IE 340/440

**PROCESS IMPROVEMENT
THROUGH PLANNED EXPERIMENTATION**

**IE 406
Simulation**

Numerical Descriptive Measures

Dr. Xueping Li
University of Tennessee

Chapter Topics

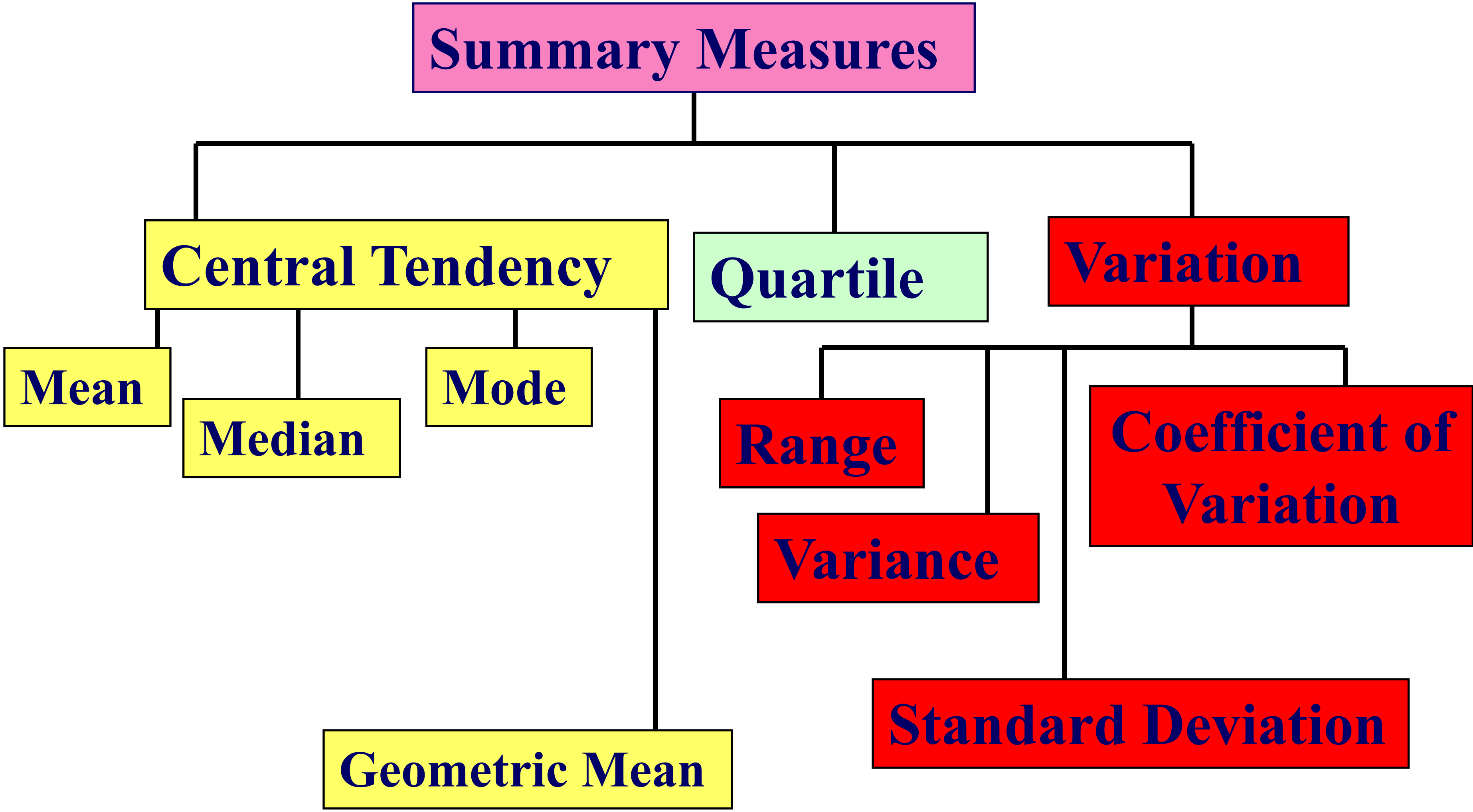
- Measures of Central Tendency
  - Mean, Median, Mode, Geometric Mean
- Quartile
- Measure of Variation
  - Range, Interquartile Range, Variance and Standard Deviation, Coefficient of Variation
- Shape
  - Symmetric, Skewed, Using Box-and-Whisker Plots

Chapter Topics

(continued)

- The Empirical Rule and the Bienayme-Chebyshev Rule
- Coefficient of Correlation
- Pitfalls in Numerical Descriptive Measures and Ethical Issues

Summary Measures



Measures of Central Tendency

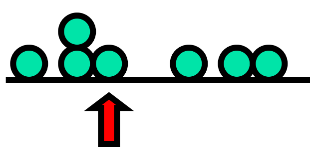
Central Tendency

- **Mean**

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

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

- **Median**



- **Mode**

- **Geometric Mean**

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

Mean (Arithmetic Mean)

- Mean (Arithmetic Mean) of Data Values
  - Sample mean

Sample Size

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

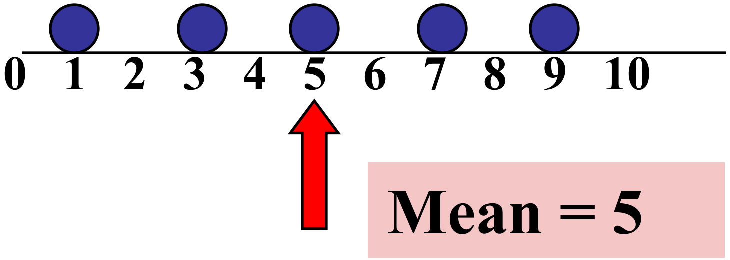
  - Population mean

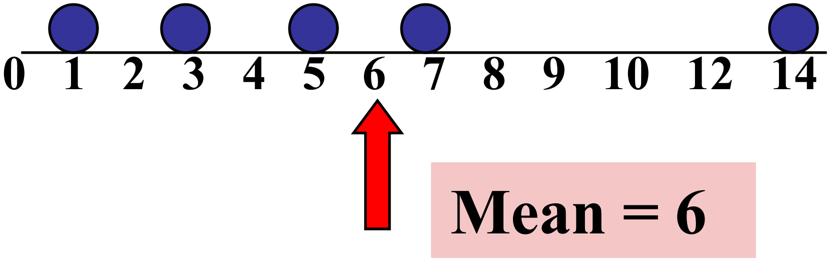
Population Size

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

Mean (Arithmetic Mean)
(continued)

- The Most Common Measure of Central Tendency
- Affected by Extreme Values (Outliers)





Mean (Arithmetic Mean)

(continued)

- Approximating the Arithmetic Mean
  - Used when raw data are not available
  - $$\bar{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

$n =$ sample size

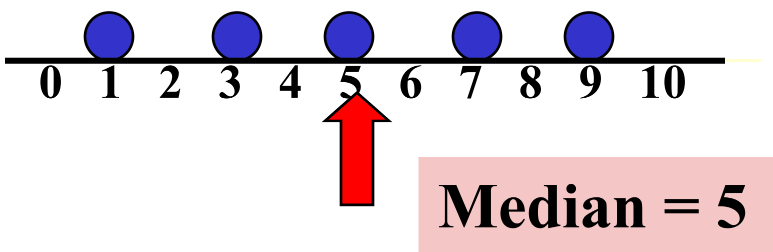
$c =$ number of classes in the frequency distribution

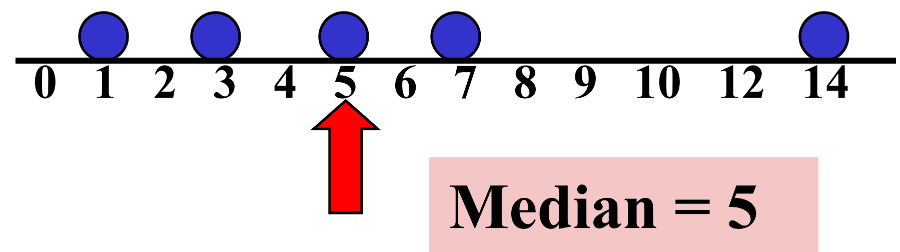
$m_j =$ midpoint of the jth class

$f_j =$ frequencies of the jth class

Median

- Robust Measure of Central Tendency
- Not Affected by Extreme Values

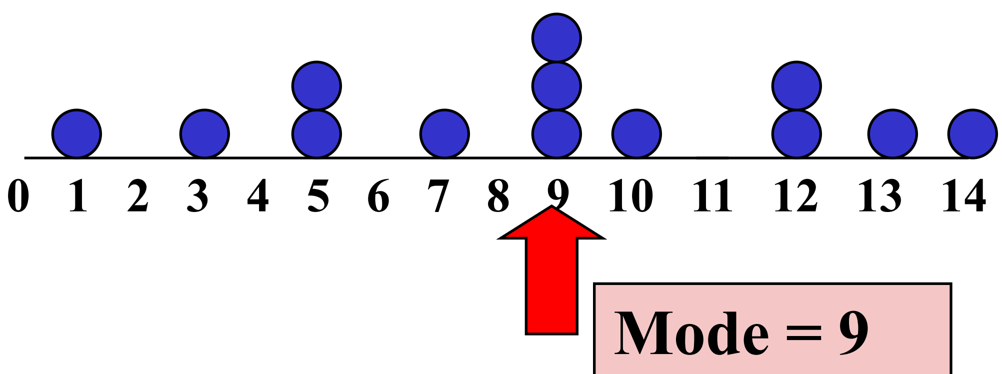


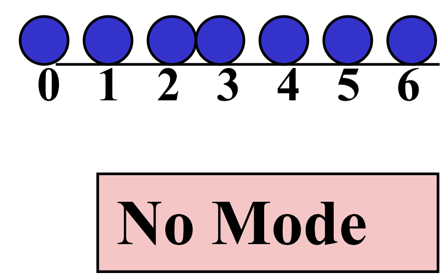


- In an Ordered Array, the Median is the 'Middle' Number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the 2 middle numbers

Mode

- A Measure of Central Tendency
- Value that Occurs Most Often
- Not Affected by Extreme Values
- There May Not Be a Mode
- There May Be Several Modes
- Used for Either Numerical or Categorical Data





Geometric Mean

- Useful in the Measure of Rate of Change of a Variable Over Time

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

- Geometric Mean Rate of Return
  - Measures the status of an investment over time

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

Example

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

$$R_1 = -0.5 \text{ (or } -50\%) \qquad R_2 = 1 \text{ (or } 100\% \text{)}$$

Average rate of return:

$$\bar{R} = \frac{(-0.5)+(1)}{2} = 0.25 \text{ (or } 25\%)$$

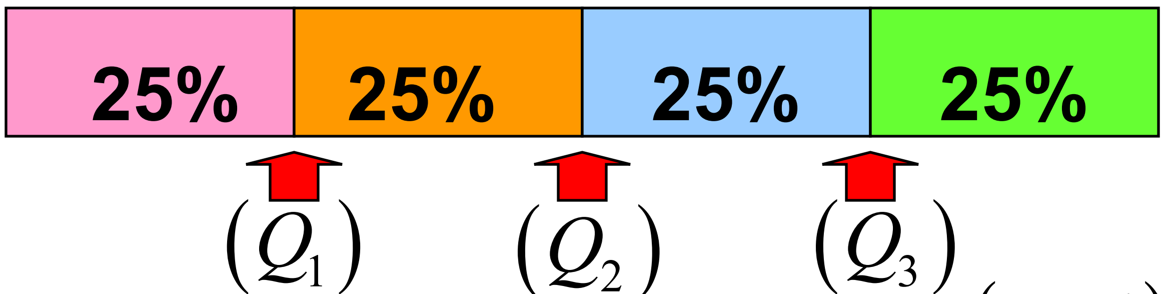
Geometric rate of return:

$$\bar{R}_G = \left[(1-0.5)\times(1+1)\right]^{1/2} - 1$$

$$= \left[(0.5)\times(2)\right]^{1/2} - 1 = 1^{1/2} - 1 = 0 \text{ (or } 0\%)$$

Quartiles

- Split Ordered Data into 4 Quarters



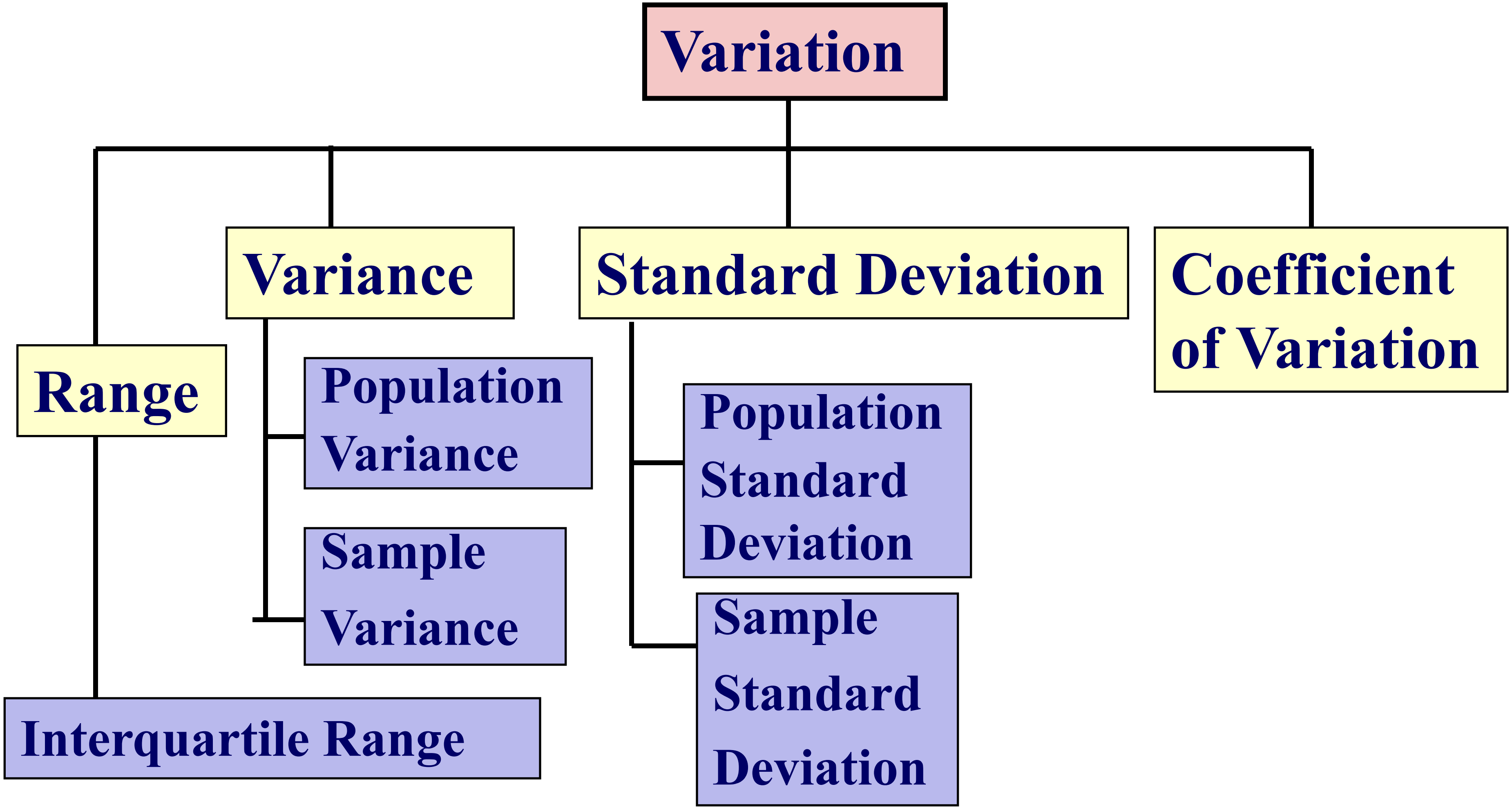
- Position of i-th Quartile $\left(Q_i\right) = \dfrac{i\left(n+1\right)}{4}$

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

$$\text{Position of } Q_1 = \frac{1\left(9+1\right)}{4} = 2.5 \qquad Q_1 = \frac{\left(12+13\right)}{2} = 12.5$$

- Q_1 and Q_3 are Measures of Noncentral Location
- Q_2 = Median, a Measure of Central Tendency

Measures of Variation



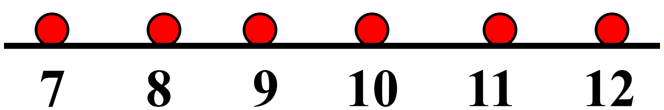
Range

- Measure of Variation
- Difference between the Largest and the Smallest Observations:

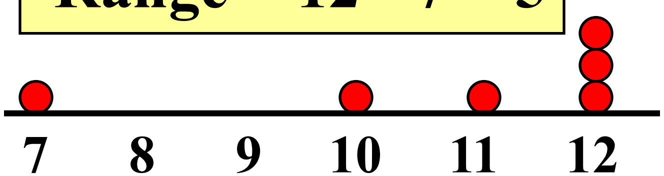
$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$

- Ignores How Data are Distributed

Range = 12 - 7 = 5



Range = 12 - 7 = 5



Interquartile Range

- Measure of Variation
- Also Known as Midspread
  - Spread in the middle 50%
- Difference between the First and Third Quartiles

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

$$\text{Interquartile Range} = Q_3 - Q_1 = 17.5 - 12.5 = 5$$

- Not Affected by Extreme Values

Variance

- Important Measure of Variation
- Shows Variation about the Mean
  - Sample Variance:

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

  - Population Variance:

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

Standard Deviation

- Most Important Measure of Variation
- Shows Variation about the Mean
- Has the Same Units as the Original Data
  - Sample Standard Deviation:

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

  - Population Standard Deviation:

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

Standard Deviation

(continued)

- Approximating the Standard Deviation
  - Used when the raw data are not available and the only source of data is a frequency distribution
  - $$S = \sqrt{\frac{\sum_{j=1}^{c} \left(m_j - \bar{X}\right)^2 f_j}{n-1}}$$

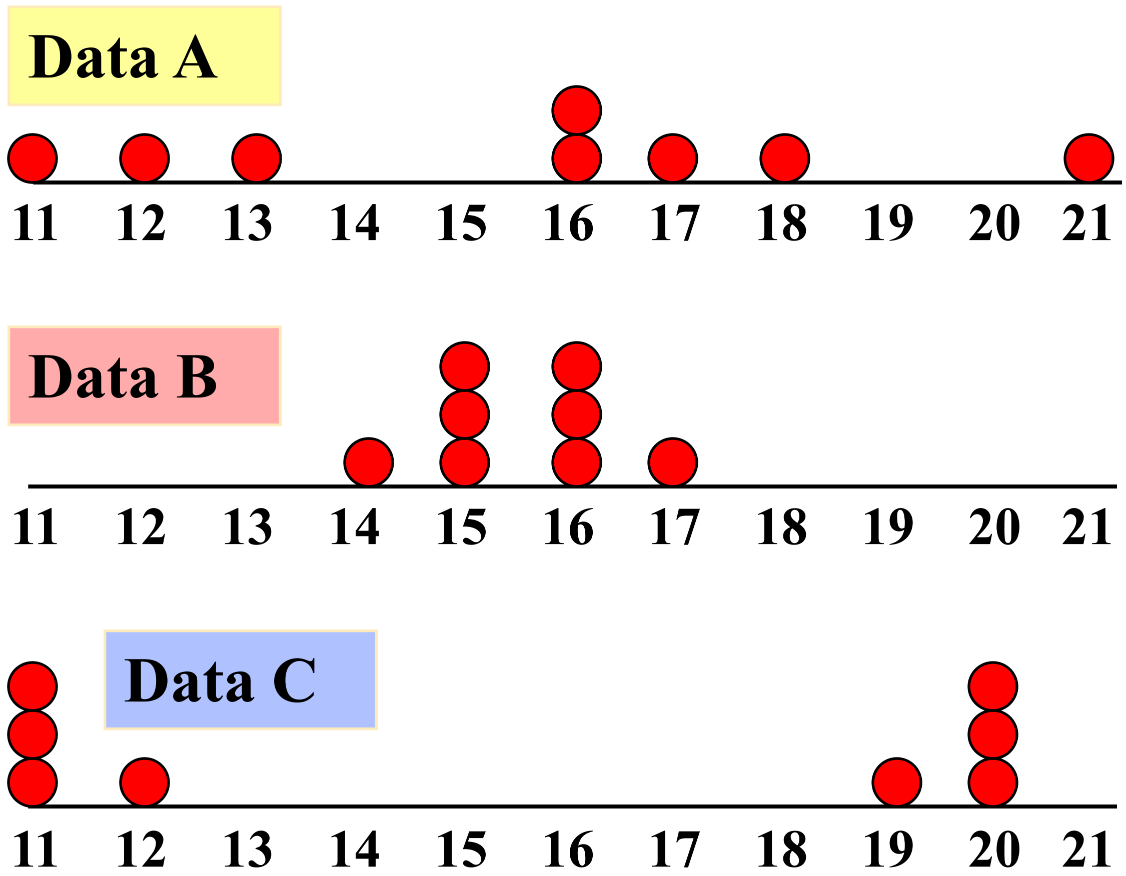
$n =$ sample size

$c =$ number of classes in the frequency distribution

$m_j =$ midpoint of the jth class

$f_j =$ frequencies of the jth class

Comparing Standard Deviations



Mean = 15.5
s = 3.338

Mean = 15.5
s = .9258

Mean = 15.5
s = 4.57

Coefficient of Variation

- Measure of Relative Variation
- Always in Percentage (%)
- Shows Variation Relative to the Mean
- Used to Compare Two or More Sets of Data Measured in Different Units
- $$CV = \left(\frac{S}{\overline{X}} \right) 100\%$$
- Sensitive to Outliers

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $2
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5
- Coefficient of Variation:
  - Stock A:

$$CV = \left(\frac{S}{\overline{X}}\right)100\% = \left(\frac{\$2}{\$50}\right)100\% = 4\%$$

  - Stock B:

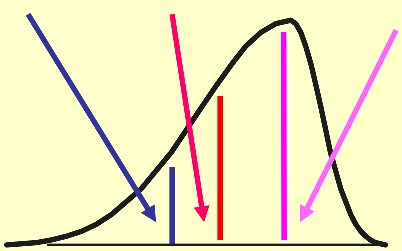
$$CV = \left(\frac{S}{\overline{X}}\right)100\% = \left(\frac{\$5}{\$100}\right)100\% = 5\%$$

Shape of a Distribution

- Describe How Data are Distributed
- Measures of Shape
  - Symmetric or skewed

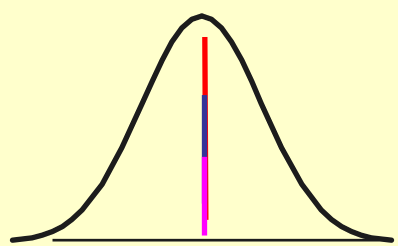
Left-Skewed

Mean < Median < Mode



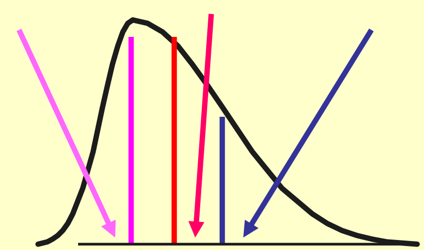
Symmetric

Mean = Median =Mode



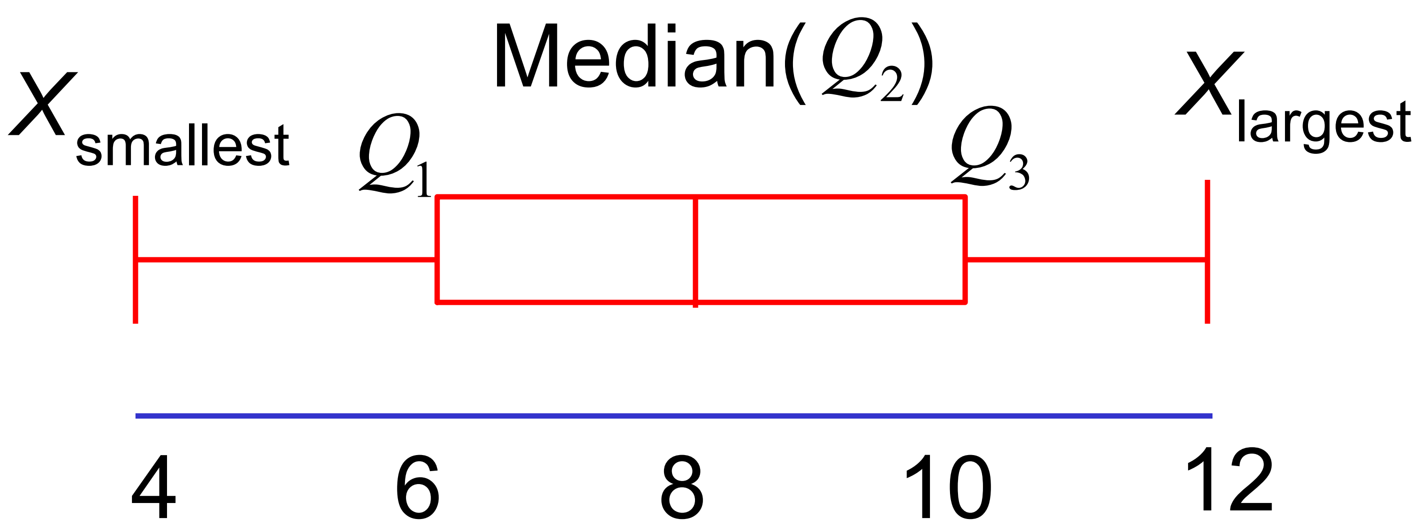
Right-Skewed

Mode < Median < Mean

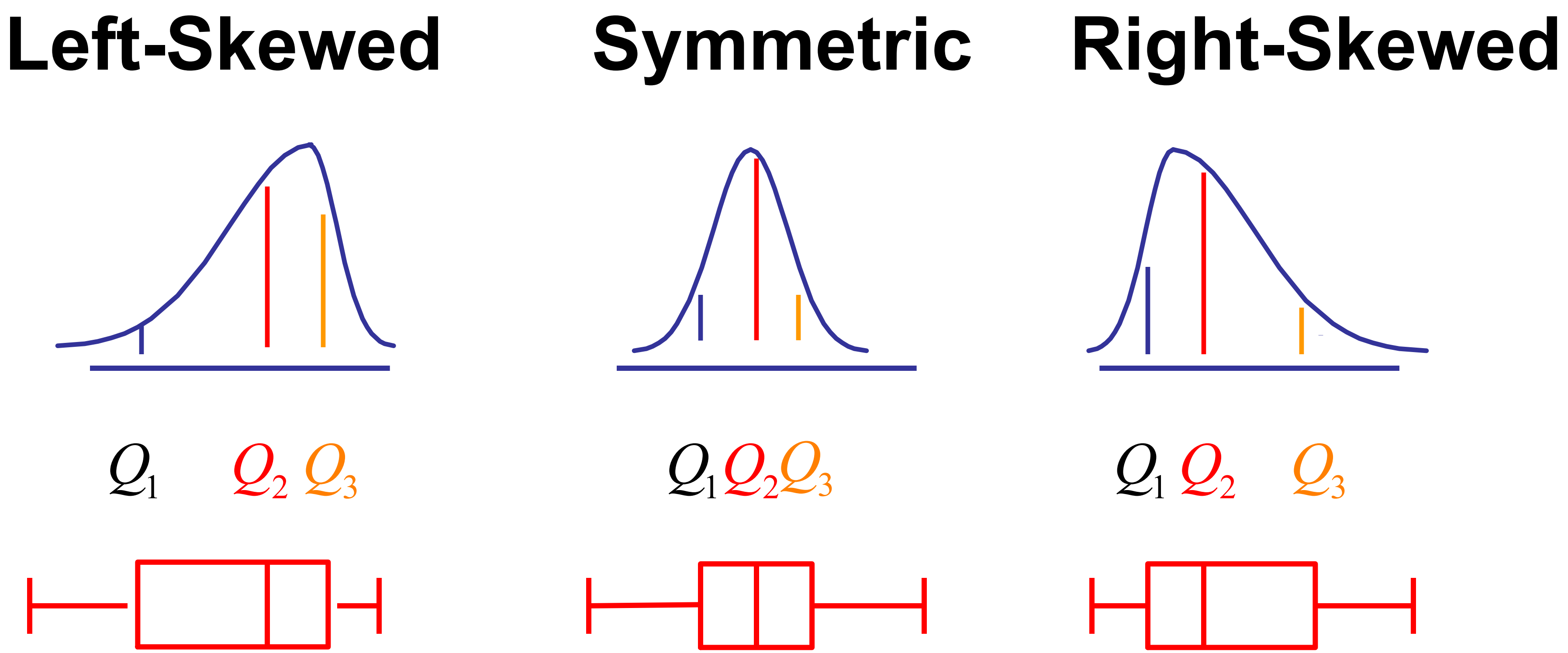


Exploratory Data Analysis

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



Distribution Shape & Box-and-Whisker



The Empirical Rule

- For Most Data Sets, Roughly 68% of the Observations Fall Within 1 Standard Deviation Around the Mean
- Roughly 95% of the Observations Fall Within 2 Standard Deviations Around the Mean
- Roughly 99.7% of the Observations Fall Within 3 Standard Deviations Around the Mean

The Bienayme-Chebyshev Rule

- The Percentage of Observations Contained Within Distances of *k* Standard Deviations Around the Mean Must Be at Least $\left(1-1/k^2\right)100\%$
  - Applies regardless of the shape of the data set
  - At least 75% of the observations must be contained within distances of 2 standard deviations around the mean
  - At least 88.89% of the observations must be contained within distances of 3 standard deviations around the mean
  - At least 93.75% of the observations must be contained within distances of 4 standard deviations around the mean

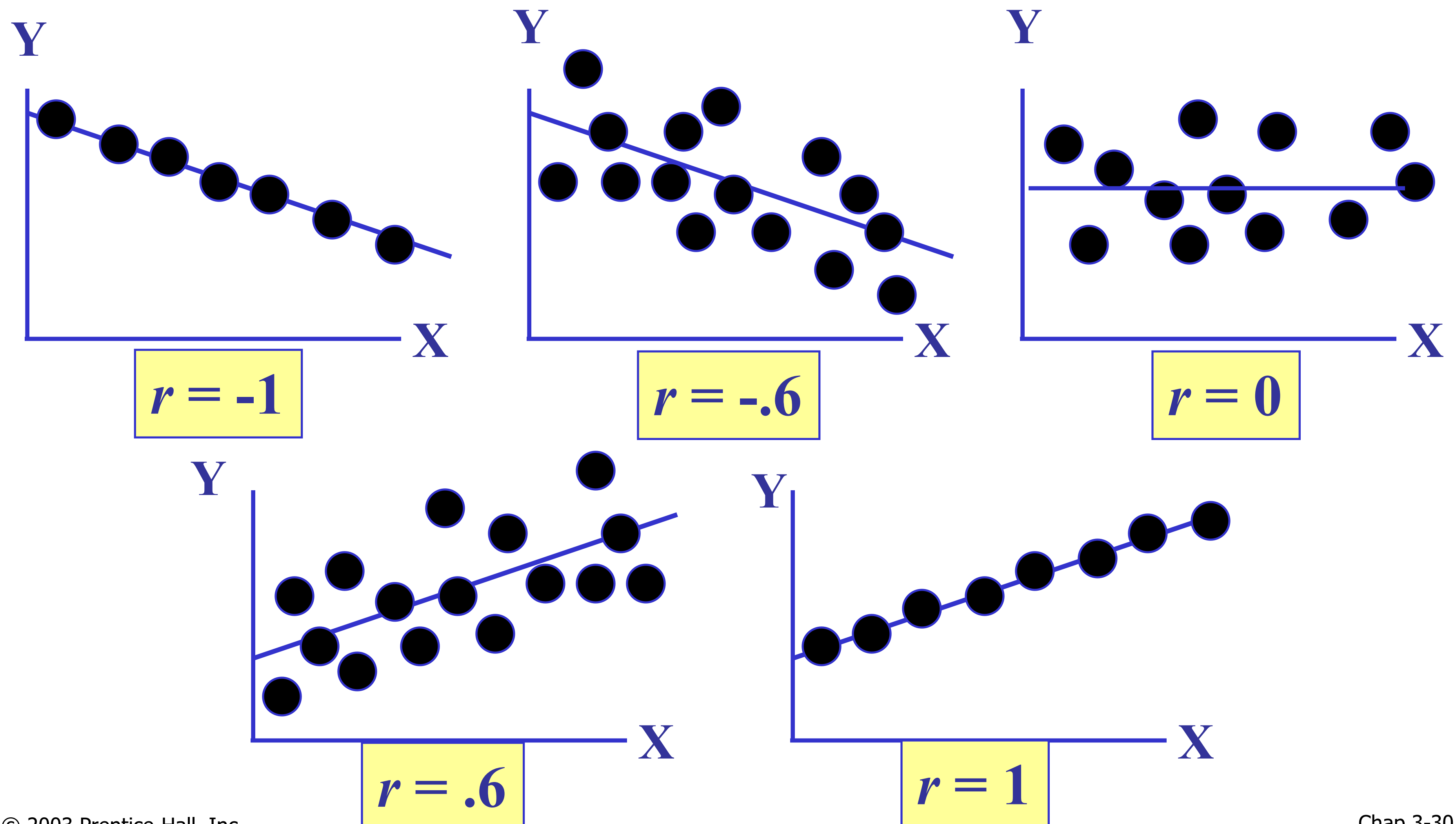
Coefficient of Correlation

- Measures the Strength of the Linear Relationship between 2 Quantitative Variables
- $r = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$

Features of Correlation Coefficient

- 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



Pitfalls in Numerical Descriptive Measures and Ethical Issues

- Data Analysis is Objective
  - Should report the summary measures that best meet the assumptions about the data set
- Data Interpretation is Subjective
  - Should be done in a fair, neutral and clear manner
- Ethical Issues
  - Should document both good and bad results
  - Presentation should be fair, objective and neutral
  - Should not use inappropriate summary measures to distort the 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, Using Box-and-Whisker Plots

Chapter Summary

(continued)

- Described the Empirical Rule and the Bienayme-Chebyshev Rule
- Discussed Correlation Coefficient
- Addressed Pitfalls in Numerical Descriptive Measures and Ethical Issues