

Final Project-2

Lecture 5

PMG497
L5- Statistical Analysis- P1



Session Outline

- Chapter 1: The Where, why, and How of Data Collection
- Chapter 2 : Graphs, Charts, and Tables – Describing Your Data
- Chapter 3 : Describing Data Using Numerical Measures
- Chapter 9: Introduction to Hypothesis Testing
- Chapter 14: Introduction to Linear Regression and Correlation Analysis
- Chapter 15: Multiple Regression Analysis and Model Building

Chapter 1-
The Where, why, and How
of Data Collection

1



Chapter Goals

After completing this chapter, you should be able to:

- Describe key data collection methods
- Know key definitions:
 - ◆ Population vs. Sample
 - ◆ Primary vs. Secondary data types
 - ◆ Qualitative vs. Quantitative data
 - ◆ Time Series vs. Cross-Sectional data
- Explain the difference between descriptive and inferential statistics
- Describe different sampling methods

Tools of Business Statistics

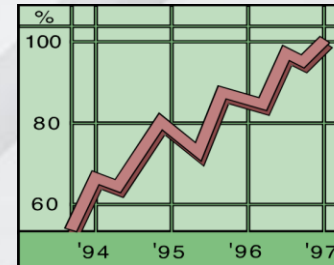
- **Descriptive statistics**
 - Collecting, presenting, and describing data
- **Inferential statistics**
 - Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics

- **Collect data**
 - e.g., Survey, Observation, Experiments



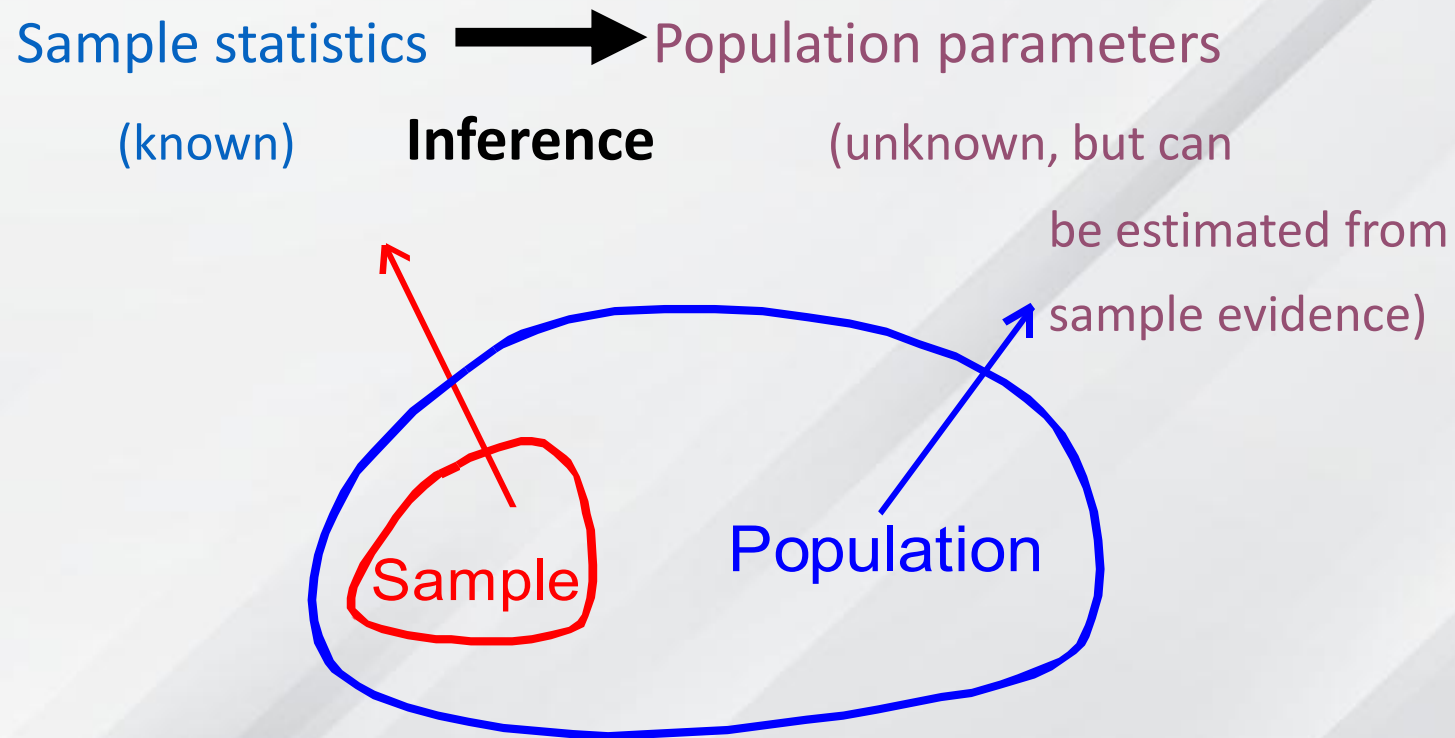
- **Present data**
 - e.g., Charts and graphs



- **Characterize data**
 - e.g., Sample mean =
$$\frac{\sum x_i}{n}$$

Inferential Statistics

- Making statements about a population by examining sample results



Inferential Statistics

Drawing conclusions and/or making decisions concerning a **population based on **sample** results.**

- **Estimation**

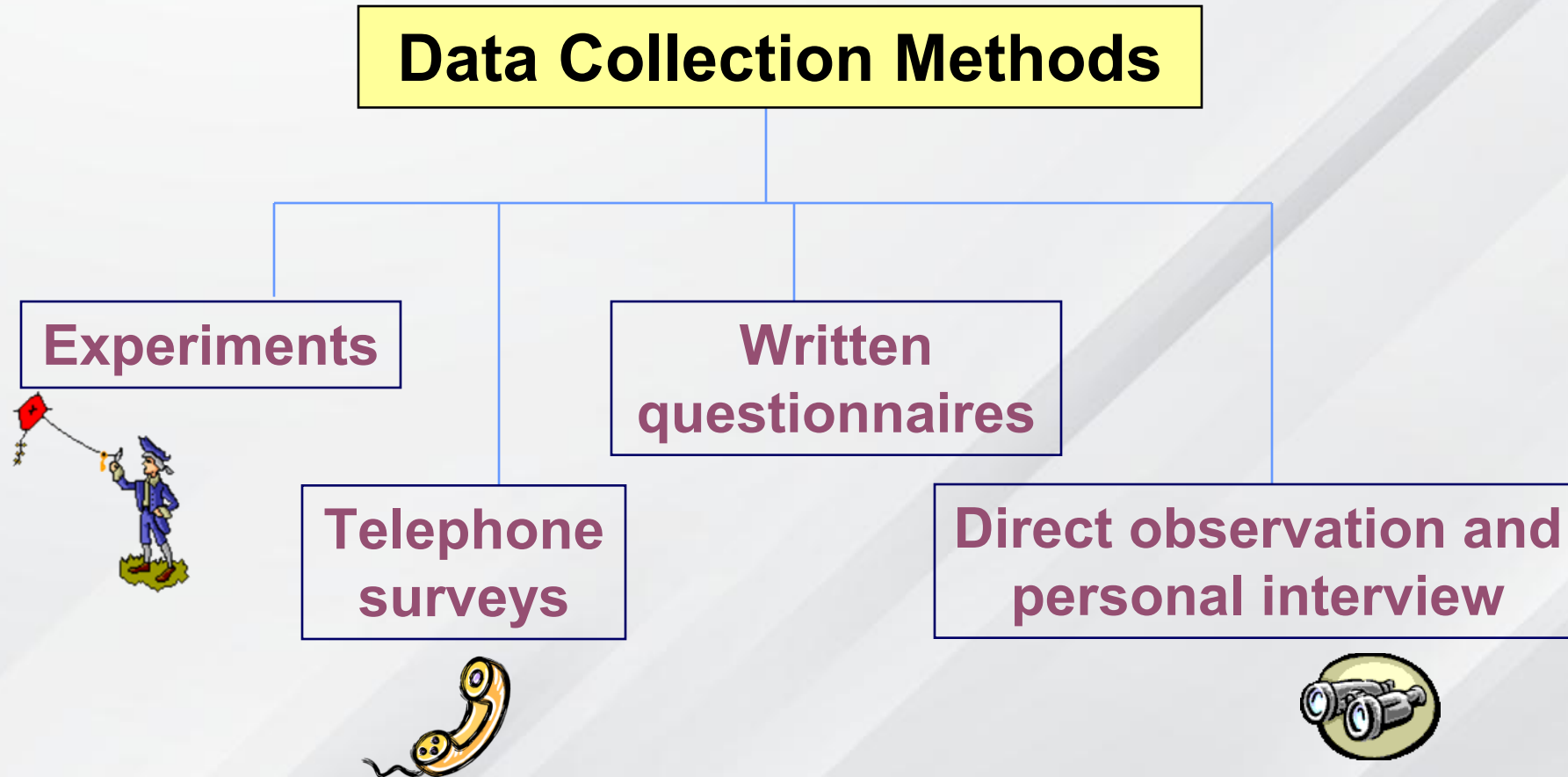
- e.g., Estimate the population mean weight using the sample mean weight

- **Hypothesis Testing**

- e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds



Tools for Collecting Data



Survey Design Steps

- Define the issue
 - what are the purpose and objectives of the survey?
- Define the population of interest
- Develop survey questions
 - make questions clear and unambiguous
 - use universally-accepted definitions
 - limit the number of questions

Survey Design Steps

(continued)

- Pre-test the survey
 - pilot test with a small group of participants
 - assess clarity and length
- Determine the sample size and sampling method
- Select sample and administer the survey

Types of Questions

- **Closed-end Questions**

- Select from a short list of defined choices

Example:

Major: ☐ business ☐ liberal arts
☐ science ☐ other

- **Open-end Questions**

- Respondents are free to respond with any value, words, or statement

Example: What did you like best about this course?

- **Demographic Questions**

- Questions about the respondents' personal characteristics

Example: Gender: ☐ Female ☐ Male

Populations and Samples

- A **Population** is the set of all items or individuals of interest

- **Examples:**
 - All likely voters in the next election
 - All parts produced today
 - All sales receipts for November

- A **Sample** is a subset of the population

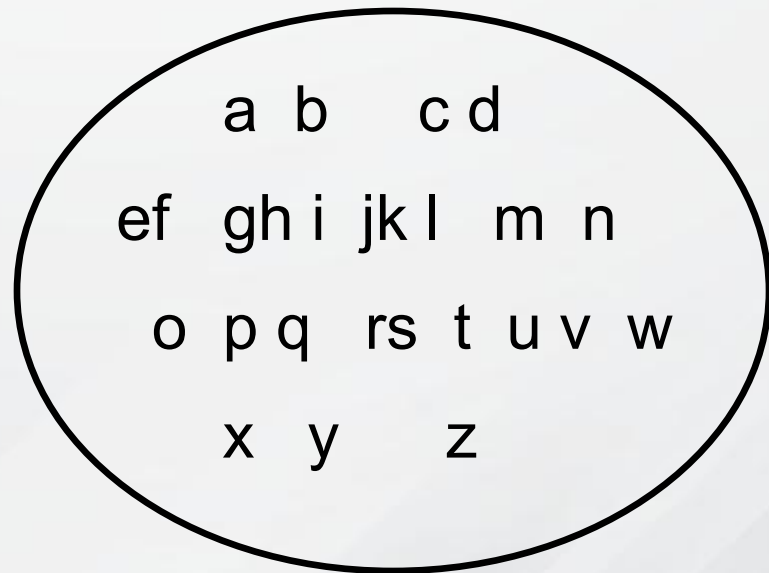
- **Examples:**
 - 1000 voters selected at random for interview
 - A few parts selected for destructive testing
 - Every 100th receipt selected for audit

Key Definitions

- A **population** is the entire collection of things under consideration
 - A **parameter** is a summary measure computed to describe a characteristic of the population
- A **sample** is a portion of the population selected for analysis
 - A **statistic** is a summary measure computed to describe a characteristic of the sample

Population vs. Sample

Population



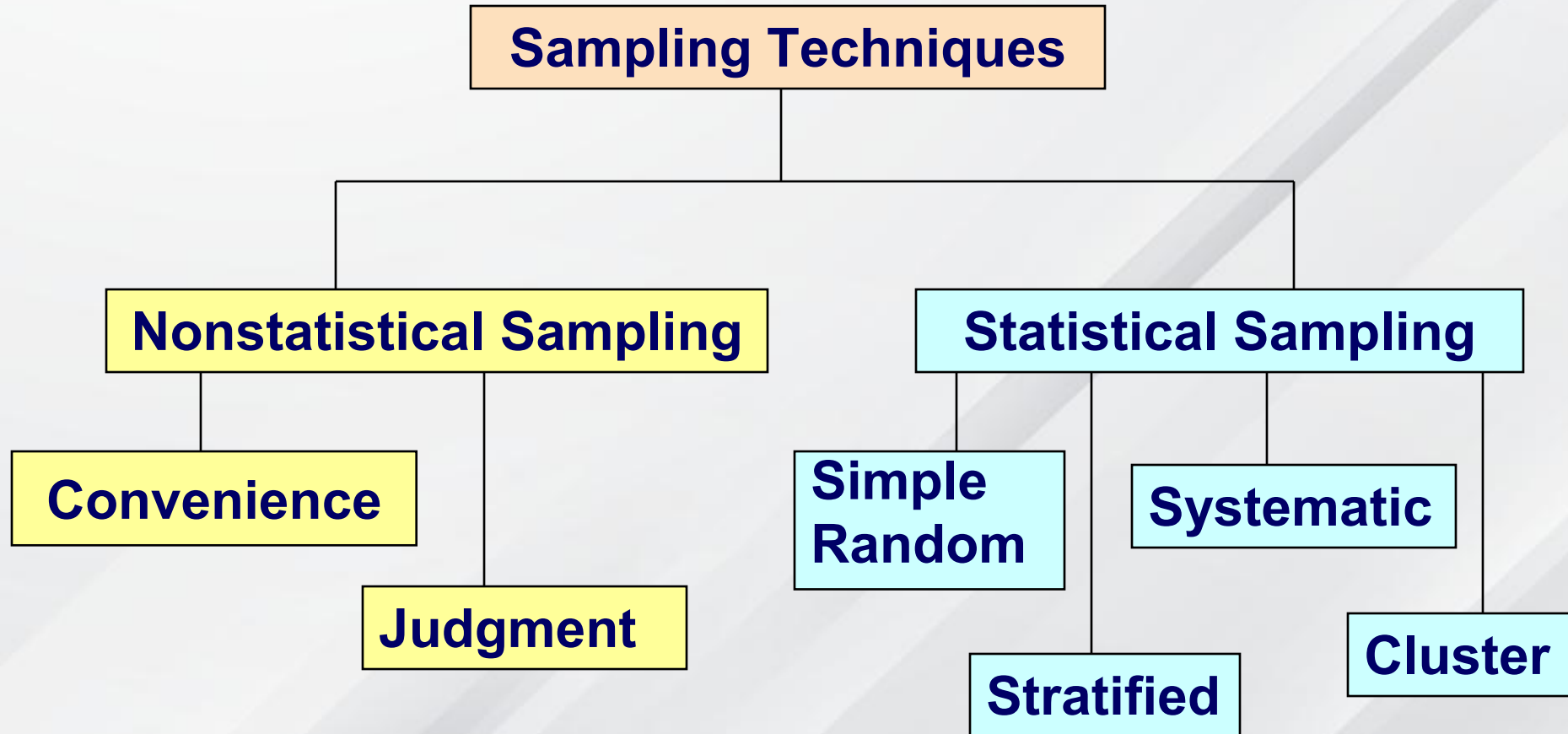
Sample



Why Sample?

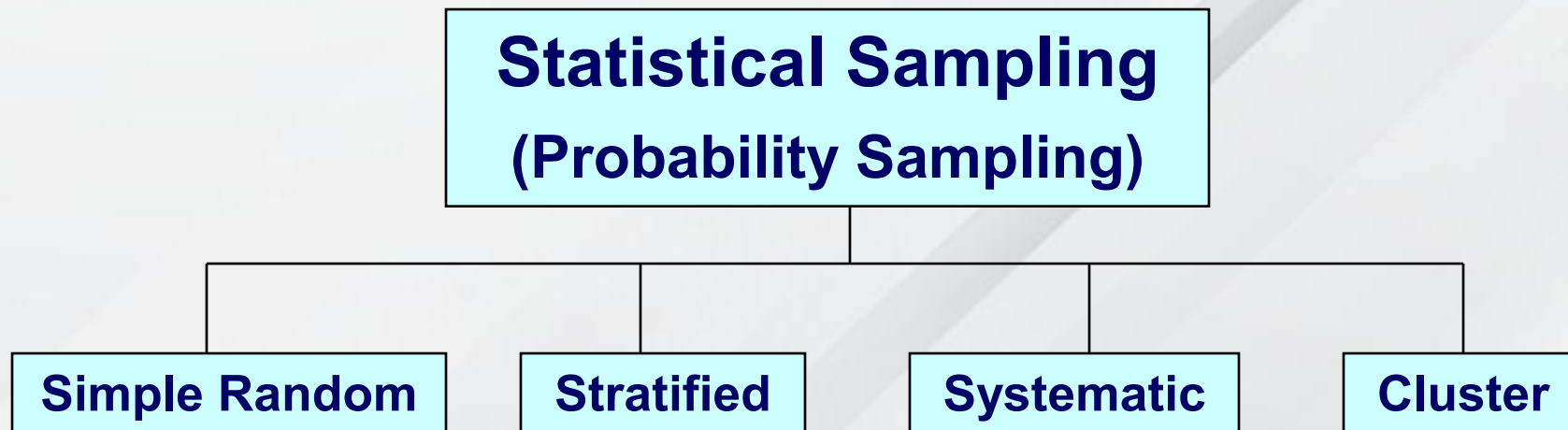
- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

Sampling Techniques



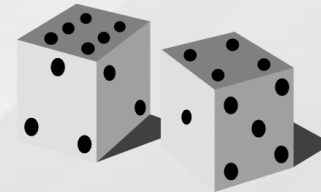
Statistical Sampling

- Items of the sample are chosen based on known or calculable probabilities



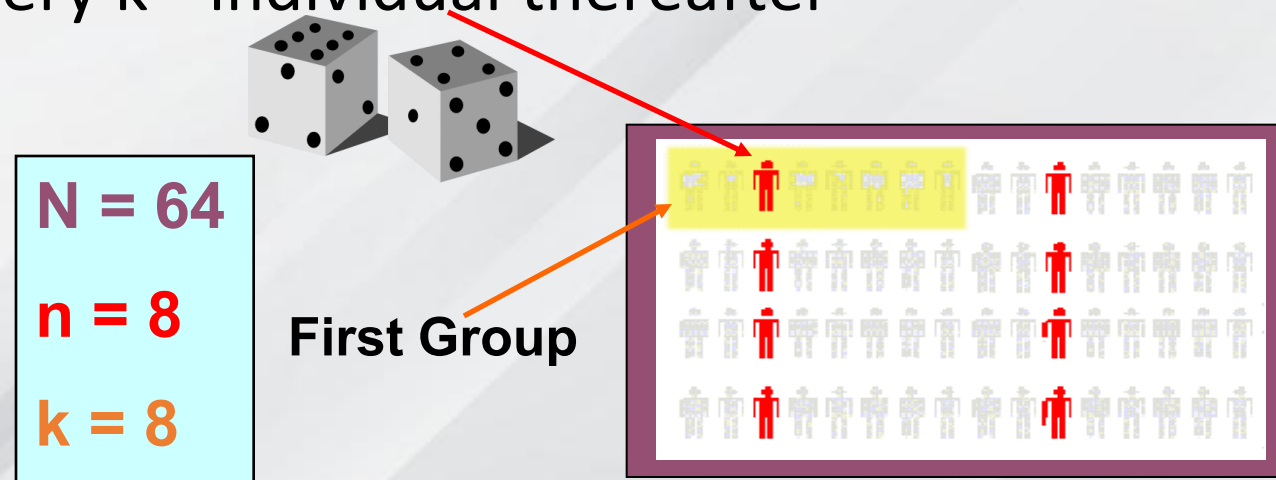
Simple Random Sampling

- Every possible sample of a given size has an **equal chance** of being selected
- Selection may be with replacement or without replacement
- The sample can be obtained using a table of random numbers or computer random number generator



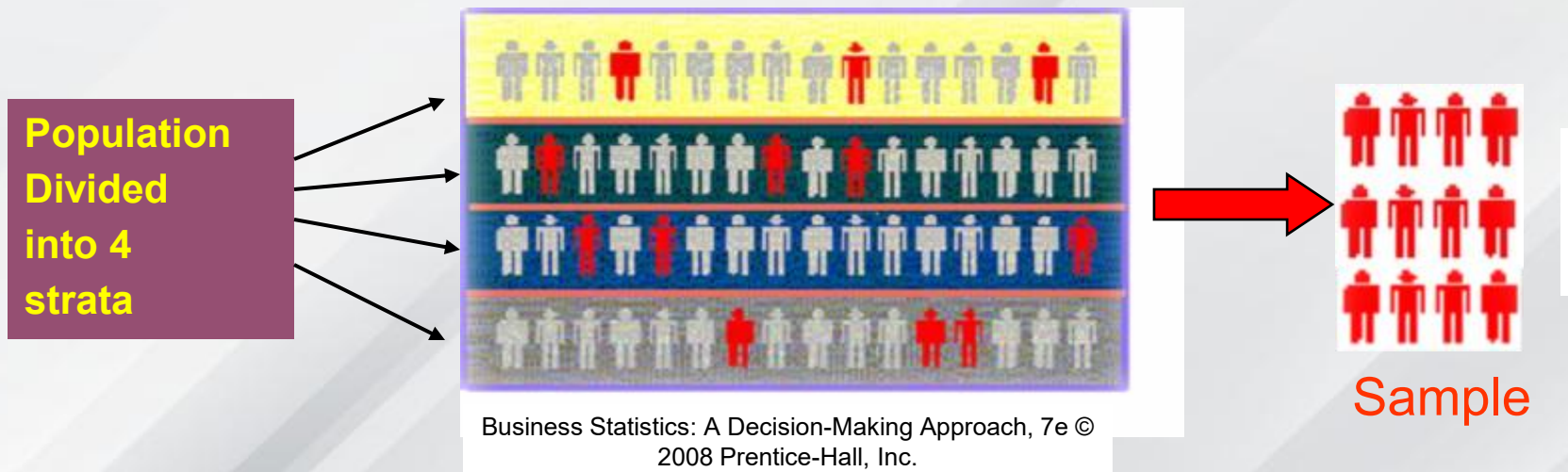
Systematic Random Sampling

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: $k = N/n$
- Randomly select one individual from the 1st group
- Select every k^{th} individual thereafter

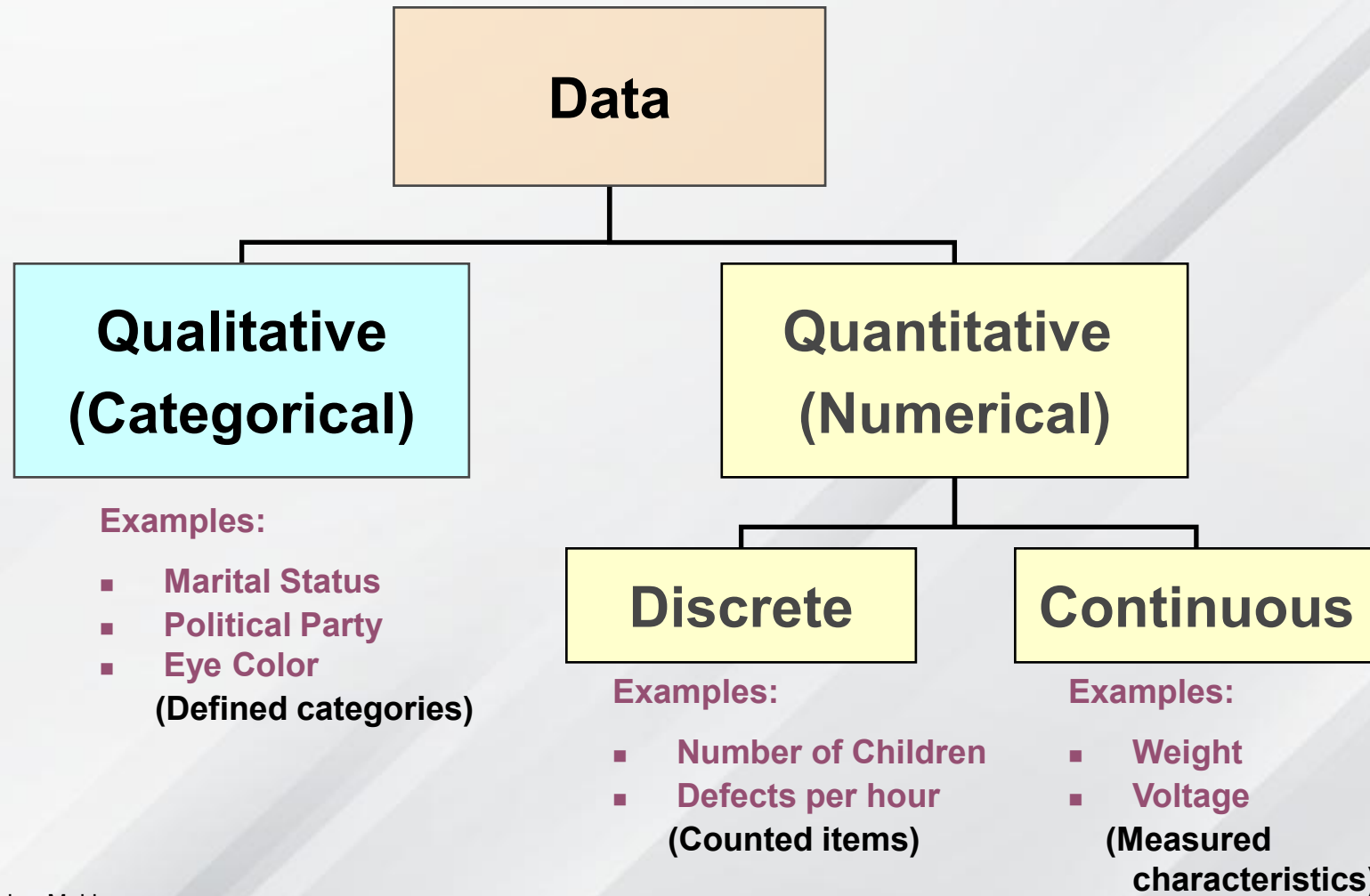


Stratified Random Sampling

- Divide population into subgroups (called *strata*) according to some common characteristic
- Select a simple random sample from each subgroup
- Combine samples from subgroups into one.



Data Types



Data Types

- **Time Series Data**
 - Ordered data values observed over time
- **Cross Section Data**
 - Data values observed at a fixed point in time

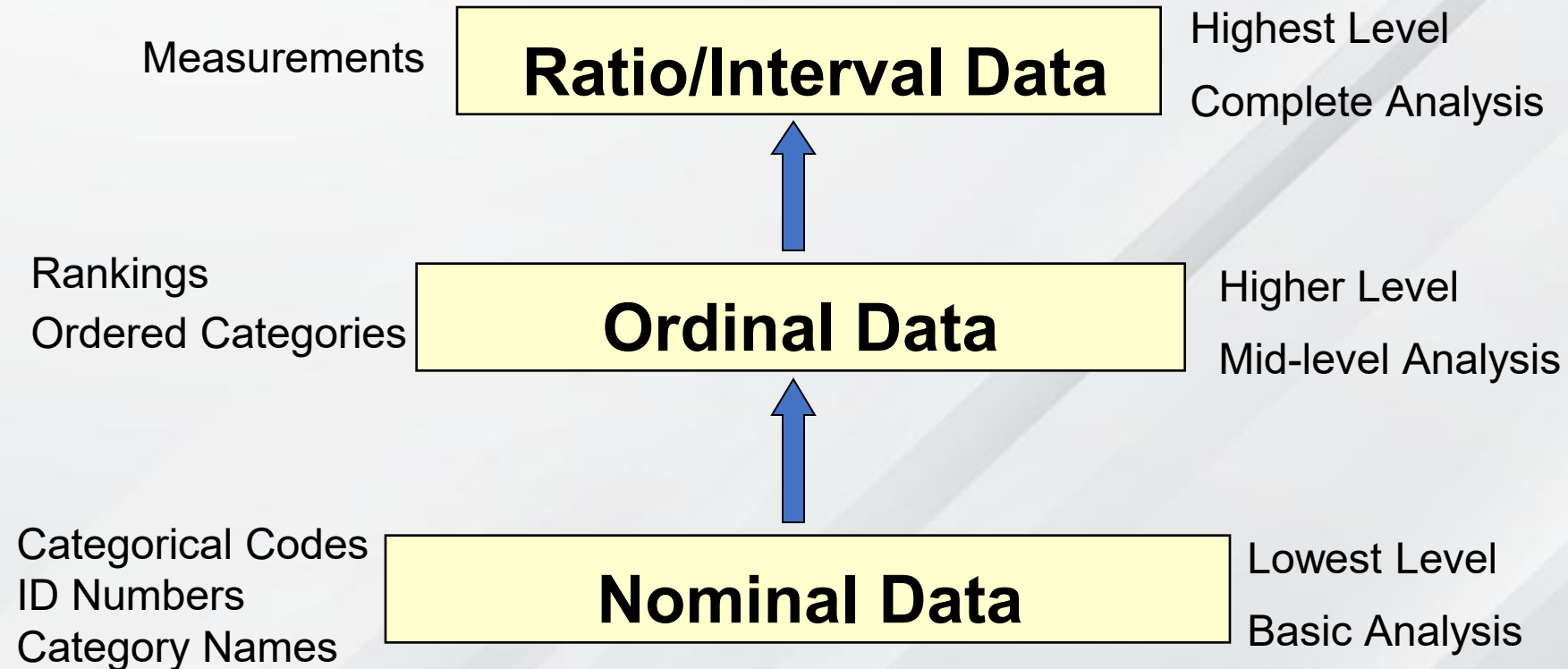
Data Types

	Sales (in \$1000's)			
	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

**Time
Series
Data**

**Cross Section
Data**

Data Measurement Levels



Chapter 2

Graphs, Charts, and Tables

– Describing Your Data

2



Chapter Goals

After completing this chapter, you should be able to:

- Construct a frequency distribution
- Construct and interpret a histogram

Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a **list or a table** ...
- containing the **values of a variable** (or a set of ranges within which the data fall) ...
- and the **corresponding frequencies** with which each value occurs (or frequencies with which data fall within each range)

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

Frequency Distribution: Discrete Data

- **Discrete data:** possible values are countable

Example: An advertiser asks 200 customers how many days per week they read the daily newspaper.



Number of days read	Frequency
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30
Total	200

Relative Frequency

Relative Frequency: What proportion is in each category?

Number of days read	Frequency	Relative Frequency
0	44	.22
1	24	.12
2	18	.09
3	16	.08
4	20	.10
5	22	.11
6	26	.13
7	30	.15
Total	200	1.00

$$\frac{44}{200} = .22$$

22% of the people in the sample report that they read the newspaper 0 days per week



Frequency Distribution: Continuous Data

- **Continuous Data:** may take on any value in some interval

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

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

(Temperature is a continuous variable because it could be measured to any degree of precision desired)

Grouping Data by Classes

Sort raw data from low to high:

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

- Find range: $58 - 12 = 46$
- Select number of classes: 5 (usually between 5 and 20)
- Compute class width: 10 (46/5 then round off)
- Determine class boundaries: 10, 20, 30, 40, 50
(Sometimes class midpoints are reported: 15, 25, 35, 45, 55)
- Count the number of values in each class

Frequency Distribution Example

Data from low to high:

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

Frequency Distribution		
Class	Frequency	Relative Frequency
10 but under 20	3	.15
20 but under 30	6	.30
30 but under 40	5	.25
40 but under 50	4	.20
50 but under 60	2	.10
Total	20	1.00

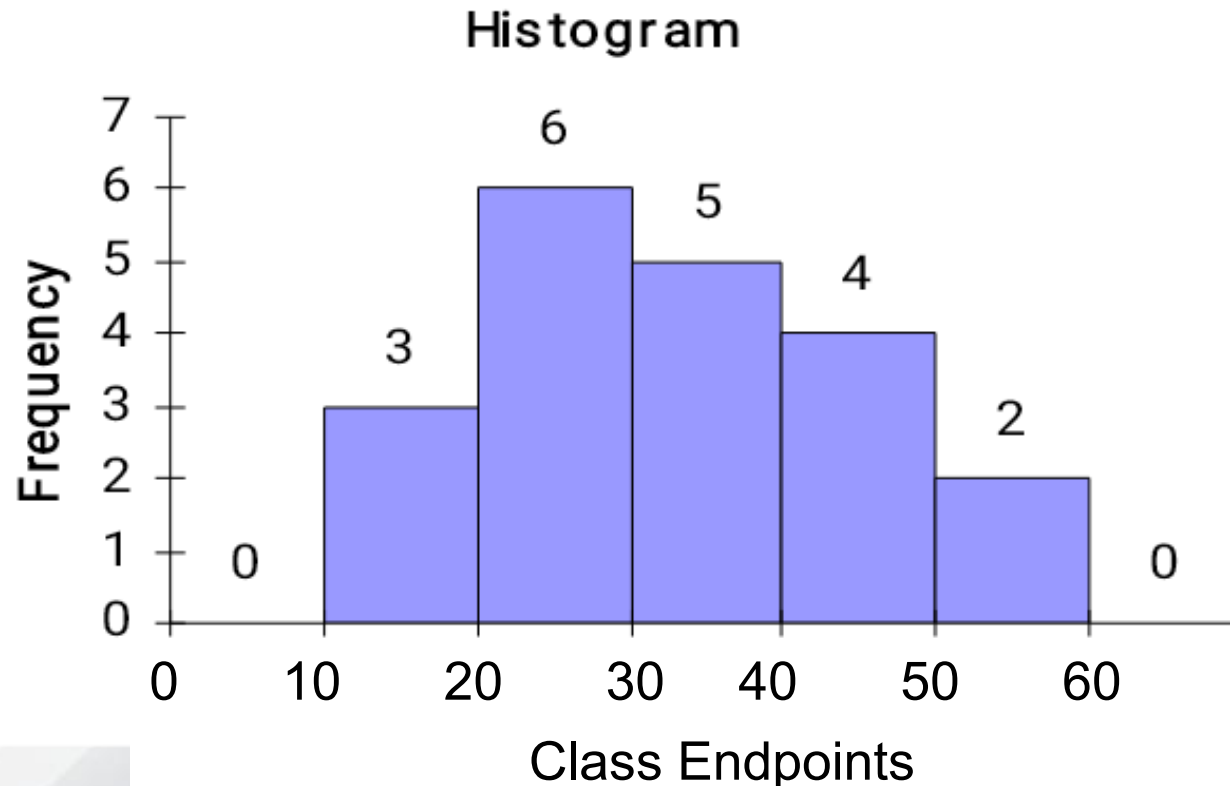
Frequency Histograms

- The **classes** or **intervals** are shown on the **horizontal axis**
- **frequency** is measured on the **vertical axis**
- Bars of the appropriate heights can be used to represent the number of observations within each class
- Such a graph is called a **histogram**

Histogram Example

Data in ordered array:

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



No gaps
between
bars, since
continuous
data

Chapter 3

Describing Data Using Numerical Measures

3



Chapter Goals

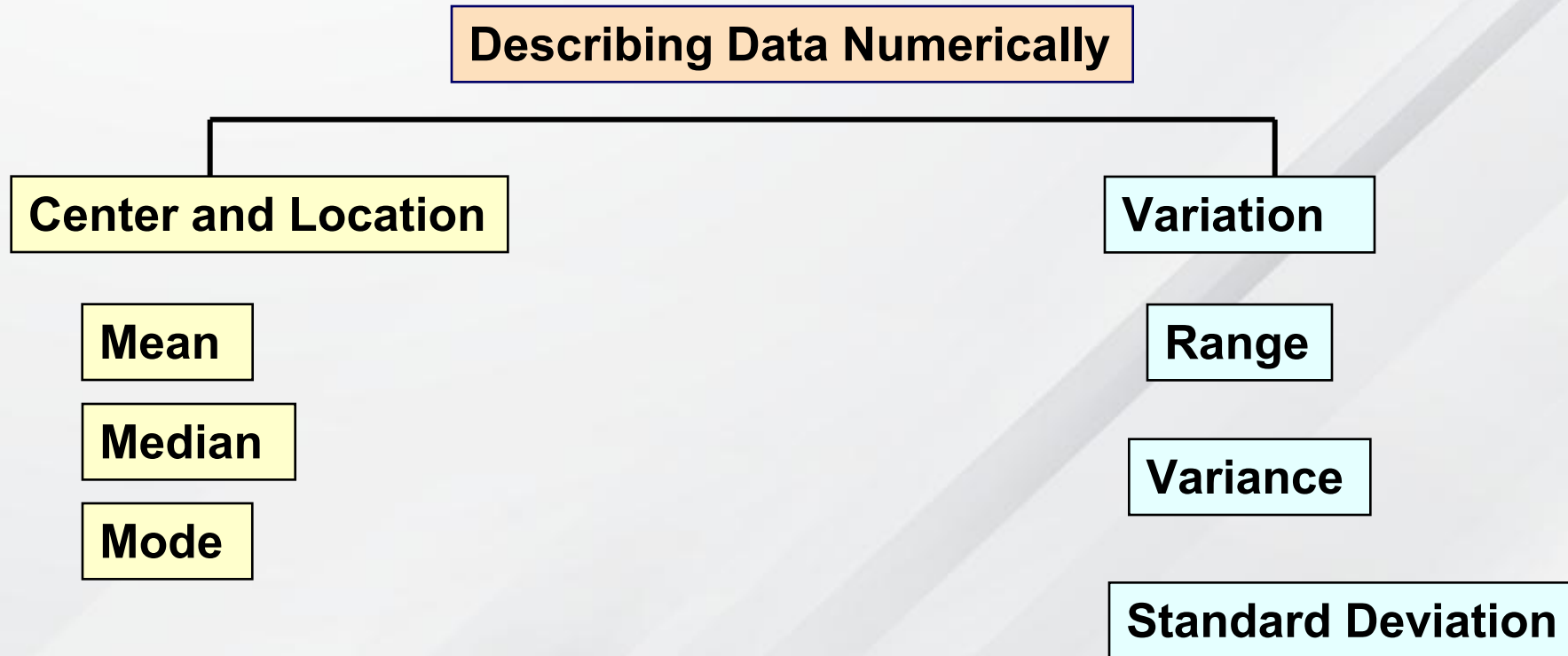
After completing this chapter, you should be able to:

- 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

Chapter Topics

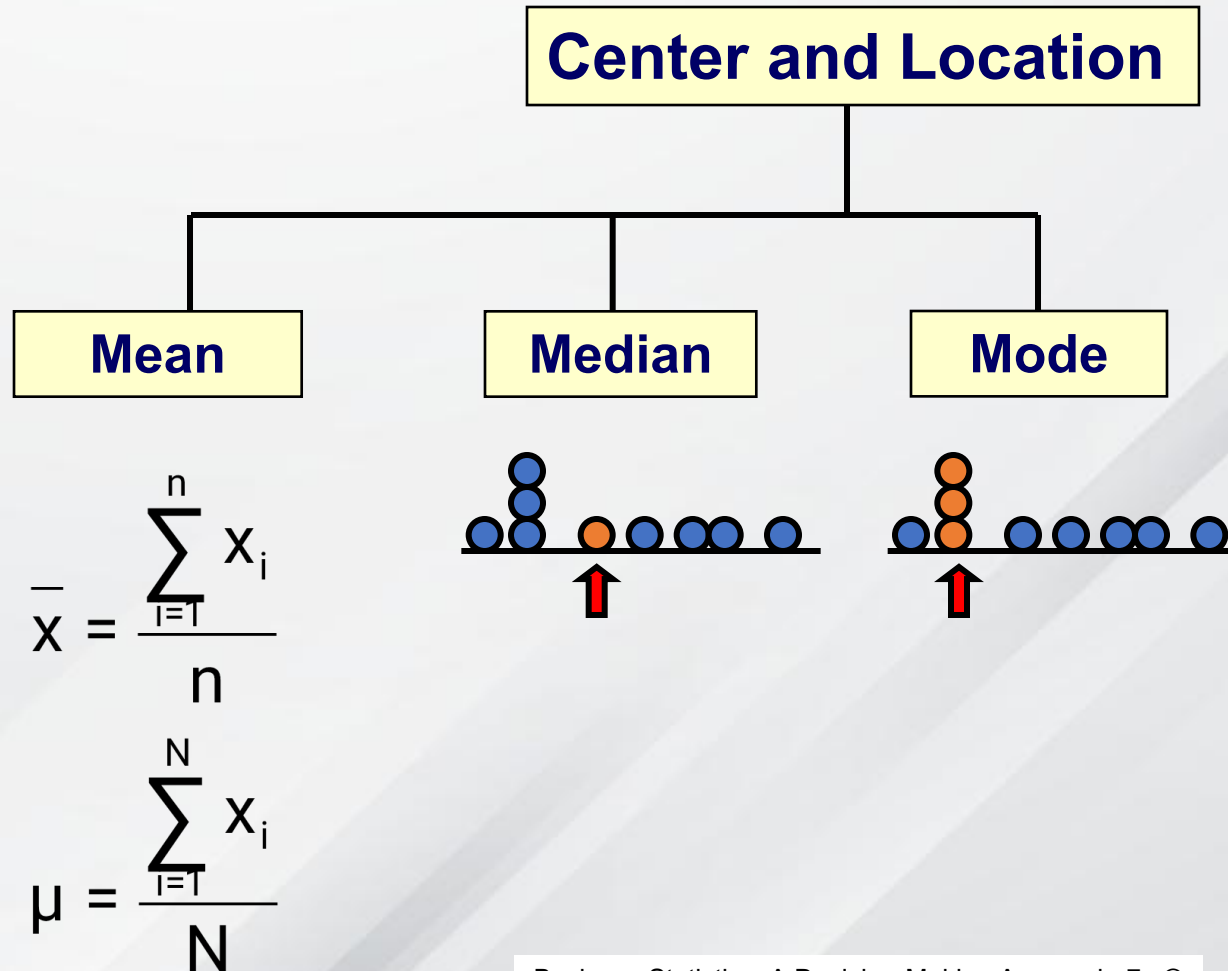
- Measures of Center and Location
 - Mean, median, mode
- Measures of Variation
 - Range, variance and standard deviation

Summary Measures



Measures of Center and Location

Overview



Mean (Arithmetic Average)

- The **Mean** is the arithmetic average of data values

- Population mean

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

N = Population Size

- Sample mean

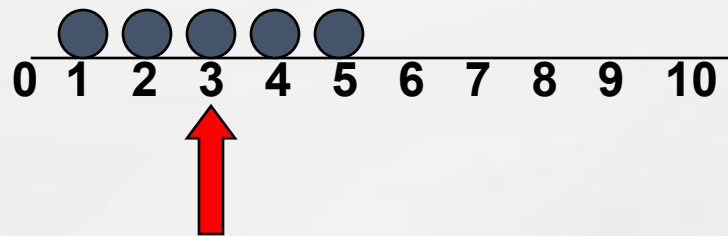
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

n = Sample Size

Mean (Arithmetic Average)

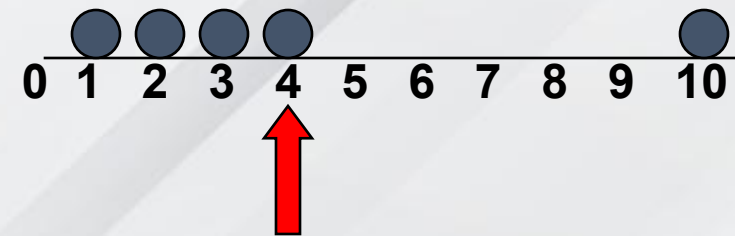
(continued)

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



Mean = 3

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

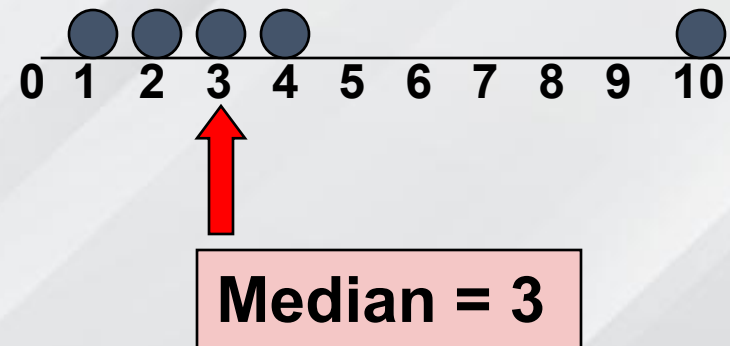
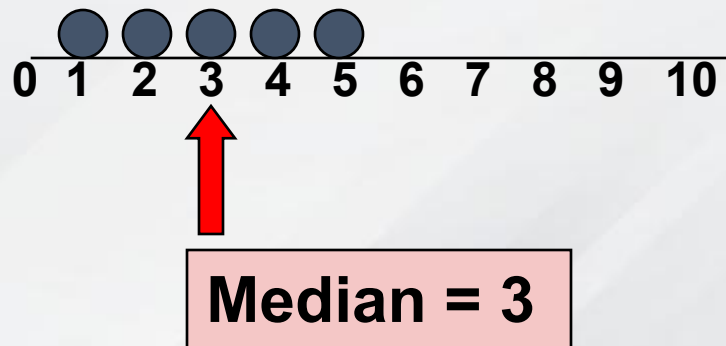


Mean = 4

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

Median

- In an ordered array, the median is the “middle” number, i.e., the number that splits the distribution in half
- The median is not affected by extreme values



Median

(continued)

- To find the median, sort the n data values from low to high (sorted data is called a data array)
- Find the value in the $i = (1/2)n$ position
- The i^{th} position is called the Median Index Point
 - If i is not an integer, round up to next highest integer

Median Example

(continued)

Data array:

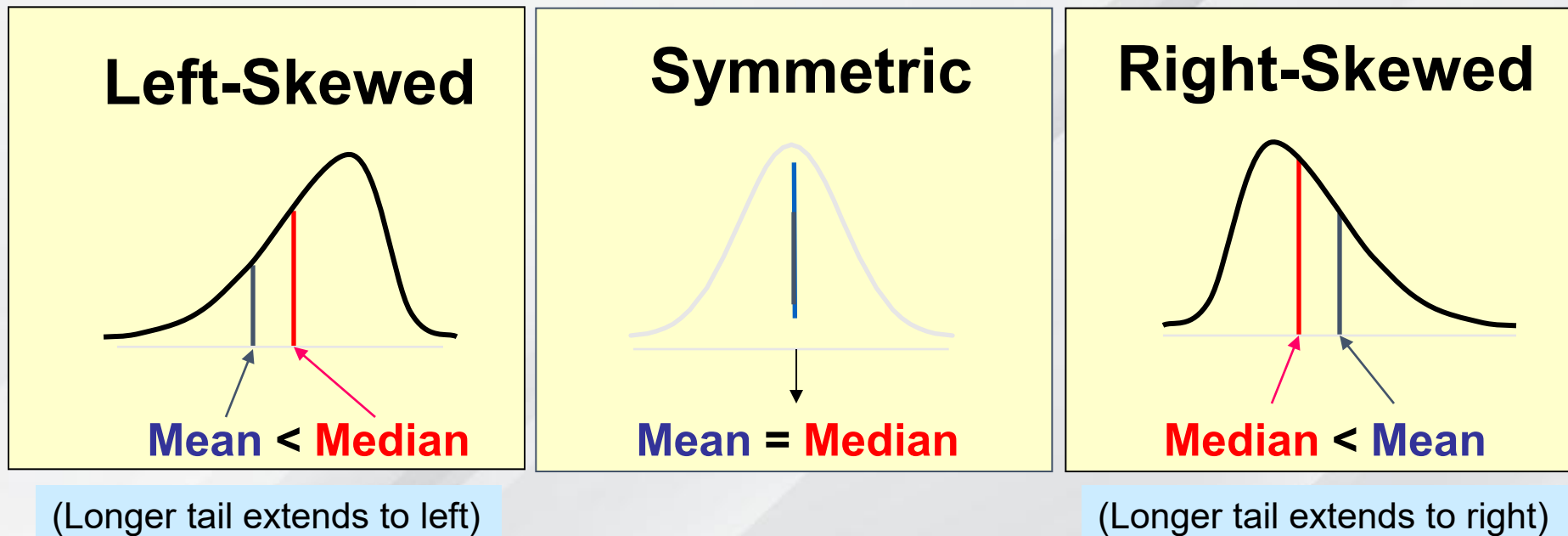
4, 4, 5, 5, 9, 11, 12, 14, 16, 19, 22, 23, 24

- Note that $n = 13$
- Find the $i = (1/2)n$ position:
$$i = (1/2)(13) = 6.5$$
- Since 6.5 is not an integer, round up to 7
- The median is the value in the 7th position:

$$M_d = 12$$

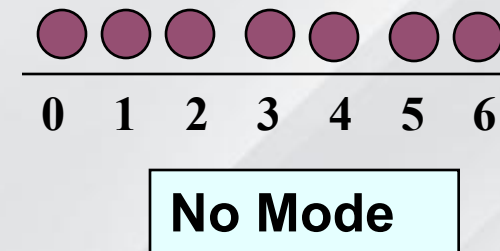
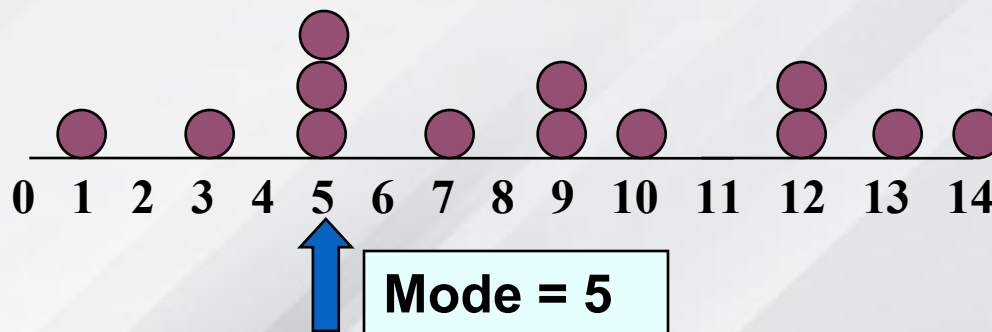
Shape of a Distribution

- Describes how data is distributed
- Symmetric or skewed



Mode

- A measure of location
- The 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

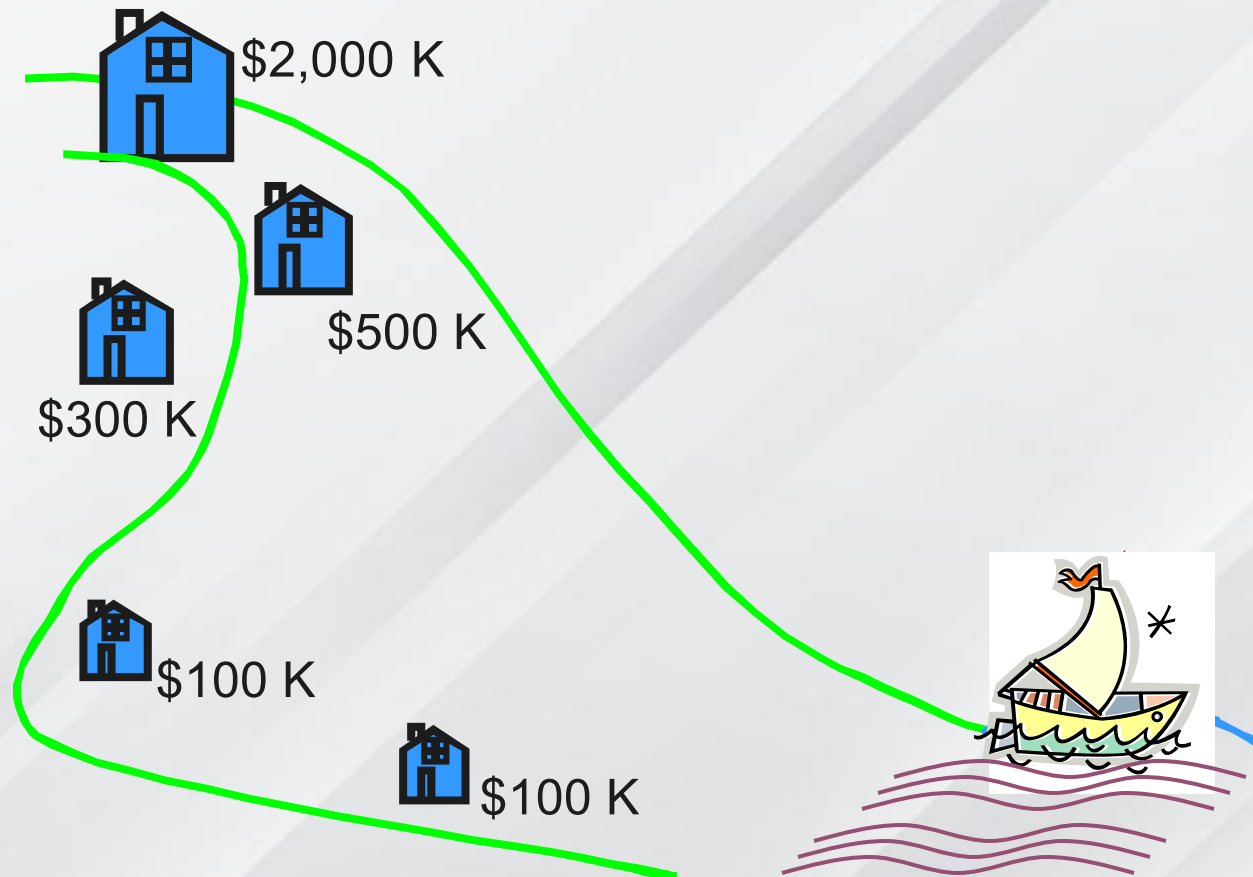


Review Example

- Five houses on a hill by the beach

House Prices:

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



Summary Statistics

House Prices:

\$2,000,000
500,000
300,000
100,000
<u>100,000</u>

Sum 3,000,000

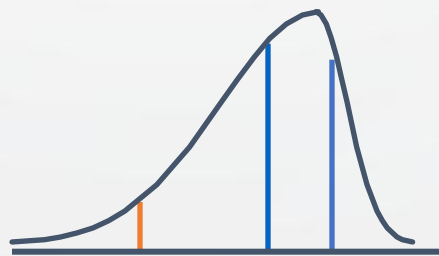
- **Mean:** $(\$3,000,000/5)$
= **\$600,000**
- **Median:** middle value of ranked data
= **\$300,000**
- **Mode:** most frequent value
= **\$100,000**

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

Distribution Shape and Box and Whisker Plot

Left-Skewed



Q1 Q2 Q3



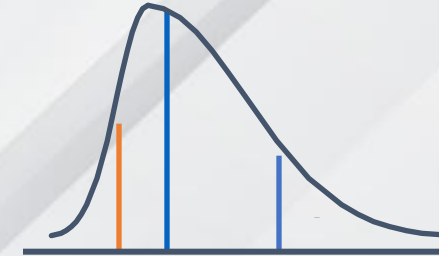
Symmetric



Q1 Q2 Q3



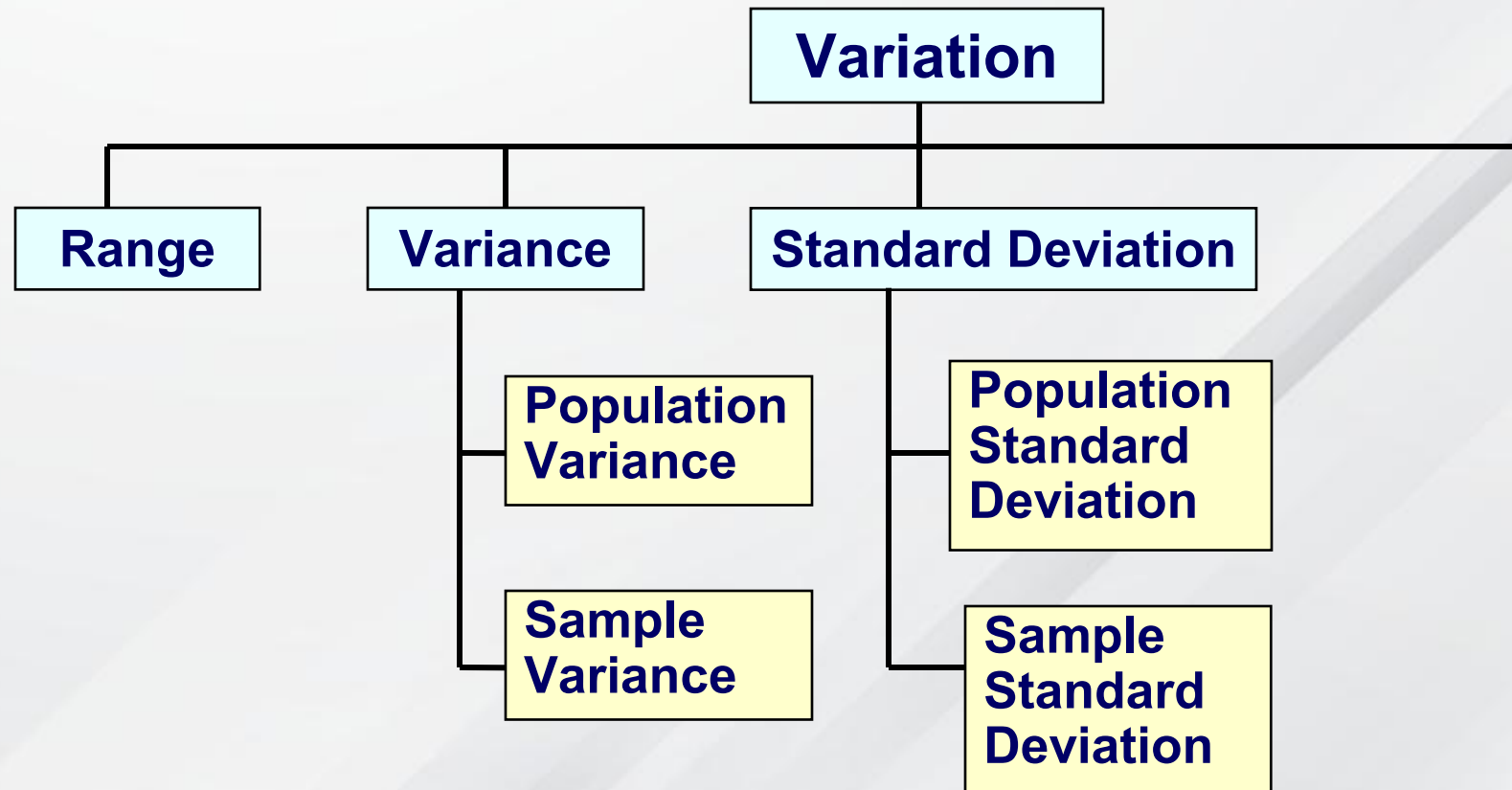
Right-Skewed



Q1 Q2 Q3

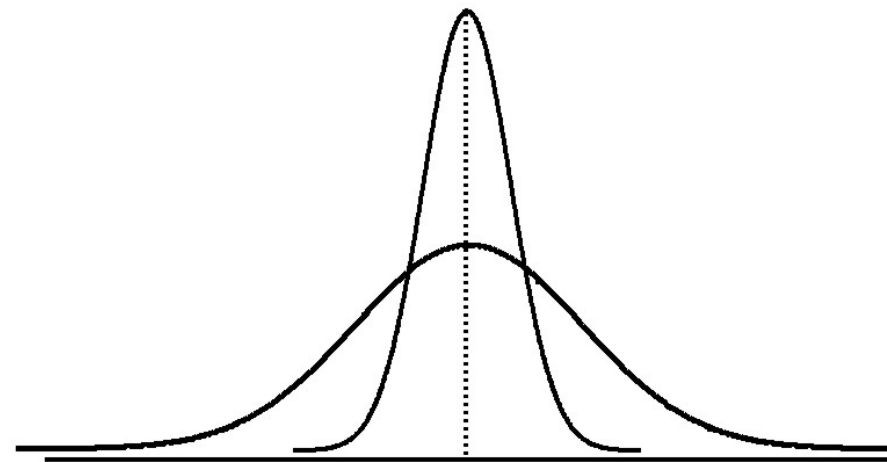


Measures of Variation



Variation

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



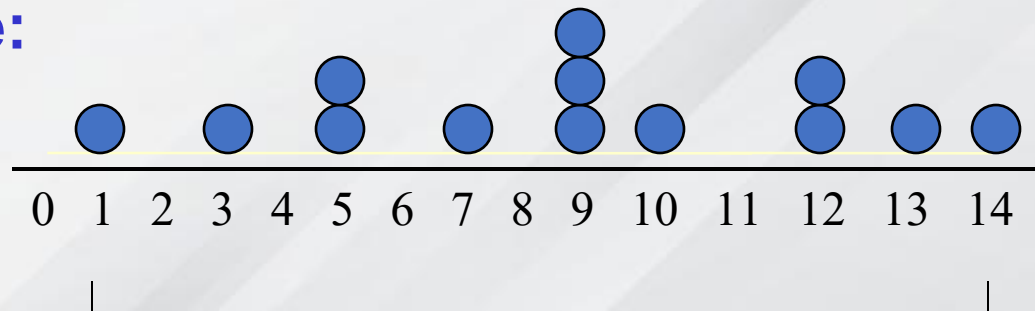
**Same center,
different variation**

Range

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

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

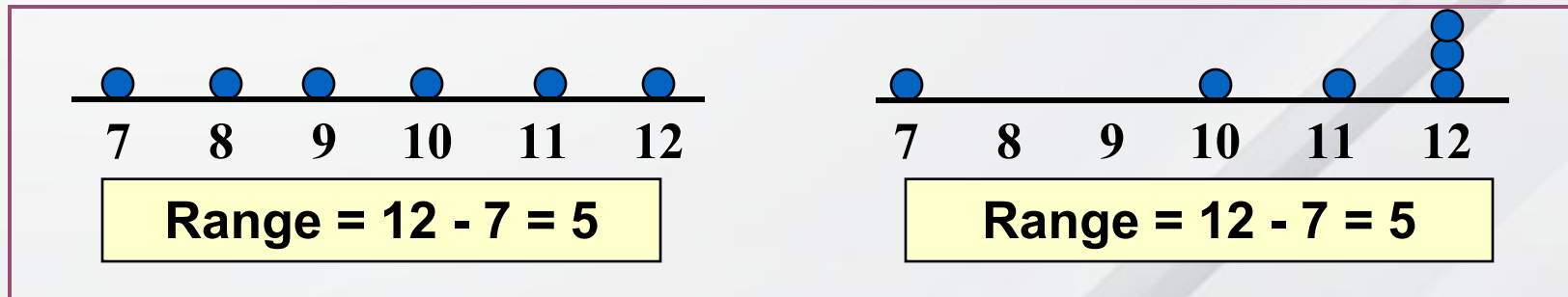
Example:



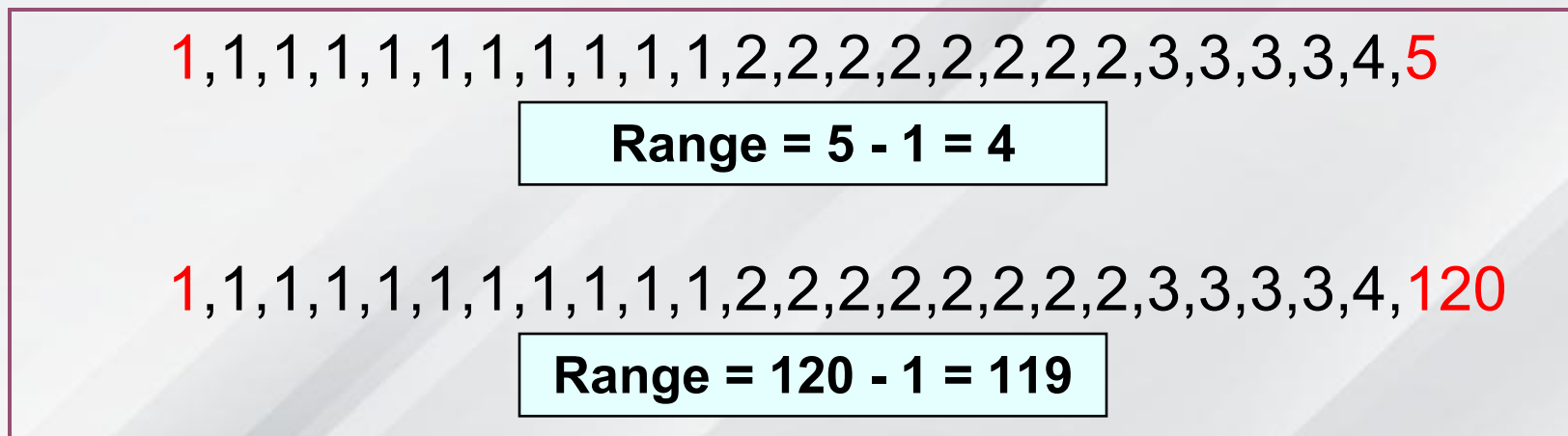
$$\text{Range} = 14 - 1 = 13$$

Disadvantages of the Range

- Ignores the way in which data are distributed



- Sensitive to outliers



Variance

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

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

- **Sample** variance:

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

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:

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

Calculation Example: Sample Standard Deviation

Sample

Data (X_i) :

10 12 14 15 17 18 18 24

$n = 8$

Mean = $\bar{x} = 16$

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

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

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

Chapter 9

Introduction to Hypothesis Testing

4



Chapter Goals

After completing this chapter, you should be able to:

- Formulate null and alternative hypotheses for applications involving a single population mean or proportion
- Formulate a decision rule for testing a hypothesis
- Know how to use the test statistic, critical value, and p-value approaches to test the null hypothesis
- Know what Type I and Type II errors are
- Compute the probability of a Type II error

What is a Hypothesis?

- A hypothesis is a claim (assumption) about a population parameter:



The Null Hypothesis, H_0

(continued)

- Begin with the assumption that the null hypothesis is true
 - Similar to the notion of innocent until proven guilty
- Refers to the **status quo**
- May or may not be rejected



The Alternative Hypothesis, H_A

- Is the opposite of the null hypothesis
 - e.g.: Challenges the status quo
- May or may not be accepted
- Is generally the hypothesis that is believed (or needs to be supported) by the researcher – a **research hypothesis**

Errors in Making Decisions

- **Type I Error**
 - Reject a true null hypothesis
 - Considered a serious type of error

The probability of Type I Error is α

- Called **level of significance** of the test
- Set by researcher in advance

Errors in Making Decisions

(continued)

- **Type II Error**
 - Fail to reject a false null hypothesis

The probability of Type II Error is β

- β is a calculated value, the formula is discussed later in the chapter

Outcomes and Probabilities

Possible Hypothesis Test Outcomes

	State of Nature	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No error ($1 - \alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	No Error ($1 - \beta$)

Key:
Outcome
(Probability)

Type I & II Error Relationship

- Type I and Type II errors cannot happen at the same time
 - Type I error can only occur if H_0 is true
 - Type II error can only occur if H_0 is false

Level of Significance, α

- Defines unlikely values of sample statistic if null hypothesis is true
 - Defines rejection region of the sampling distribution
- Is designated by α , (level of significance)
 - Typical values are .01, .05, or .10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

Process of Hypothesis Testing

- 1. Specify population parameter of interest
- 2. Formulate the null and alternative hypotheses
- 3. Specify the desired significance level, α
- 4. Define the rejection region
- 5. Take a random sample and determine whether or not the sample result is in the rejection region
- 6. Reach a decision and draw a conclusion

P-Value Approach to Testing

- Determine the **p-value** from a table or computer
- Compare the **p-value** with **α**
 - If $p\text{-value} < \alpha$, reject H_0
 - If $p\text{-value} \geq \alpha$, do not reject H_0



Thank You