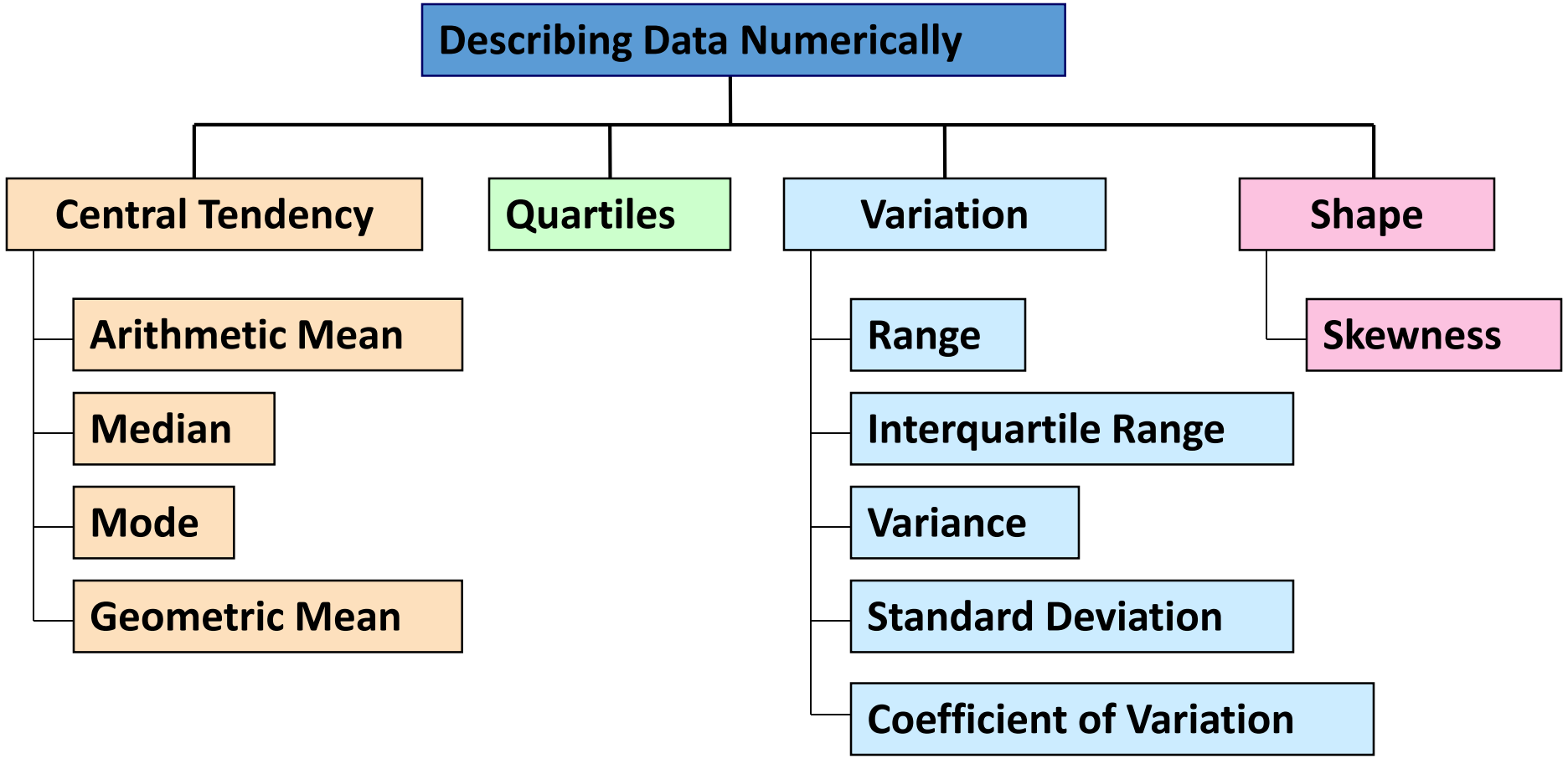
Measures of variation

Dr. Bornwell Sikateyo

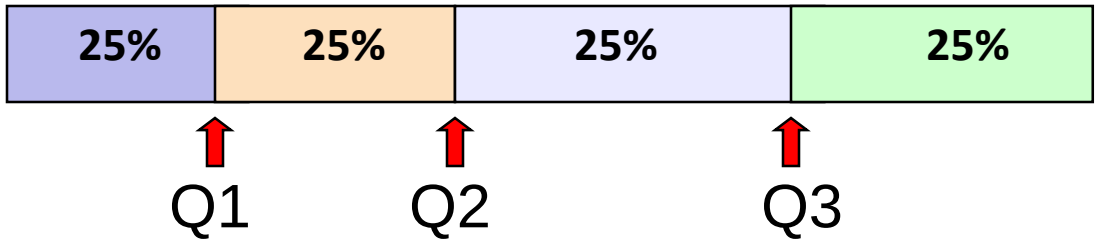
Department of Medical Education Development

Summary Measures



Quartiles

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



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Formulas

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $Q_1 = (n+1)/4$

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

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

where n is the number of observed values

Quartiles

- Example: Find the first quartile

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

(n = 9)

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

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

so **$Q_1 = 12.5$**

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

Quartiles

(continued)

- Example:

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

(n = 9)

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

so **$Q_1 = 12.5$**

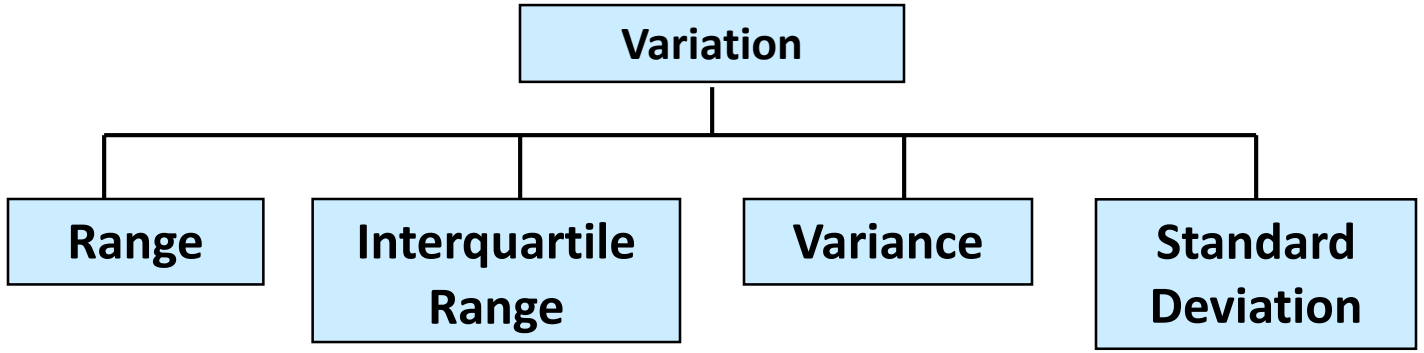
Q_2 is in the (9+1)/2 = 5^{th} position of the ranked data,

so **Q_2 = median = 16**

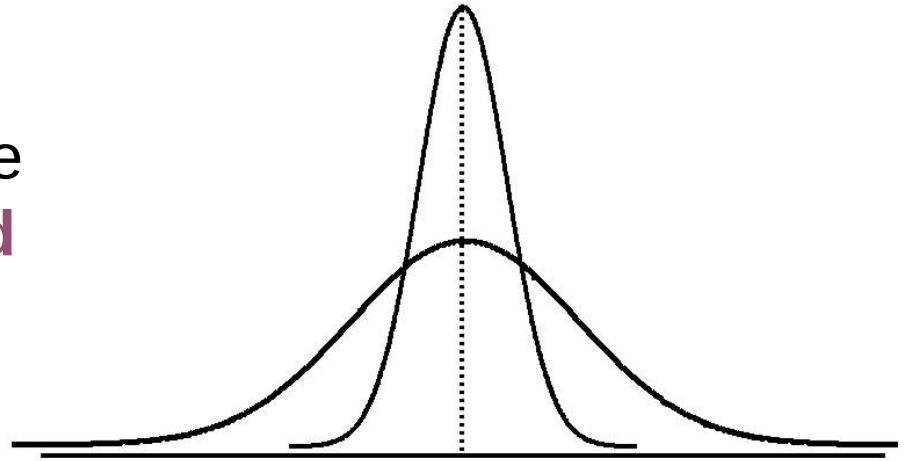
Q_3 is in the 3(9+1)/4 = 7.5 position of the ranked data,

so **$Q_3 = 19.5$**

Measures of Variation



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



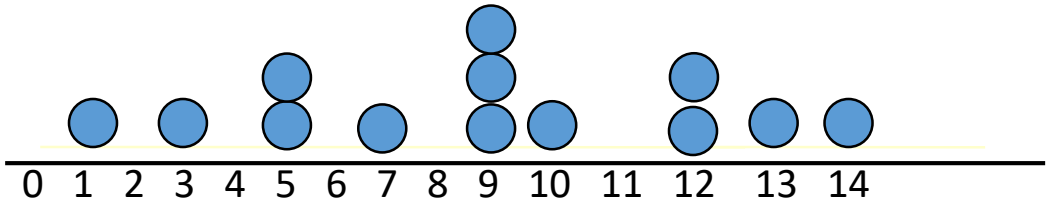
Same center, different variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest values in a set of data:

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

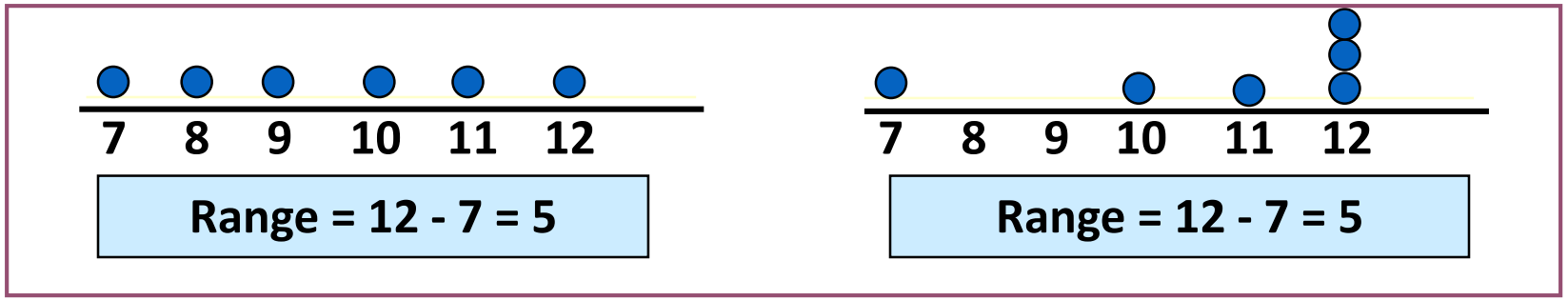
Example:



Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



- 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

- Can eliminate some outlier problems by using the **interquartile range**

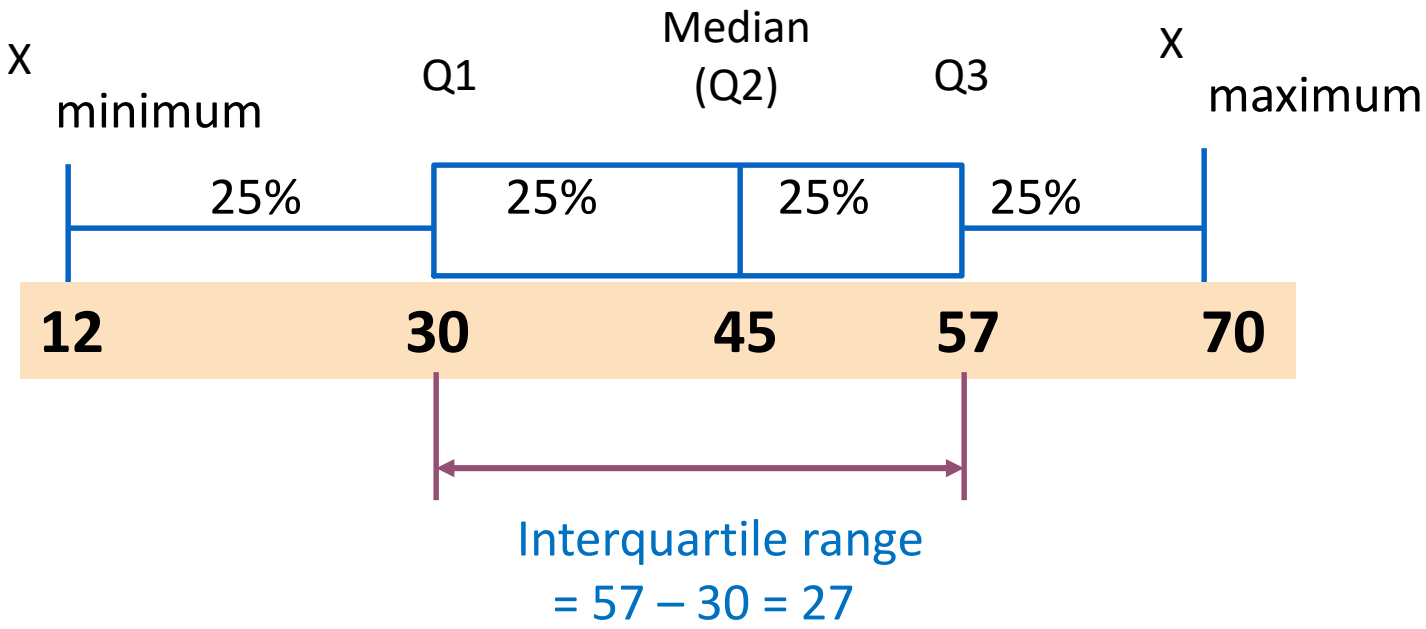
- Eliminate some high- and low-valued observations and calculate the range from the remaining values

- Interquartile range = 3^{rd} quartile – 1^{st} quartile

  $= Q_3 - Q_1$

Interquartile Range

Example:



- **Quartiles:** The values which divide a series of observations, arranged in ascending order, into 4 equal parts. (Thus the 2nd Quartile is the Median).

- **The <u>Interquartile Range</u>** represents the central portion of the distribution and is calculated as the difference between the third quartile and the first quartile.

- This range includes about one-half of the observations in the set, leaving one quarter of the observations on each side.

So how do we get a single mathematical measure to summarise the variability of an observed set of values?

- The most frequent and most informative measure is the VARIANCE and its related functions

- The variance is computed in stages:

Steps in calculation Standard deviation

- 1. Calculate the mean as a measure of central location (MEAN)

- 2. Calculate the difference between each observation and the mean (DEVIATION)

  $(x-\bar{x})$

- 3. Next square the differences (SQUARED DEVIATION)

  $(x-\bar{x})^2$

- Q. What is the effect of this ?

  - Negative and positive deviations will not cancel each other out.
  - Values further from the mean have a bigger impact.

Steps in calculation Standard deviation

- **4. Sum up these squared deviations (SUM OF THE SQUARED DEVIATIONS)**

$$\Sigma (x - \bar{x})^2$$

- **5. Divide this SUM OF THE SQUARED DEVIATIONS by the total number of observations minus 1 (n-1) to give the VARIANCE**

$$\frac{\Sigma (x - \bar{x})^2}{n - 1}$$

This is a measure of the variability of the data

Why divide by n - 1 ?

This is an adjustment for the fact that the mean is just an estimate of the true population mean. It tends to make the variance bigger.

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 $\overline{X}$ = mean

n = sample size

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

Standard Deviation

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

- Sample standard deviation:

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

Calculation Example: Sample Standard Deviation

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

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

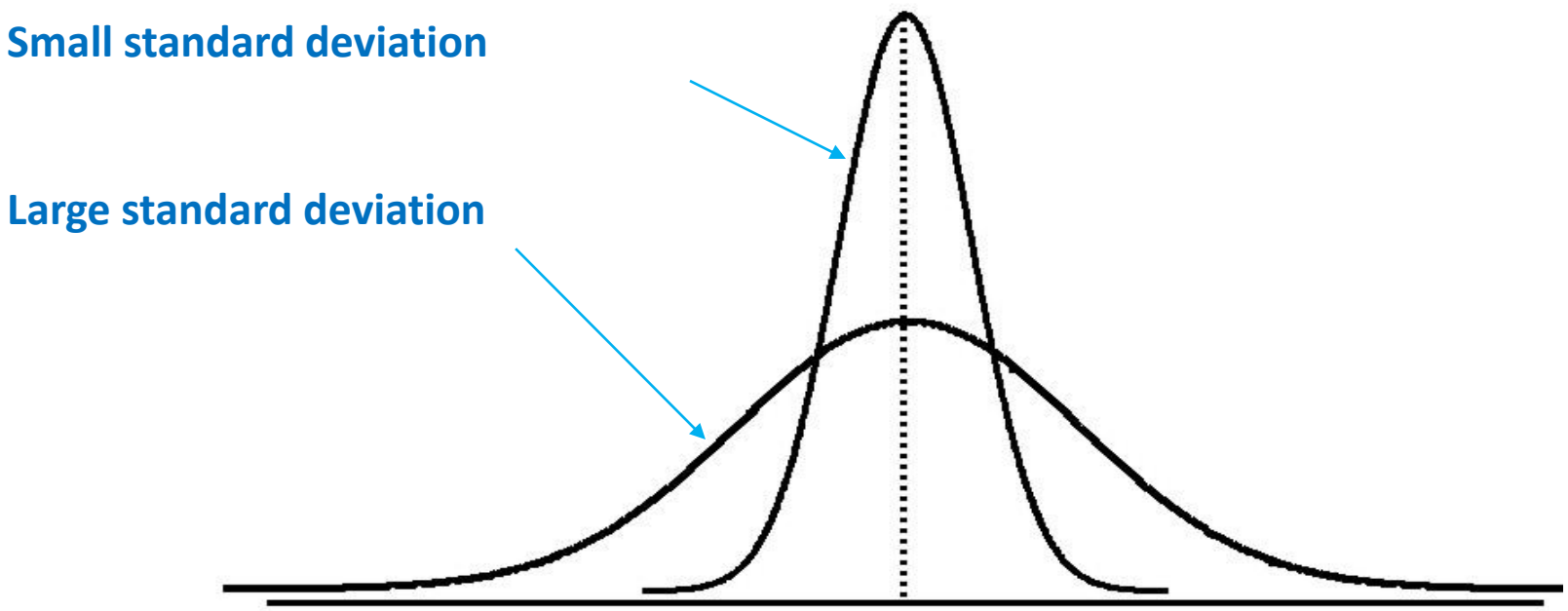
$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

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

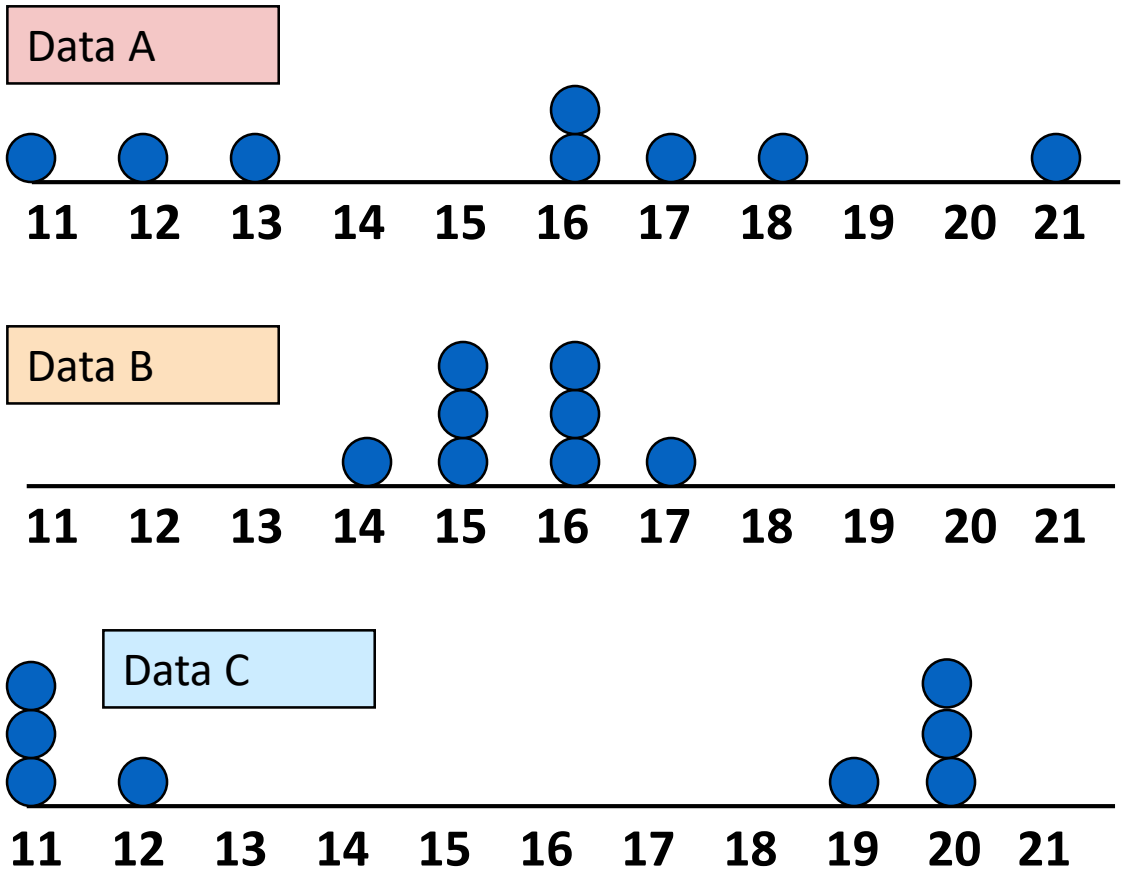
$$= \sqrt{\frac{130}{7}} = 4.3095$$

A measure of the "average" scatter around the mean

Comparing Standard Deviations



Comparing Standard Deviations



Mean = 15.5
S = 3.338

Mean = 15.5
S = 0.926

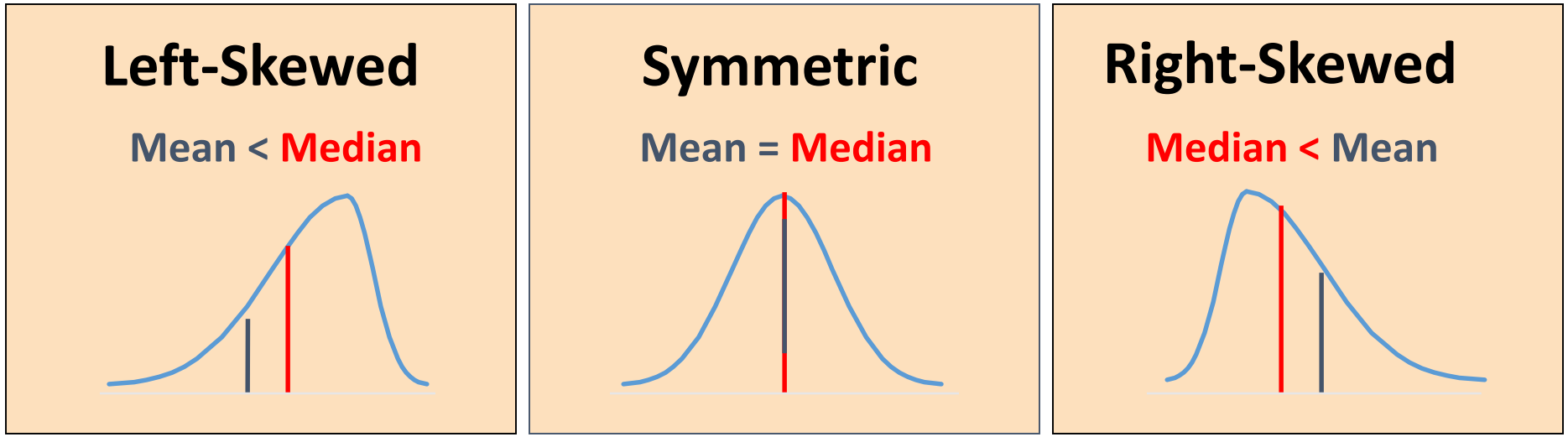
Mean = 15.5
S = 4.567

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)

Shape of a Distribution

- Describes how data are distributed
- Measures of shape
  - Symmetric or skewed



Population Variance

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

Where μ = population mean

N = population size

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

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the population variance
- Has the same units as the original data

- Population standard deviation:

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

Approximating the Standard Deviation from a Frequency Distribution

- Assume that all values within each class interval are located at the midpoint of the class

  - Approximation for the standard deviation from a frequency distribution:

$$S = \sqrt{\frac{\sum_{j=1}^{c} (m_j - \overline{X})^2 \, f_j}{n-1}}$$

	Measure		
Type of Distribution	**Central location**	**Variation or dispersion**	**Example**
Normal	Arithmetic Mean	Standard Deviation	Mean Age: M (SD) 34.0 (3)
Skewed	Median	Interquartile range	Median Age: Med (IQR) 34 (23, 37)

The Empirical Rule

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

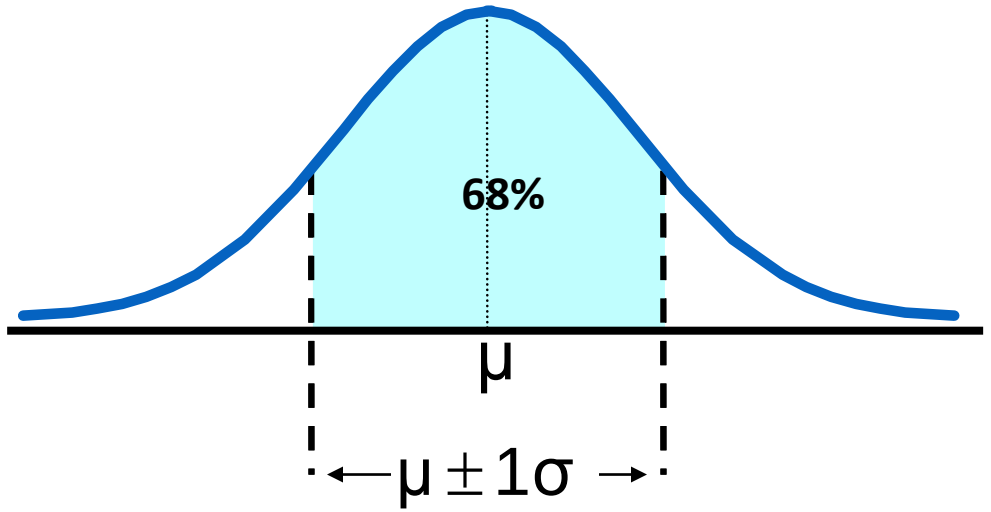
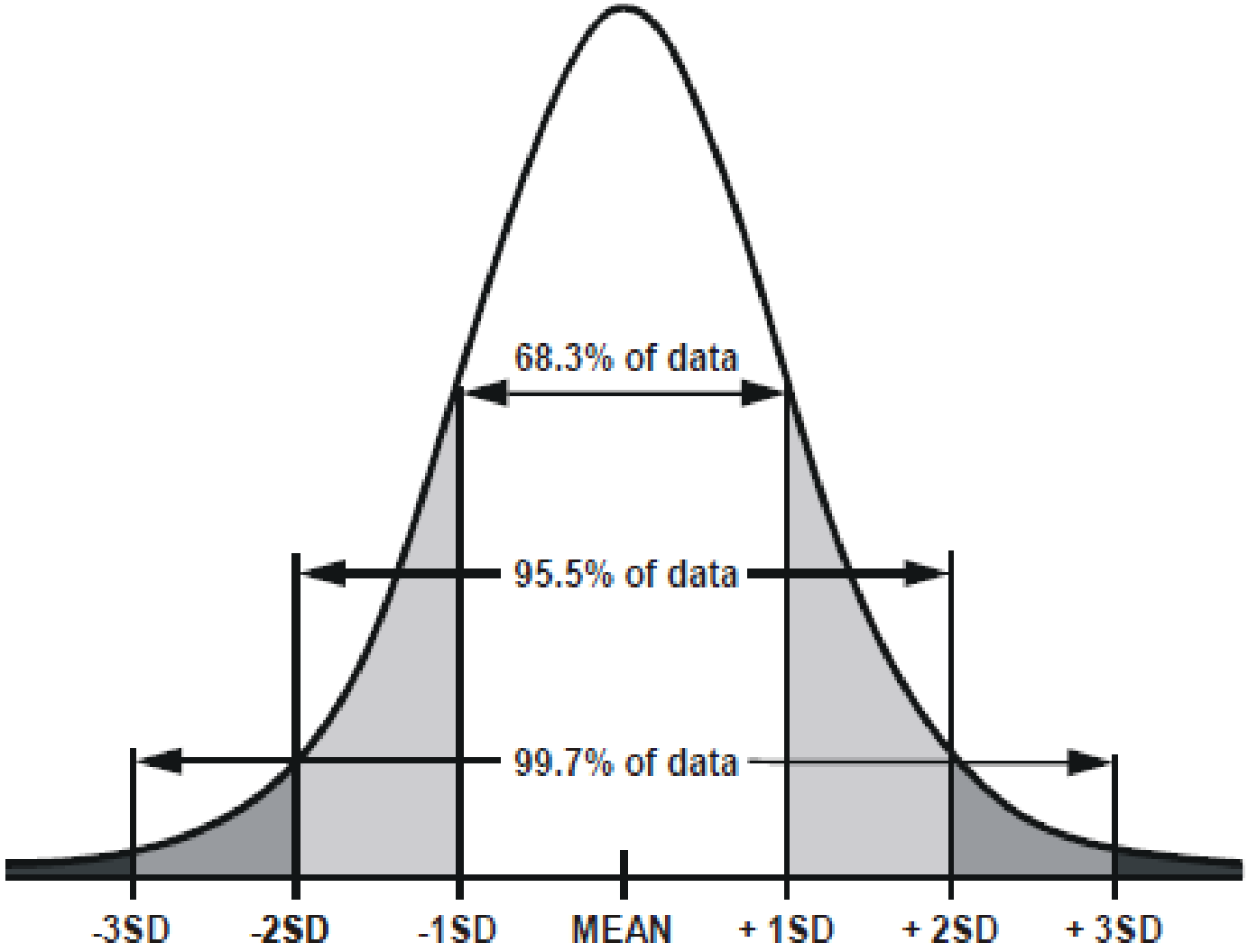
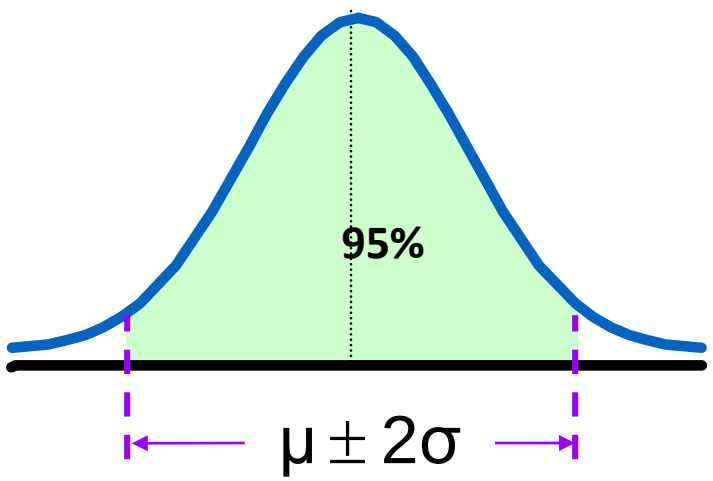


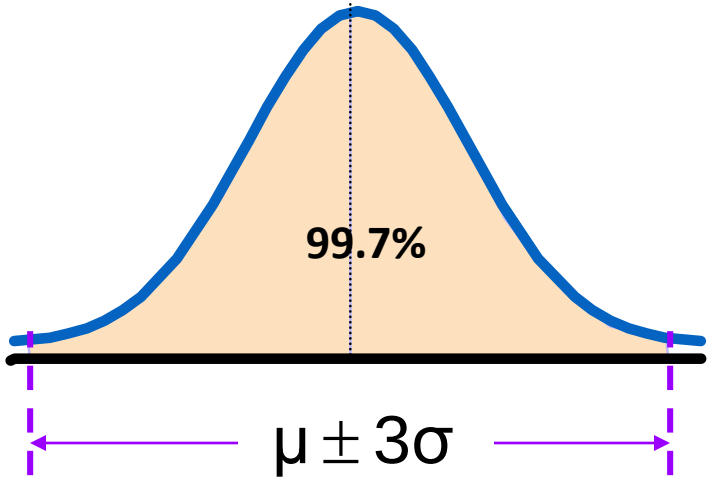
Figure 3.9
Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean



The Empirical Rule

- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains about 99.7% of the values in the population or the sample





Pitfalls in Numerical Descriptive Measures

- **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 fair, neutral and clear manner

Lecture Summary

- Described measures of central tendency
  - Mean, median, mode
- Discussed quartiles
- Described measures of variation
  - Range, interquartile range, variance and standard deviation
- Illustrated shape of distribution
  - Symmetric, skewed