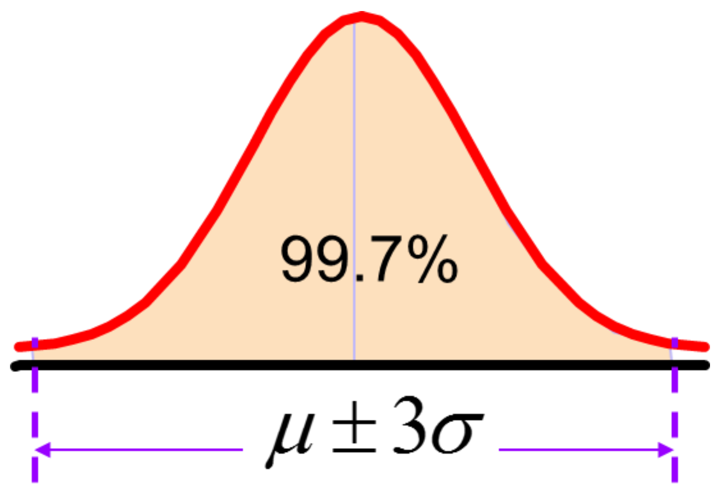
- If the data distribution is bell-shaped, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample



The Empirical Rule (2 of 2)

- $\mu \pm 2\sigma$ 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 (1 of 3)

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 (2 of 3)

- 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 (3 of 3)

- 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.

Section 2.3 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 (1 of 2)

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

- For a sample of *n* observations, the mean is

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

Approximations for Grouped Data (2 of 2)

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

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

Section 2.4 Measures of Relationships Between Variables

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:

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

- 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

Interpreting Covariance

- **Covariance** between two variables:

$\text{Cov}(x, y) > 0 \rightarrow$ *x* and *y* tend to move in the same direction

$\text{Cov}(x, y) < 0 \rightarrow$ *x* and *y* tend to move in opposite directions

$\text{Cov}(x, y) = 0 \rightarrow$ *x* and *y* are independent

Coefficient of Correlation

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

- Population correlation coefficient:

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

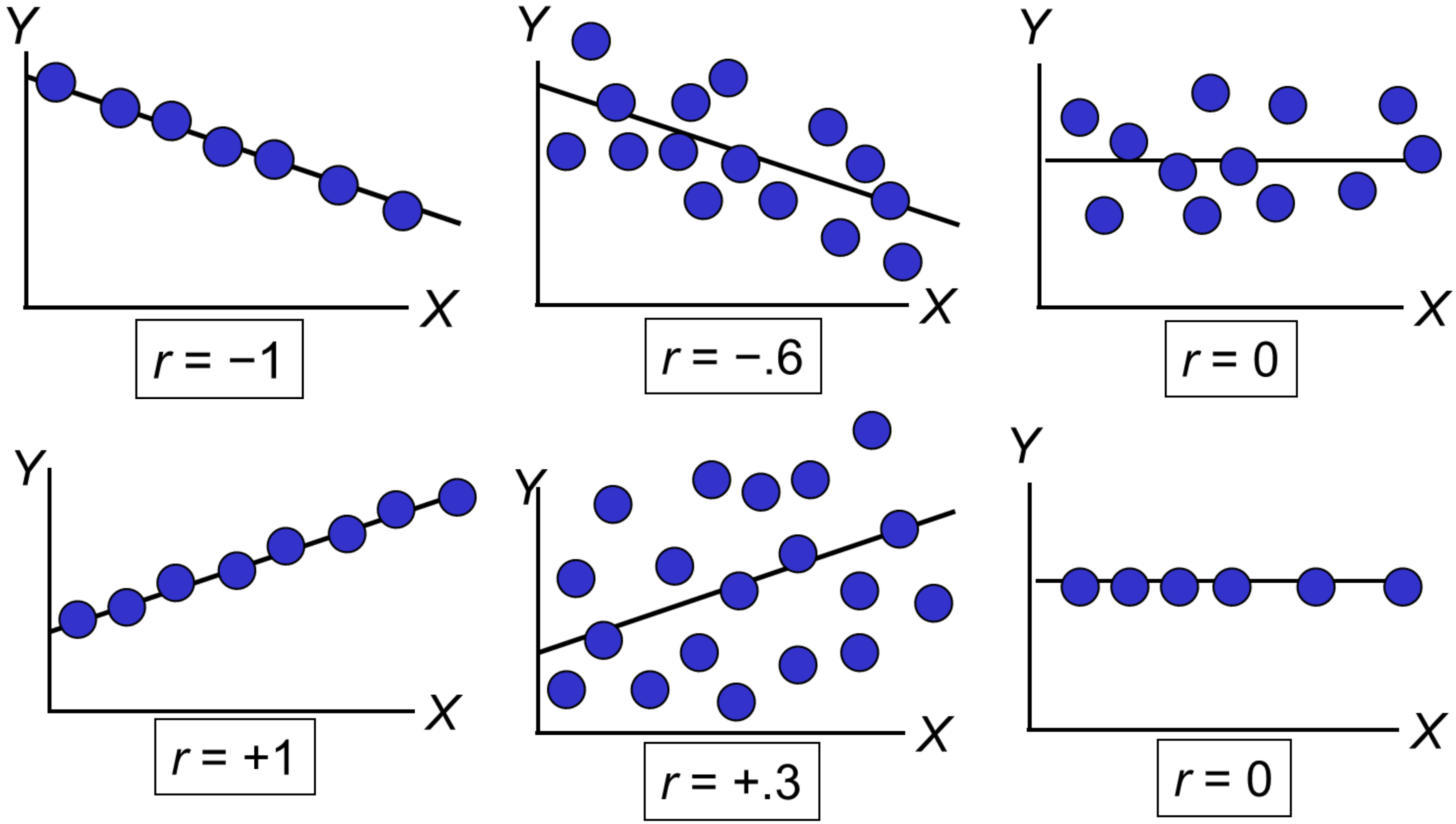
- Sample correlation coefficient:

$$r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

Features of Correlation Coefficient, *r*

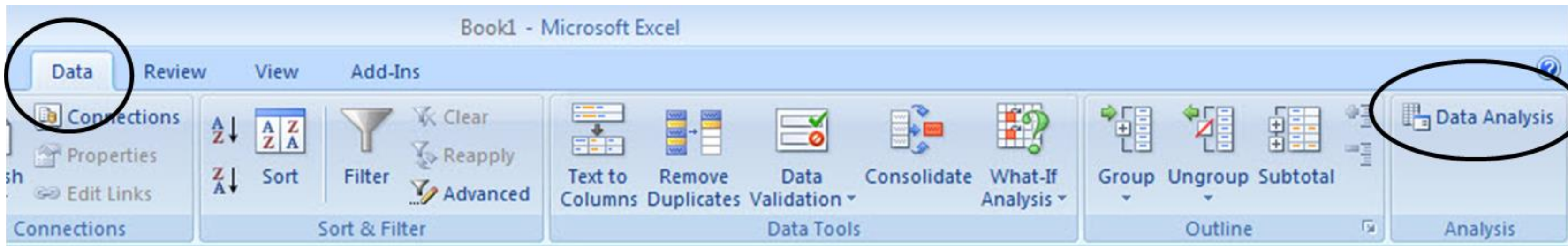
- Unit free
- Ranges between −1 and 1
- The closer to −1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

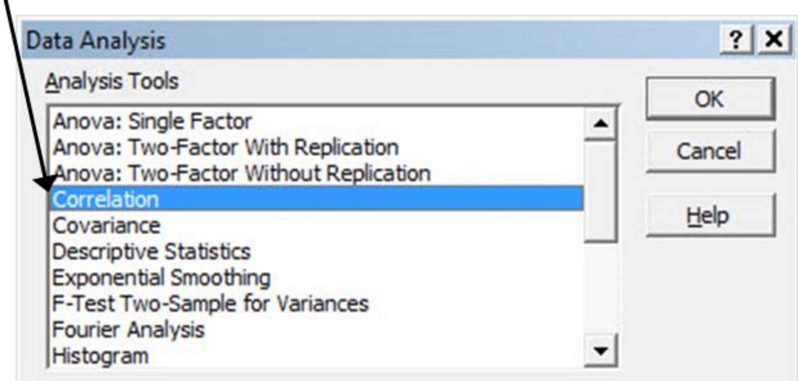


Using Excel to Find the Correlation Coefficient (1 of 2)

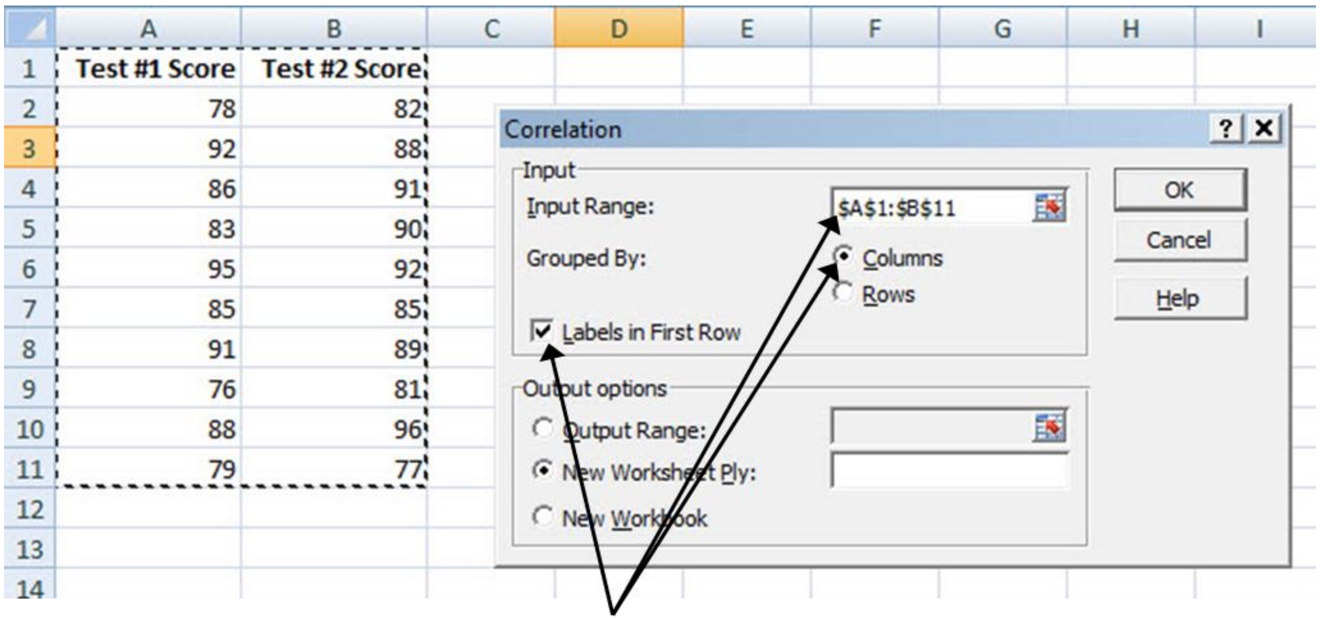
- Select Data/Data Analysis



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

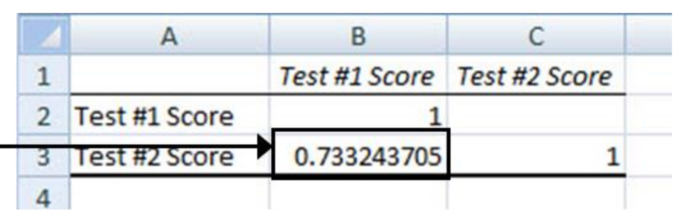


Using Excel to Find the Correlation Coefficient (2 of 2)



- Input data range and select appropriate options

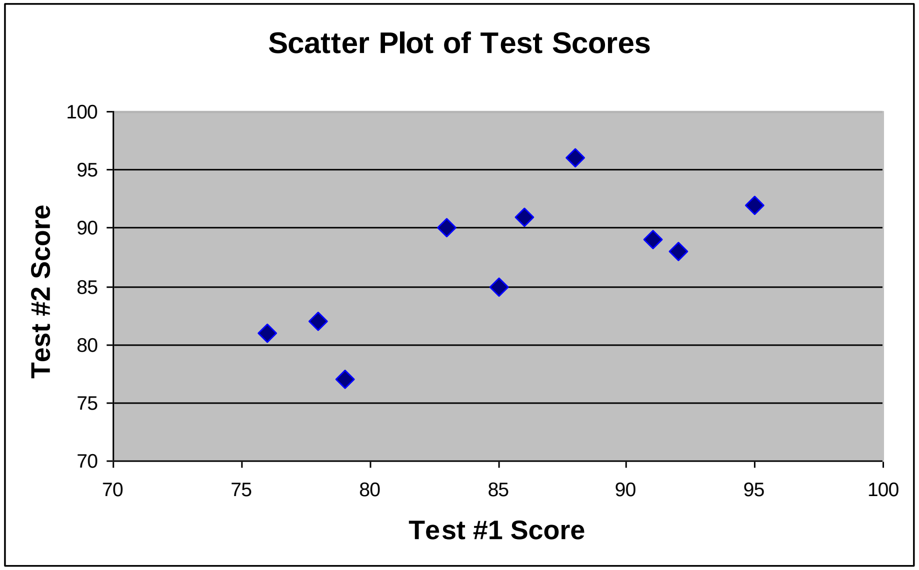
- Click OK to get output



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