STATISTICAL METHODS

Master in Industrial Management,

Operations and Sustainability (MIMOS)

2nd year/1st Semester

2025/2026

CONTACT

Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

PROGRAM

1. Fundamental Concepts of Statistics
2. Descriptive Data Analysis
3. Introduction to Inferential Analysis
4. Parametric Hypothesis Testing
5. Non-Parametric Hypothesis Testing
6. Linear Regression Analysis

LECTURE 2: DESCRIPTIVE DATA ANALYSIS – TABLES AND GRAPHICS (CONTINUATION)

EXERCISE 1.36

1.36 The following table shows the ages of competitors in a charity tennis event in Rome:

Age	Percent
18–24	18.26
25–34	16.25
35–44	25.88
45–54	19.26
55+	20.35

a. Construct a relative cumulative frequency distribution.
b. What percent of competitors were under the age of 35?
c. What percent of competitors were 45 or older?

Newbold et al (2013)

EXERCISE 1.36 A): SOLUTION

✅ Answer:

a) Relative cumulative frequency distribution

To get the **cumulative percentage**, we add the percentages successively:

Age	Percent	Cumulative Percent
18–24	18.26	18.26
25–34	16.25	18.26 + 16.25 = 34.51
35–44	25.88	34.51 + 25.88 = 60.39
45–54	19.26	60.39 + 19.26 = 79.65
55+	20.35	79.65 + 20.35 = 100.00

The **cumulative percentage column** is the **relative cumulative frequency (in percentage)**.

EXERCISE 1.36 B): SOLUTION

✅ Answer:

b) Percent of competitors under the age of 35

- Competitors under 35 include **18–24** and **25–34**.
- Add the percentages:

$$18.26 + 16.25 = 34.51\%$$

34.51% of competitors were under 35.

EXERCISE 1.36 C): SOLUTION

✅ Answer:

c) Percent of competitors 45 or older

- Competitors 45 or older include **45–54** and **55+**.
- Add the percentages:

$$19.26 + 20.35 = 39.61\%$$

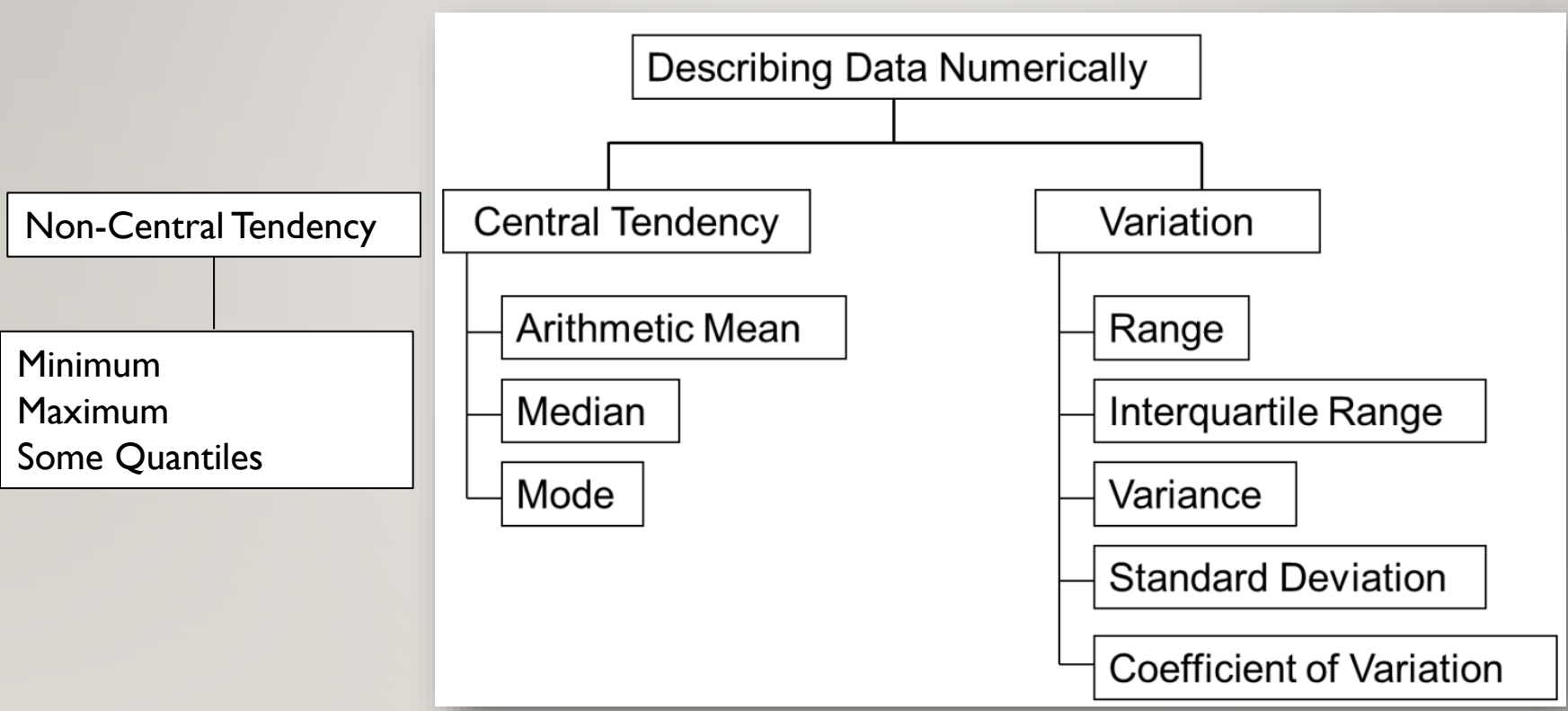
39.61% of competitors were 45 or older.

LECTURE 2: DESCRIPTIVE DATA ANALYSIS - MEASURES

DESCRIBING DATA NUMERICALLY

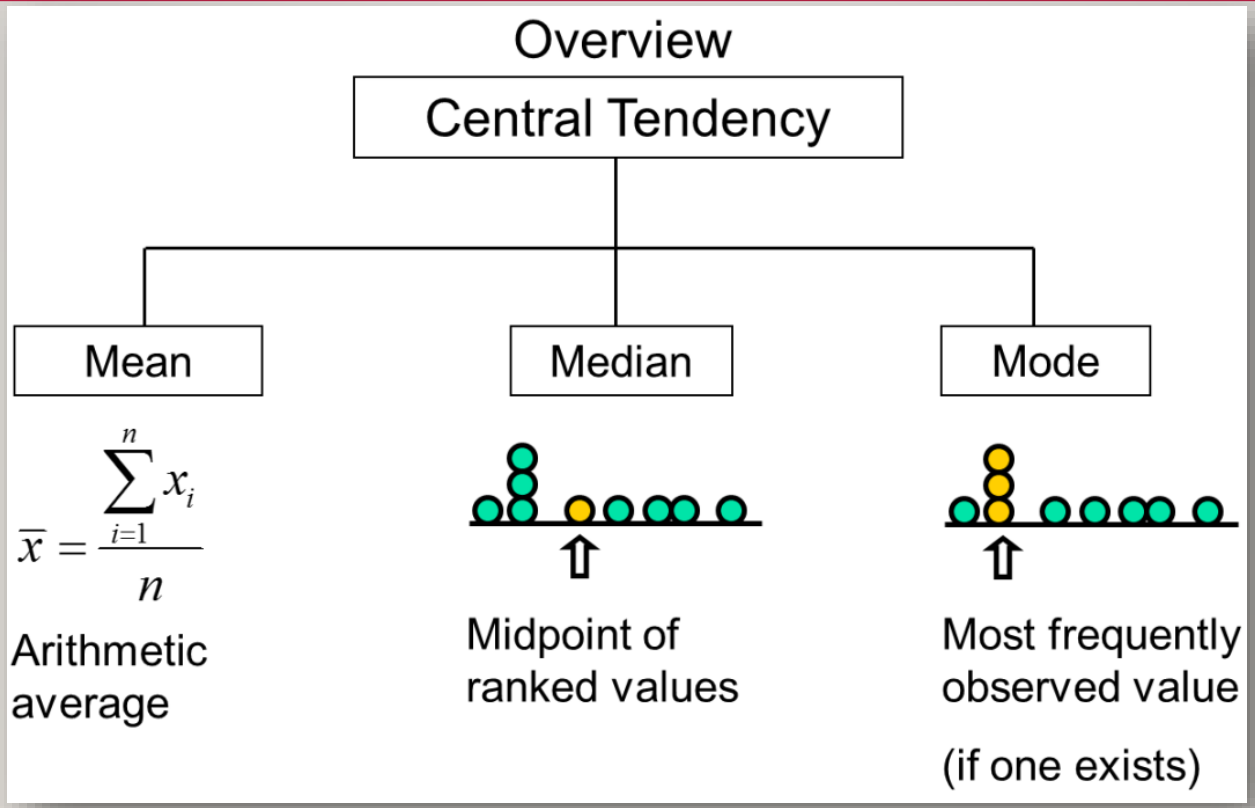
- **Measures:**
  - Central and Non-Central Location / Tendency
  - Dispersion / Variation
  - Skewness
  - Kurtosis

MEASURES OF TENDENCY AND VARIATION



Newbold et al (2013)

MEASURES OF CENTRAL TENDENCY



Newbold et al (2013)

ARITHMETIC MEAN

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*:

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

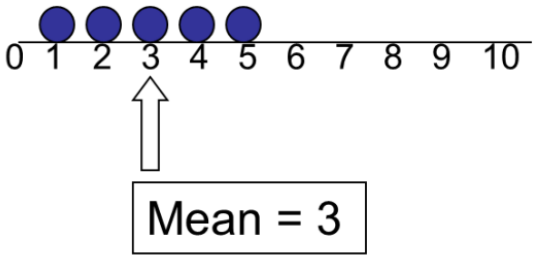
Observed values

Sample size

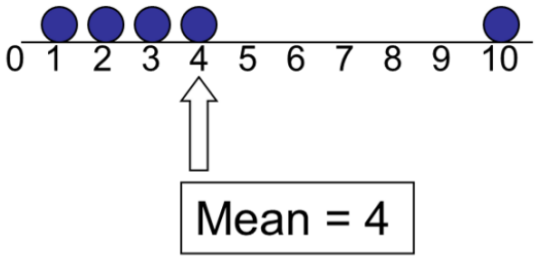
Newbold et al (2013)

ARITHMETIC MEAN: EXAMPLES

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



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

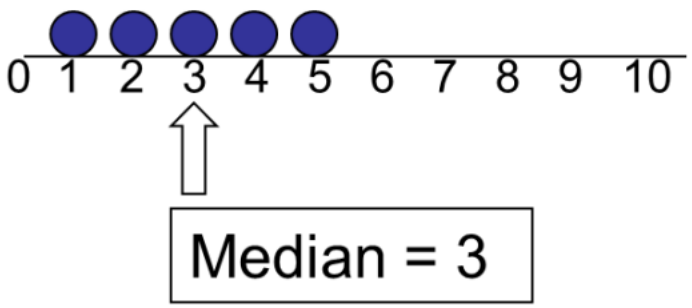


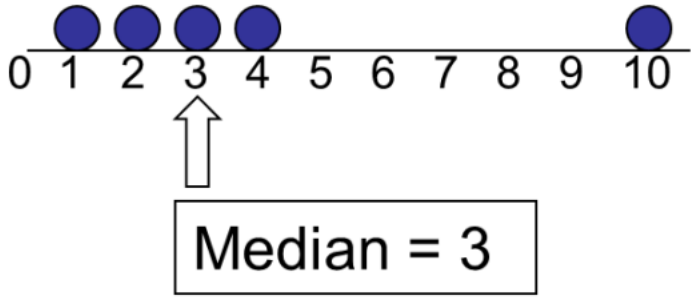
$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Newbold et al (2013)

MEDIAN

- In an ordered list, the median is the “middle” number (50% above, 50% below)





- Not affected by extreme values

Newbold et al (2013)

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

Newbold et al (2013)

CALCULATING THE MEDIAN: EXAMPLES

Formula to Find Median Position

$$\text{Position} = \frac{n+1}{2}$$

- n = number of observations
- If **Position is integer** → Median = value at that position
- If **Position is not integer** → Median = average of values at floor and ceil(Position)

Examples

Example 1 – Position is integer

- Data: 2, 4, 6, 8, 10 ($n = 5$)

$$\text{Position} = \frac{5+1}{2} = 3 \quad \text{(integer)}$$

$$\text{Median} = 3\text{rd value} = 6$$

Example 2 – Position is not integer

- Data: 3, 5, 8, 12, 15, 18 ($n = 6$)

$$\text{Position} = \frac{6+1}{2} = 3.5 \quad \text{(not integer)}$$

$$\text{Median} = \frac{3\text{rd value} + 4\text{th value}}{2} = \frac{8+12}{2} = 10$$

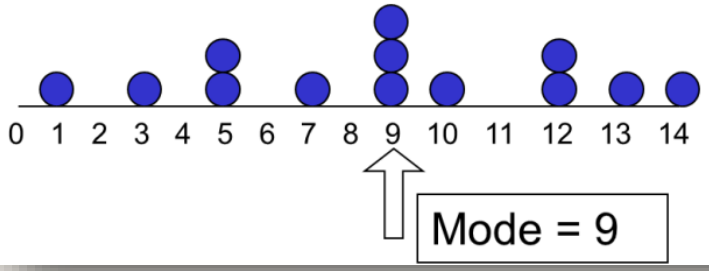
WHICH MEASURE OF LOCATION IS THE "BEST"?

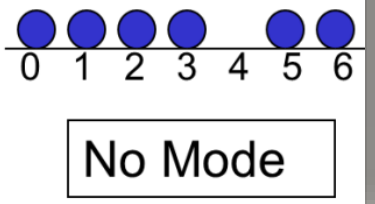
- **Mean** is generally used, unless extreme values (outliers) exist …
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Newbold et al (2013)

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





Newbold et al (2013)

DISTRIBUTIONS BY MODE

1 Amodal Distribution (No Mode)

- **Definition:** No value repeats → no mode
- **Example:** 2, 3, 5, 7, 11

2 Unimodal Distribution (One Mode)

- **Definition:** One value appears most frequently
- **Example:** 1, 2, 2, 3, 4 → **Mode = 2**

3 Bimodal Distribution (Two Modes)

- **Definition:** Two values appear with the same highest frequency
- **Example:** 1, 2, 2, 3, 3, 4 → **Modes = 2, 3**

4 Multimodal Distribution (More than Two Modes)

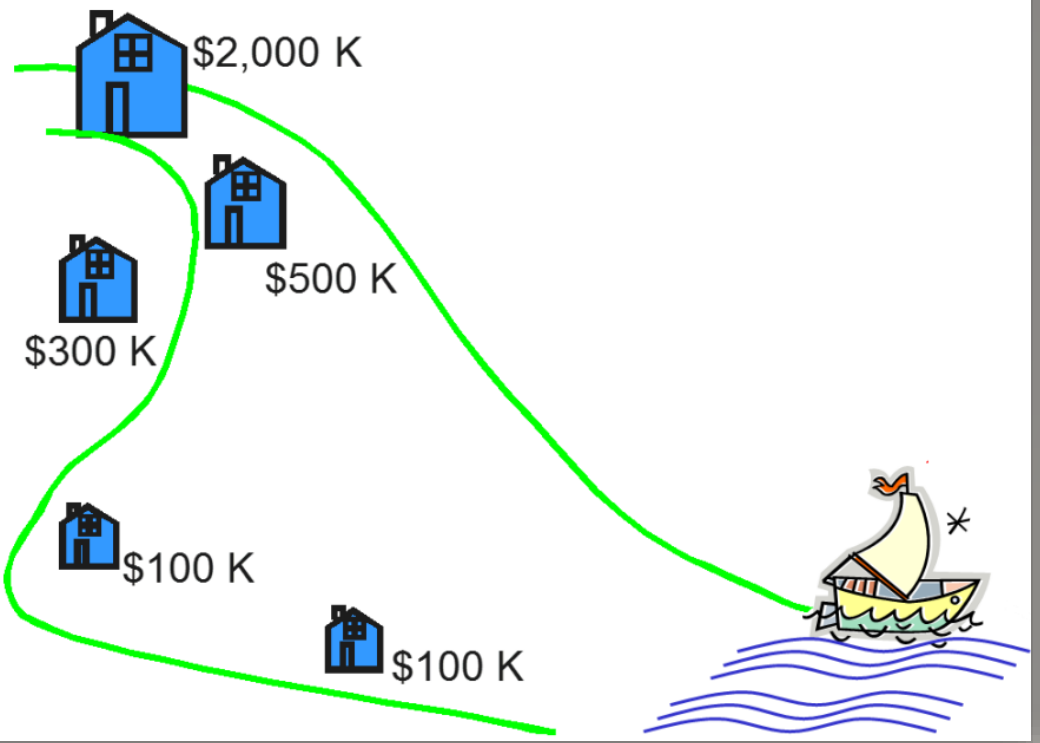
- **Definition:** More than two values appear with the same highest frequency
- **Example:** 1, 1, 2, 2, 3, 3, 4 → **Modes = 1, 2, 3**

REVIEW EXAMPLE

- Five houses on a hill by the beach

House Prices:

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



Newbold et al (2013)

REVIEW EXAMPLE: SUMMARY STATISTICS

House Prices:

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

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

Newbold et al (2013)

QUANTILES: DEFINITION

1️⃣ What are Quantiles?

- Quantiles are **values that divide a dataset into equal parts**.
- Special cases:
  - **Quartiles** → Q1, Q2, Q3, Q4 (divide data into 4 equal parts)
    - **Median = Q2**
  - **Deciles** → D1, D2, ..., D10 (divide data into 10 equal parts)
    - **Median = D5**
  - **Percentiles** → P1, P2, ..., P100 (divide data into 100 equal parts)
    - **Median = P50**

QUANTILES: CALCULATION STEPS (NEWBOLD METHOD)

2 Steps to Calculate a Quantile (Newbold Method)

1. **Order the data** (ascending).
2. **Calculate the position:**

$$\text{Position} = \alpha \cdot (n + 1)$$

- n = number of observations
- α = proportion of the quantile (e.g., 0.25 for Q1/P25, 0.50 for median/P50)

3. **Determine the quantile value:**

- **If Position is integer:**

$$\text{Quantile} = \text{value at that position}$$

- **If Position is not integer:**

$$\text{Quantile} = \frac{\text{value at floor(Position)} + \text{value at ceil(Position)}}{2}$$

QUANTILES: EXAMPLES

3 Examples

Example 1 – Position is integer (Median / Q2 / D5 / P50):

Data: 3, 5, 8, 12, 15, 18, 20 ($n = 7$)

$$\text{Position} = 0.5 \cdot (7 + 1) = 4 \qquad \text{Quantile} = 4\text{th value} = 12$$

Example 2 – Position is not integer (Q1 / P25):

Data: 3, 5, 8, 12, 15, 18, 20, 25 ($n = 8$)

$$\text{Position} = 0.25 \cdot (8 + 1) = 2.25 \qquad \text{Quantile} = (5 + 8)/2 = 6.5$$

4 Note

- Other formulas exist (e.g., SPSS, Excel) → results may **differ slightly**.
- Key: choose **one method** and apply consistently.

PERCENTILES AND QUARTILES

Percentiles and Quartiles

To find percentiles and quartiles, data must first be arranged in order from the smallest to the largest values.

The ***P*th percentile** is a value such that approximately P% of the observations are at or below that number. **Percentiles** separate large ordered data sets into 100ths. The 50th percentile is the median.

The Pth percentile is found as follows:

$$P\text{th percentile} = \text{value located in the } (P/100)(n+1)\text{th ordered position} \qquad (2.6)$$

Quartiles are descriptive measures that separate large data sets into four quarters. The **first quartile**, Q_1, (or 25th *percentile*) separates approximately the smallest 25% of the data from the remainder of the data. The **second quartile**, Q_2, (or 50th *percentile*) is the median (see Equation 2.3).

Newbold et al (2013)

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

Newbold et al (2013)

QUARTILES

- 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

Newbold et al (2013)

QUARTILE FORMULAS

$$\text{Position} = \alpha \cdot (n+1)$$

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

Newbold et al (2013)

QUARTILE: EXAMPLE

- 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 $Q_1 = 12.5$

Newbold et al (2013)

FIVE-NUMBER SUMMARY

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

minimum

first quartile

median

third quartile

maximum

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

Newbold et al (2013)

EXERCISE 2.6

2.6 During the last 3 years Consolidated Oil Company expanded its gasoline stations into convenience food stores (CFSs) in an attempt to increase total sales revenue. The daily sales (in hundreds of dollars) from a random sample of 10 weekdays from one of its stores are:

6 8 10 12 14 9 11 7 13 11

a. Find the mean, median and mode for this store.
b. Find the five-number summary.

Newbold et al (2013)

EXERCISE 2.6: SOLUTION

✅ Answer:

Given data (daily sales in hundreds of dollars): 6, 8, 10, 12, 14, 9, 11, 7, 13, 11

a. Mean, Median, and Mode

- **Order the data (ascending):** 6, 7, 8, 9, 10, 11, 11, 12, 13, 14
- **Mean:**

$$\bar{x} = \frac{\text{sum of all values}}{n}$$

$$\bar{x} = \frac{6+7+8+9+10+11+11+12+13+14}{10} = \frac{101}{10} = 10.1$$

Mean = 10.1 (hundreds of dollars)

EXERCISE 2.6: SOLUTION

✅ Answer:

a. Mean, Median, and Mode

- **Order the data (ascending):** 6, 7, 8, 9, 10, 11, 11, 12, 13, 14

- **Median:** Number of observations n = 10 (even), so:

$$\text{Median} = \frac{5\text{th value} + 6\text{th value}}{2} = \frac{10 + 11}{2} = 10.5$$

- **Mode:**

  Most frequent value = 11 (appears twice)

Median = 10.5 (hundreds of dollars)

Mode = 11 (hundreds of dollars)

EXERCISE 2.6: SOLUTION

✅ Answer:

b. Five-number summary

- **Order the data (ascending):** 6, 7, 8, 9, 10, 11, 11, 12, 13, 14

Q1: $\alpha = 0.25$

$$\text{Position} = 0.25 \cdot 11 = 2.75$$

- Not integer → take **average of 2nd and 3rd values**:

$$Q1 = \frac{x_2 + x_3}{2} = \frac{7 + 8}{2} = 7.5$$

Q3: $\alpha = 0.75$

$$\text{Position} = 0.75 \cdot 11 = 8.25$$

- Not integer → average of 8th and 9th values:

$$Q3 = \frac{x_8 + x_9}{2} \downarrow \frac{12 + 13}{2} = 12.5$$

Five-number summary using the "average for non-integer position" method:

$$\text{Minimum} = 6, \quad Q1 = 7.5, \quad \text{Median} = 10.5, \quad Q3 = 12.5, \quad \text{Maximum} = 14$$

EXERCISE 2.6: SOLUTION

✅ Answer:

b. Five-number summary: Alternative solution (Both methods are correct!)

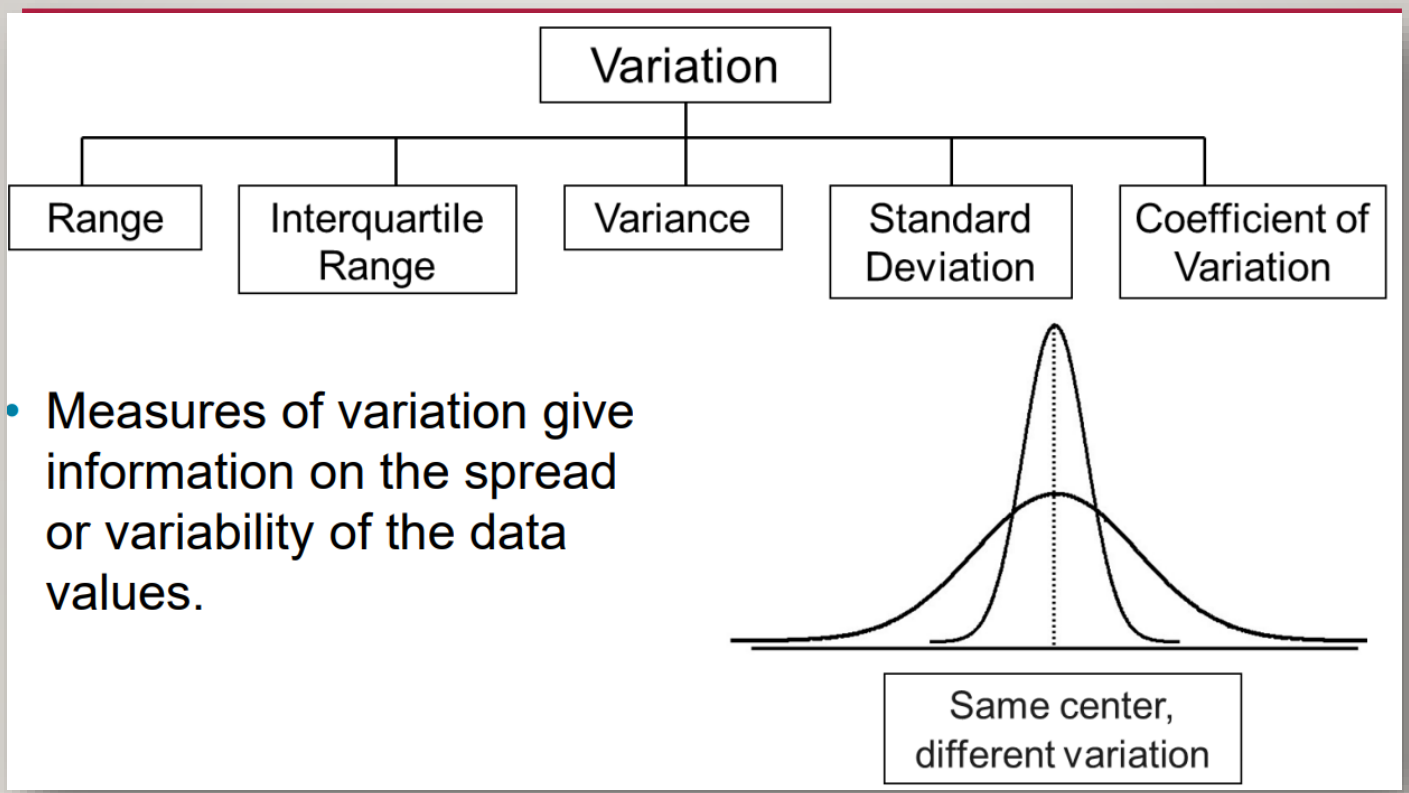
- **Order the data (ascending):** 6, 7, 8, 9, 10, 11, 11, 12, 13, 14

  - Ordered data: 6, 7, 8, 9, 10, 11, 11, 12, 13, 14
  - **Q2 (Median) = 10.5**
  - **Q1:** median of lower half (first 5 values: 6, 7, 8, 9, 10)
    - Middle value = 8 → Q1 = 8
  - **Q3:** median of upper half (last 5 values: 11, 11, 12, 13, 14)
    - Middle value = 12 → Q3 = 12

Five-number summary:

$$\text{Minimum} = 6, \quad Q1 = 8, \quad \text{Median} = 10.5, \quad Q3 = 12, \quad \text{Maximum} = 14$$

MEASURES OF VARIABILITY



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

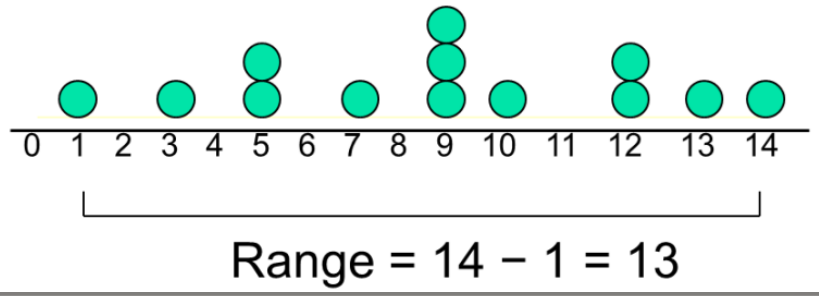
Newbold et al (2013)

RANGE

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

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

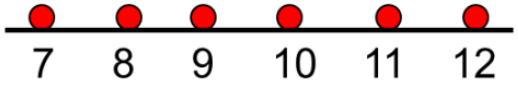
Example:



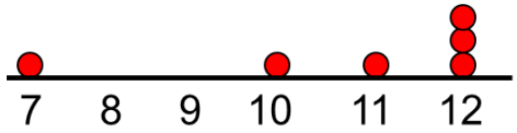
Newbold et al (2013)

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

Newbold et al (2013)

INTERQUARTILE RANGE

- 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

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

Newbold et al (2013)

INTERQUARTILE RANGE

- 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

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

Newbold et al (2013)

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

Newbold et al (2013)

SAMPLE VARIANCE

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

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

Where $\bar{x}$ = arithmetic mean

n = sample size

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

Newbold et al (2013)

SAMPLE VARIANCE – TWO FORMULAS

Formulas (for a sample)

1. **Sample Variance (divide by n):**

$$s_n^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

- Correct formula for a sample.
- Simple calculation, but on average gives a slightly smaller value than the true population variance.

---

2. **Sample Variance (divide by n−1):**

$$s_{n-1}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- Also correct for a sample.
- **Preferred formula** → provides an **unbiased estimate** of the population variance.

Note

- Both formulas are correct for sample variance.
- The version with **n−1** is standard in statistics because it adjusts for the estimation of the mean from the sample.

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:

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

Newbold et al (2013)

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:

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

Newbold et al (2013)

SAMPLE STANDARD DEVIATION: EXAMPLE

Sample Data (x_i): | 10 | 12 | 14 | 15 | 17 | 18 | 18 | 24 |

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

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

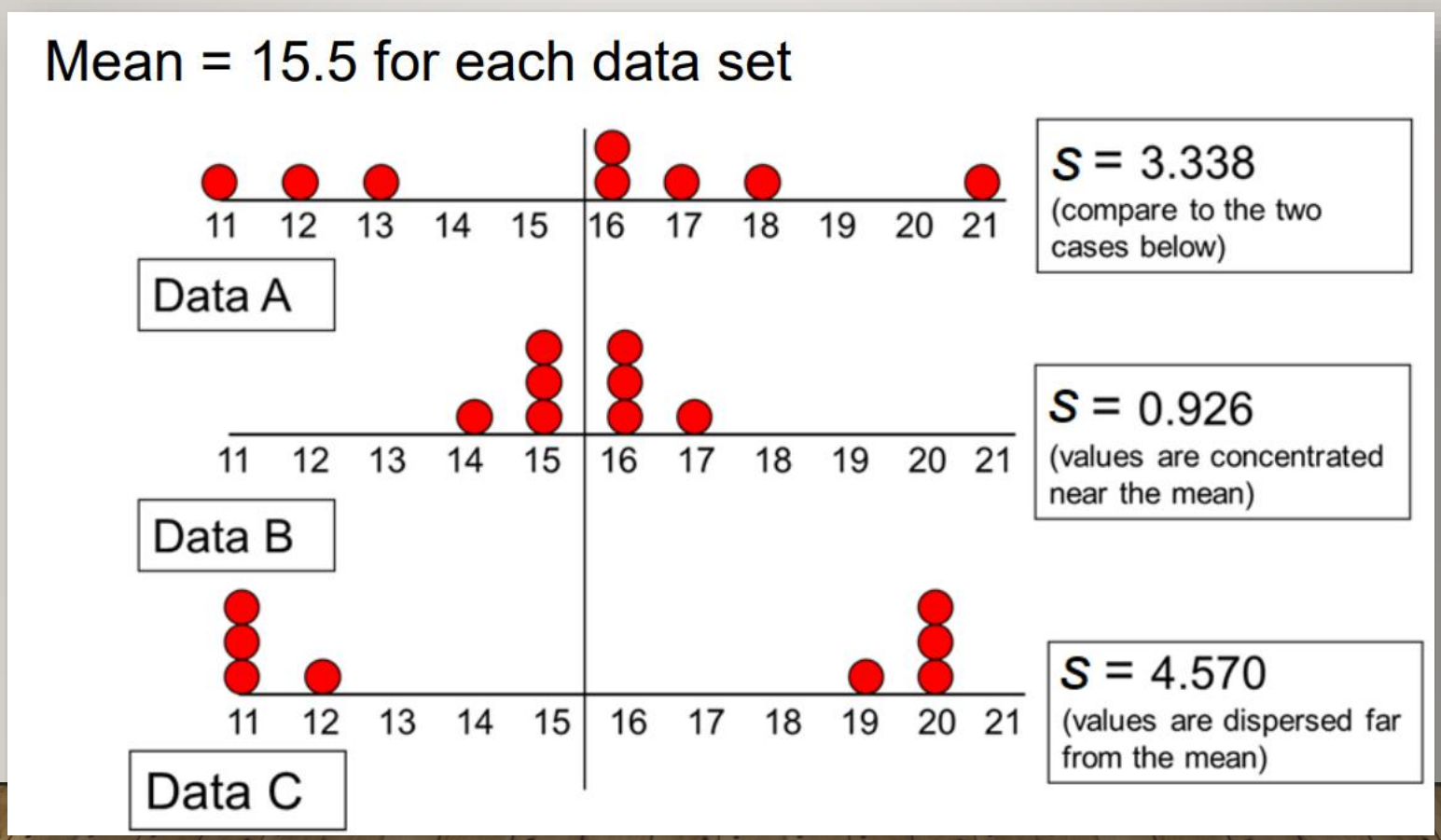
$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

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

A measure of the "average" scatter around the mean

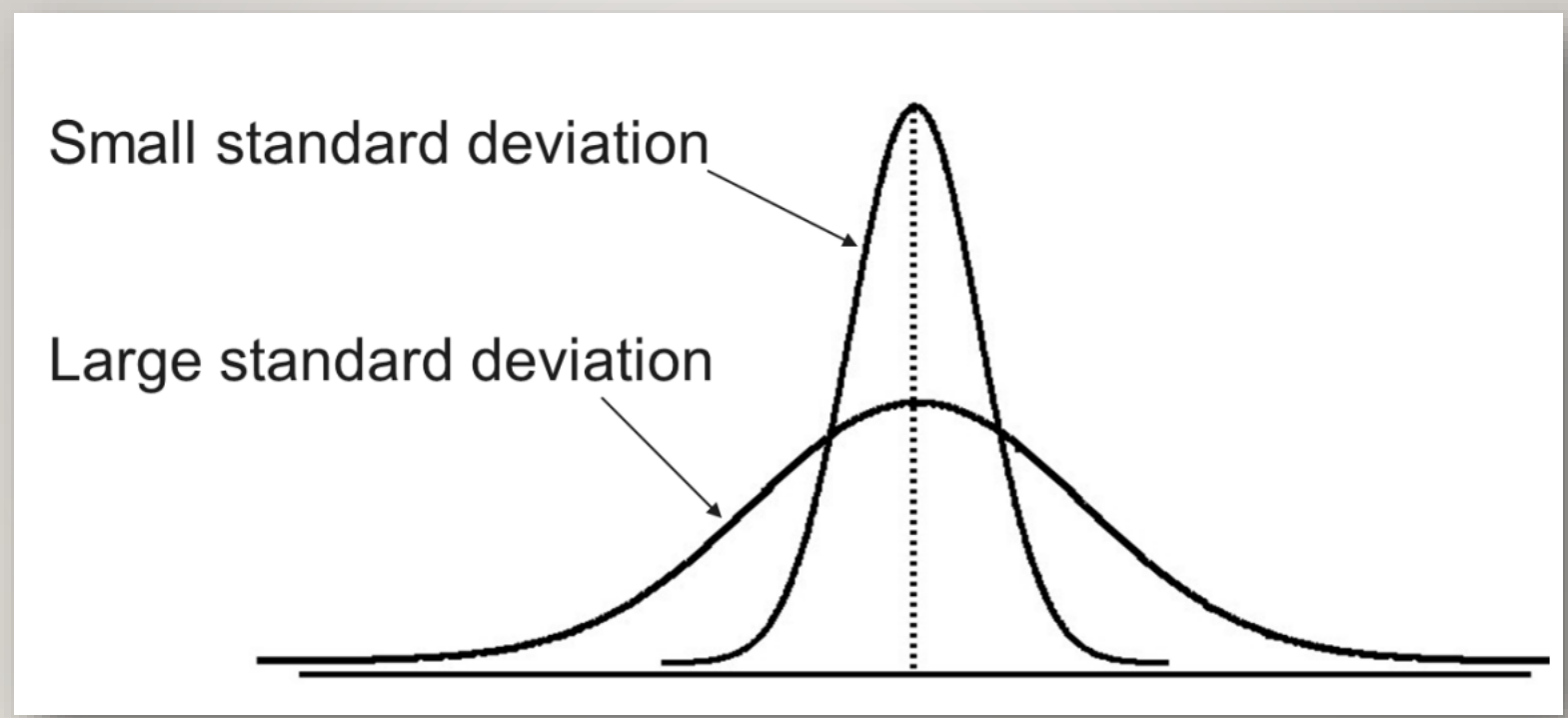
Newbold et al (2013)

COMPARING STANDARD DEVIATIONS



Newbold et al (2013)

MEASURING VARIATION



Newbold et al (2013)

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)

Newbold et al (2013)

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:

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

Sample coefficient of variation:

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

Newbold et al (2013)

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

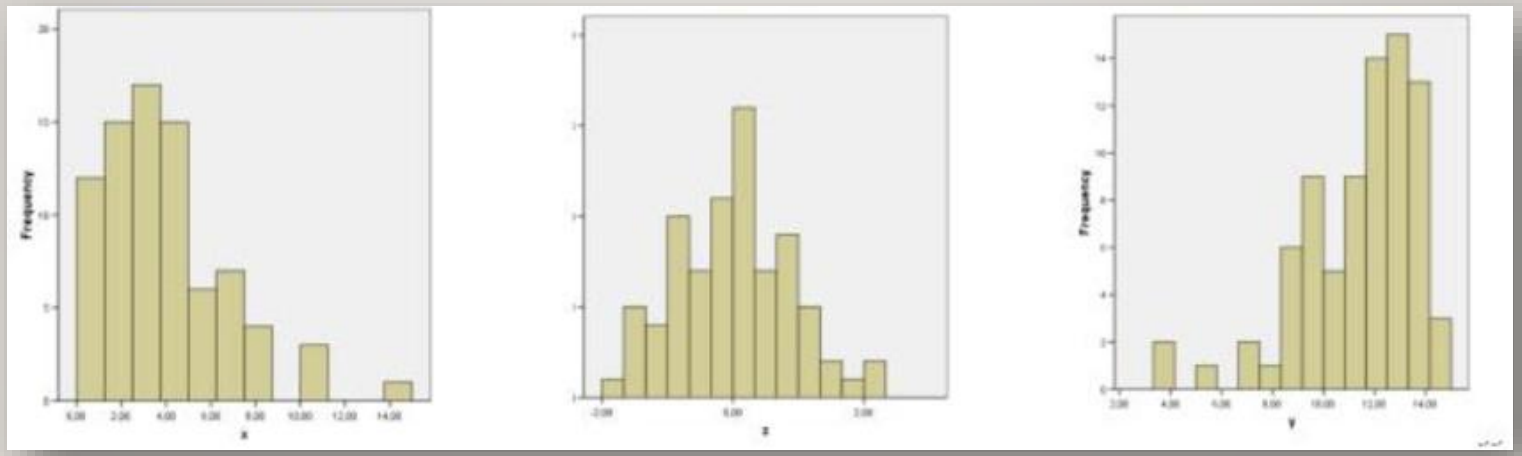
Newbold et al (2013)

SKEWNESS VS HISTOGRAM

Right-Skewed

Symmetric

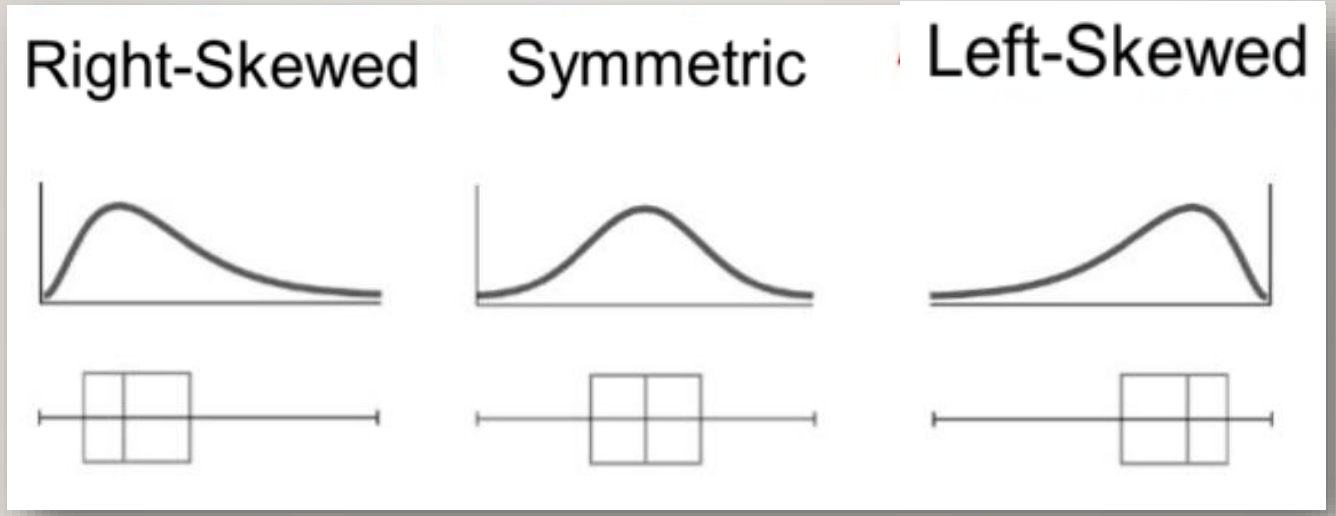
Left-Skewed



Right-skewed (positively skewed):
- The tail of the distribution extends more to the right (higher values).
- Most data are concentrated on the left.
- Skewness > 0

Left-skewed (negatively skewed):
- The tail of the distribution extends more to the left (lower values).
- Most data are concentrated on the right.
- Skewness < 0

SKEWNESS VS BOXPLOT

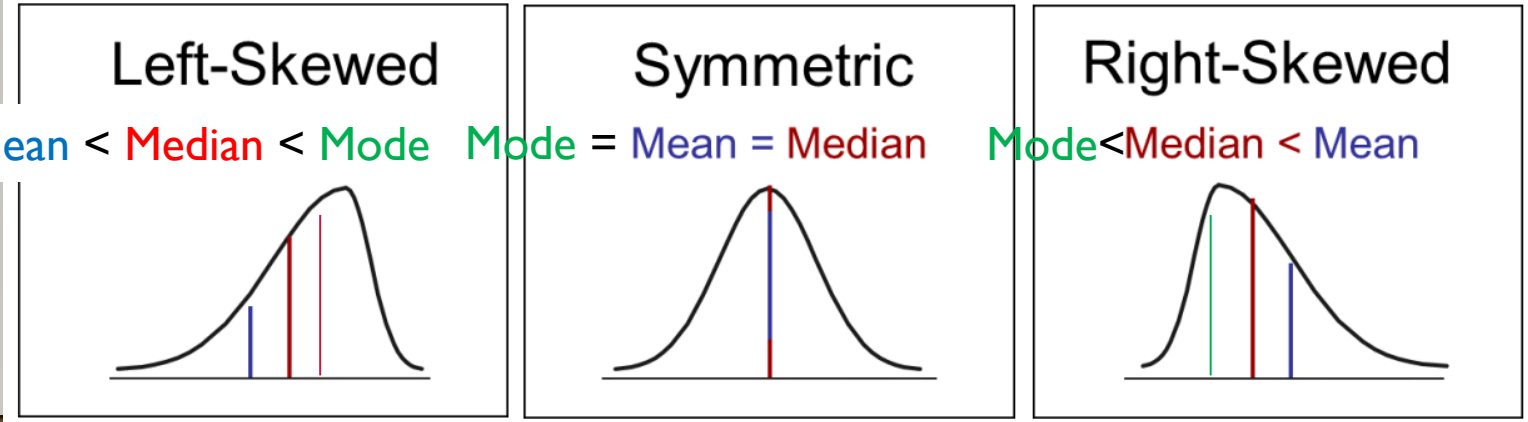


Newbold et al (2013)

IDENTIFYING SKEWNESS USING MEAN, MEDIAN, AND MODE

- Skewness can be identified by the relationship between Mode, Median, and Mean.

- Mean > Median > Mode → positively skewed (right-skewed)
- Mean < Median < Mode → negatively skewed (left-skewed)
- Mean ≈ Median ≈ Mode → approximately symmetric



SKEWNESS MEASURES: HOW TO IDENTIFY SKEWNESS WITH COEFFICIENTS

Measure	Formula (simplified)	Interpretation
Pearson's Coefficient of Skewness	$Sk_P = \frac{\bar{x} - \text{Mode}}{s}$ or $Sk_P = \frac{3(\bar{x} - \text{Median})}{s}$	> 0 → right-skewed < 0 → left-skewed = 0 → symmetric
Moment Coefficient of Skewness (SPSS)	$Sk = \frac{\frac{1}{n}\sum(x_i - \bar{x})^3}{s^3}$	> 0 → right-skewed < 0 → left-skewed = 0 → symmetric

PEARSON'S COEFFICIENT OF SKEWNESS: MODE VS MEDIAN

Two common formulas:

1. **Using the Mode:**

$$Sk_P = \frac{\bar{X} - \text{Mode}}{s}$$

- Requires a well-defined **mode**.
- Can be problematic if the distribution is **multimodal** or the mode is unclear.

2. **Using the Median:**

$$Sk_P = \frac{3(\bar{X} - \text{Median})}{s}$$

- Uses the **median**, which always exists.
- More stable in skewed distributions or with outliers.

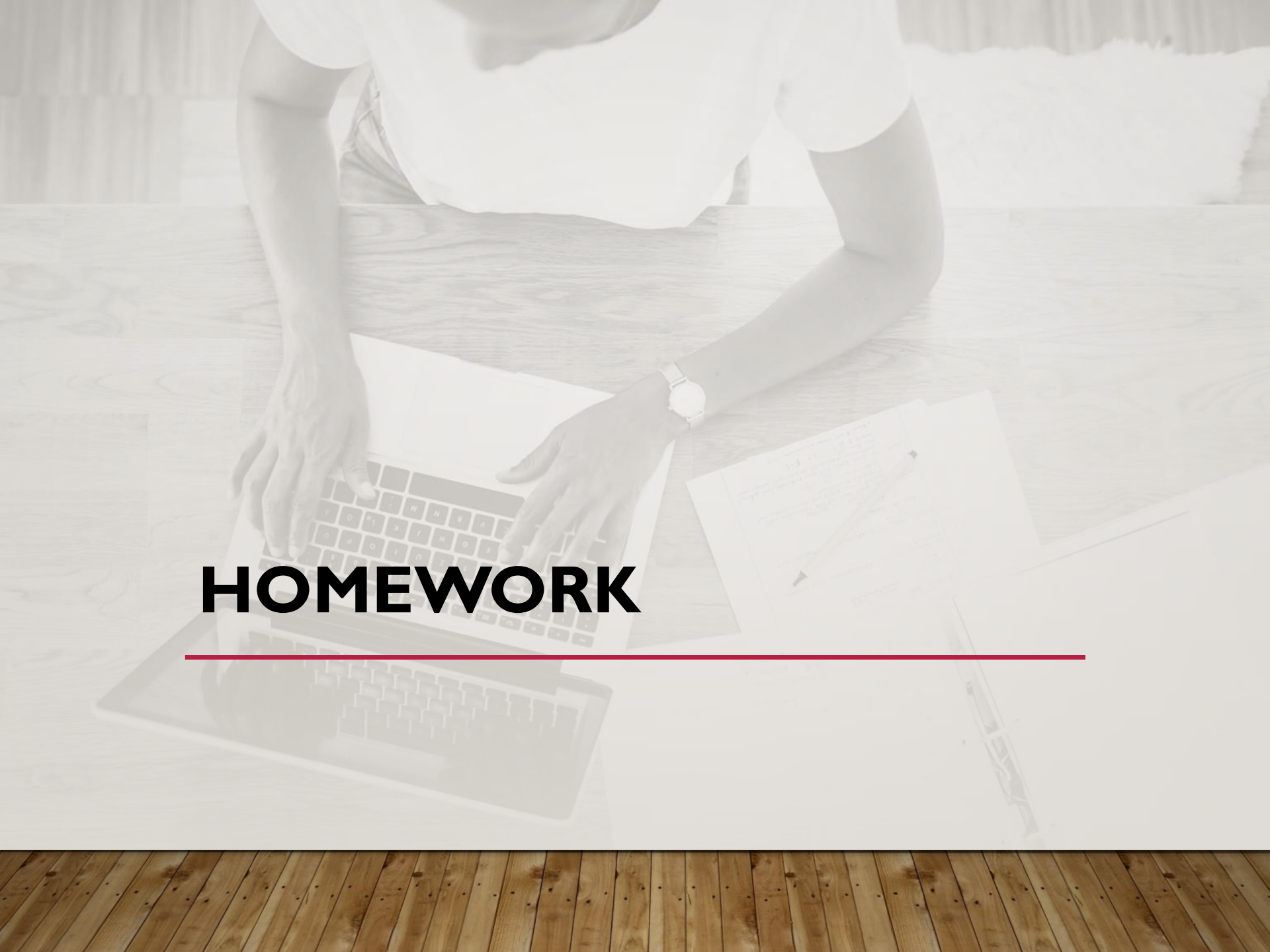
Key Points:

- The two formulas **do not necessarily give the same value**.
- The median-based formula is **more robust** and often preferred.
- For symmetric distributions, **both formulas give values close to zero**.

ADVANTAGES AND DISADVANTAGES OF SKEWNESS MEASURES

- **Pearson's Coefficient of Skewness**
  - ✅ Simple, intuitive (based on mean, mode, median).
  - ❌ Depends on mode/median (not always defined), less robust.

- **Moment Coefficient of Skewness (SPSS)**
  - ✅ Standardized, uses all data points, always available in software.
  - ❌ More abstract, sensitive to outliers.



HOMEWORK

EXERCISE 1.32

1.32 Consider the following data:

17	62	15	65
28	51	24	65
39	41	35	15
39	32	36	37
40	21	44	37
59	13	44	56
12	54	64	59

a. Construct a frequency distribution.
b. Construct a histogram.
c. Construct an ogive.
d. Construct a stem-and-leaf display.

Newbold et al (2013)

EXERCISE 2.8

2.8 The ages of a sample of 12 students enrolled in an on-line macroeconomics course are as follows:

21 22 27 36 18 19

22 23 22 28 36 33

a. What is the mean age for this sample?
b. Find the median age.
c. What is the value of the modal age?

Newbold et al (2013)

EXERCISE 2.14

2.14 Calculate the coefficient of variation for the following sample data:

10 8 11 7 9

Newbold et al (2013)

EXERCISE 2.15

2.15 The ages of a random sample of people who attended a recent soccer match are as follows:

23	35	14	37	38	15	45
12	40	27	13	18	19	23
37	20	29	49	40	65	53
18	17	23	27	29	31	42
35	38	22	20	15	17	21

a. Find the mean age.
b. Find the standard deviation.
c. Find the coefficient of variation.

Newbold et al (2013)

THANKS!

Questions?