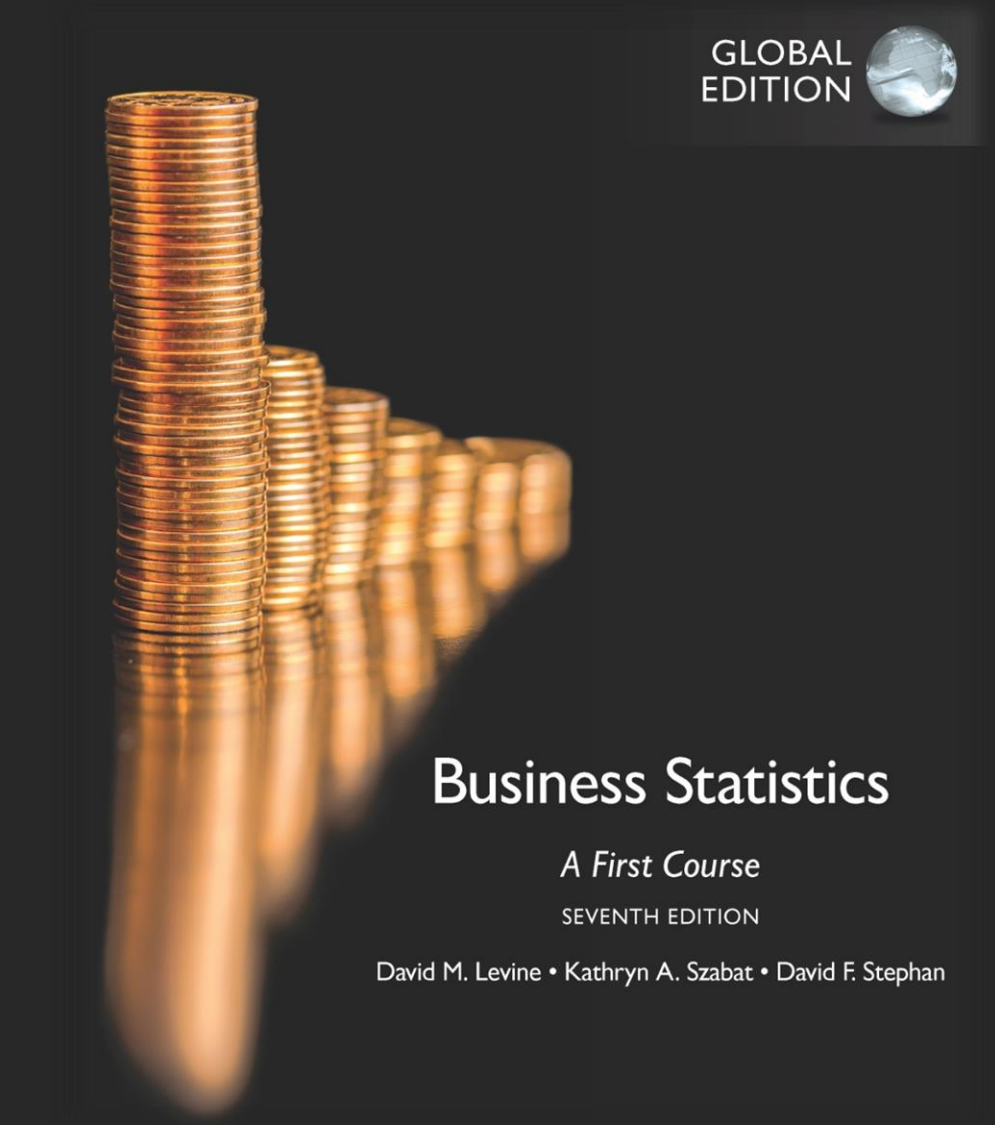



# Chapter 3

## Numerical Descriptive Measures

# Objectives

**In this chapter, you learn to:**

- Describe the properties of central tendency, variation, and shape in numerical data
- Construct and interpret a boxplot
- Compute descriptive summary measures for a population
- Calculate the covariance and the coefficient of correlation

# Summary Definitions

DCOVA

- The **central tendency** is the extent to which the values of a numerical variable group around a typical or central value.

- The **variation** is the amount of dispersion or scattering away from a central value that the values of a numerical variable show.

- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.

# Measures of Central Tendency: The Mean

DCOVA

- The arithmetic mean (often just called the “mean”) is the most common measure of central tendency
  - For a sample of size n:

Pronounced x-bar

The $i^{th}$ value

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

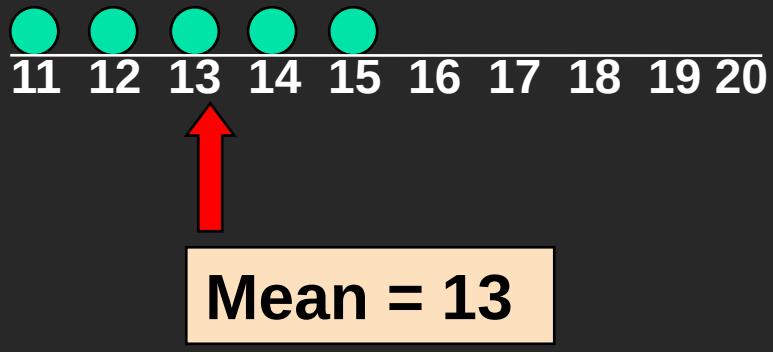
Sample size

Observed values

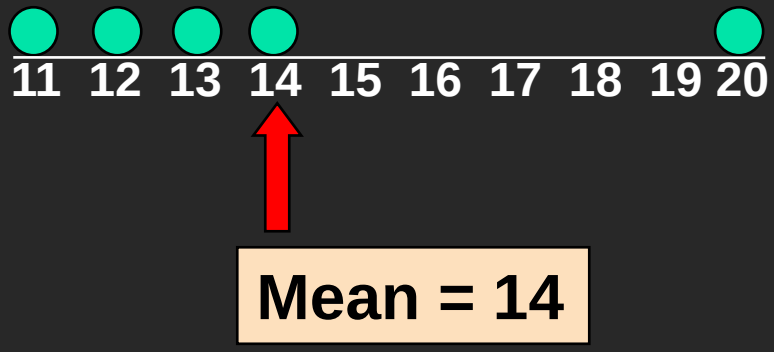
# Measures of Central Tendency: The Mean (con't)

DCOVA

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




$$\frac{11+12+13+14+15}{5}=\frac{65}{5}=13$$




$$\frac{11+12+13+14+20}{5}=\frac{70}{5}=14$$

# Numerical Descriptive Measures for a Population

DCOVA

- Descriptive statistics discussed previously described a *sample*, not the *population*.
- Summary measures describing a population, called **parameters**, are denoted with Greek letters.
- Important population parameters are the population mean, variance, and standard deviation.

# Numerical Descriptive Measures for a Population: The mean μ

DCOVA

- The population mean is the sum of the values in the population divided by the population size, N

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

Where μ = population mean

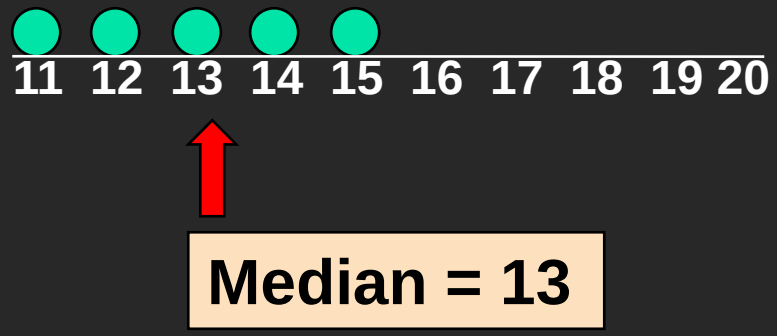
N = population size

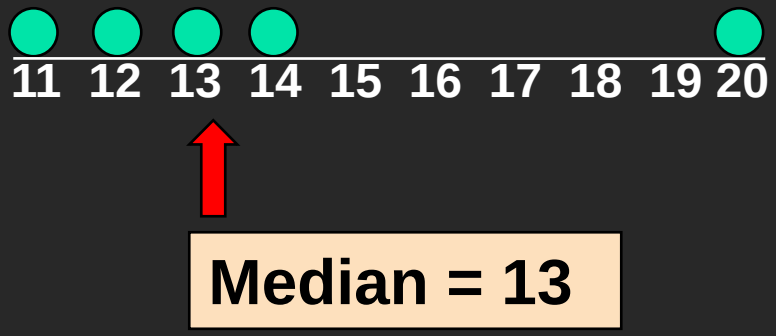
$X_i$ = $i^{th}$ value of the variable X

# Measures of Central Tendency: The Median

DCOVA

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




- Less sensitive than the mean to extreme values

# Measures of Central Tendency: Locating the Median

DCOVA

- The location of the median when the values are in numerical order (smallest to largest):

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- If the number of values is odd, the median is the middle number

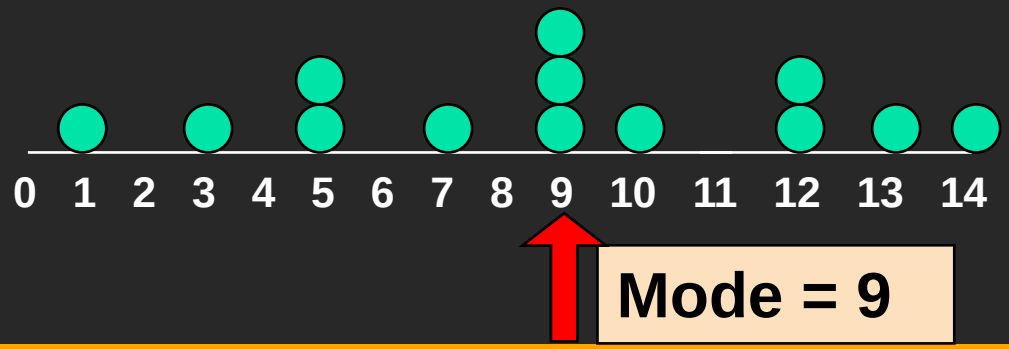
- If the number of values is even, the median is the average of the two middle numbers

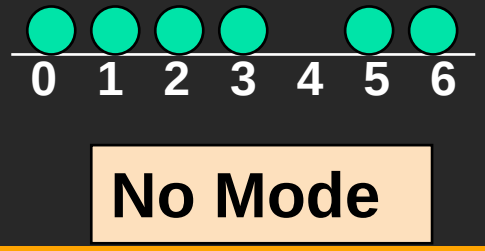
Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

# Measures of Central Tendency: The Mode

DCOVA

- 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



# Measures of Central Tendency: Review Example

DCOVA

**House Prices:**

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

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**

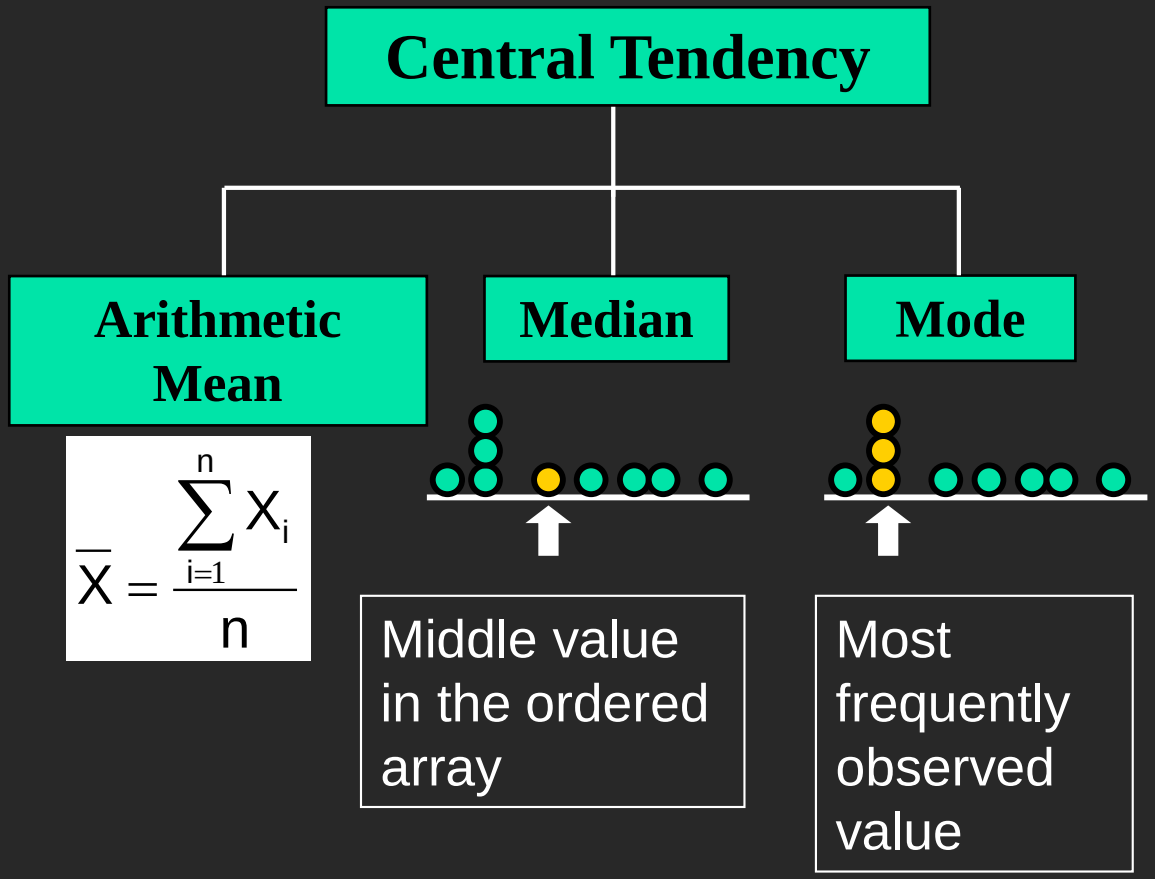
# Measures of Central Tendency: Which Measure to Choose?

DCOVA

- The **mean** is generally used, unless extreme values (outliers) exist.
- The **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.
- In some situations it makes sense to report both the **mean** and the **median**.

# Measures of Central Tendency: Summary

DCOVA



# Shape of a Distribution

DCOVA

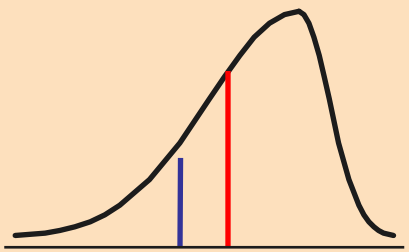
- Describes how data are distributed
- Two useful shape related statistics are:
  - Skewness
    - Measures the extent to which data values are not symmetrical
  - Kurtosis
    - Kurtosis affects the peakedness of the curve of the distribution—that is, how sharply the curve rises approaching the center of the distribution

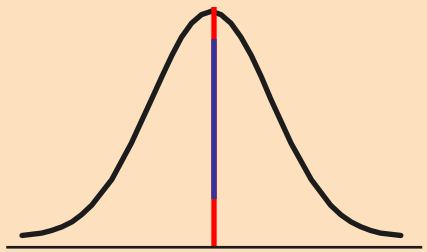
# Shape of a Distribution (Skewness)

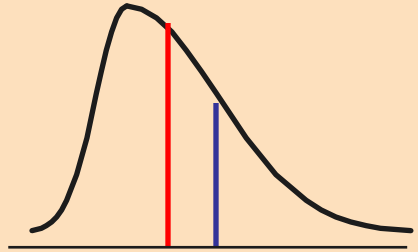
DCOVA

- Measures the extent to which data is not symmetrical

| Left-Skewed | Symmetric | Right-Skewed |
|---|---|---|
| Mean < Median | Mean = Median | Median < Mean |



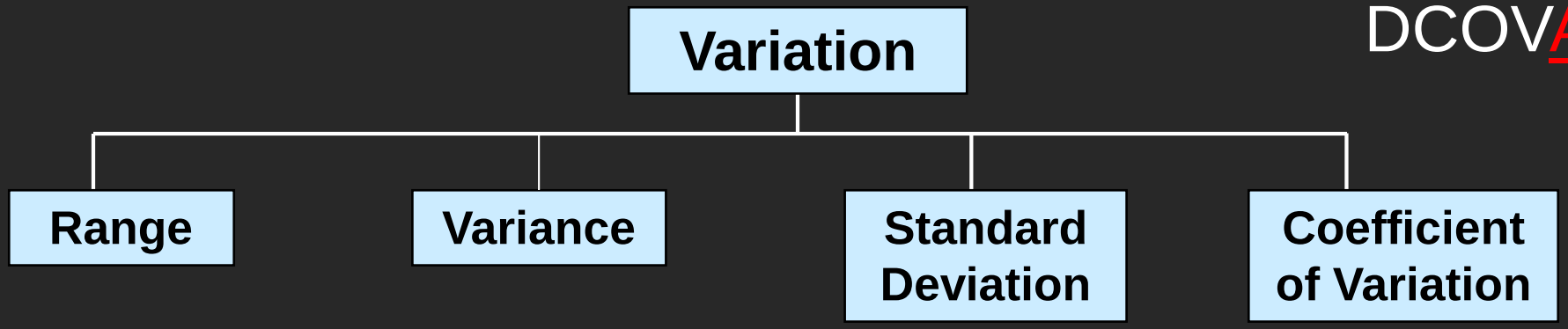




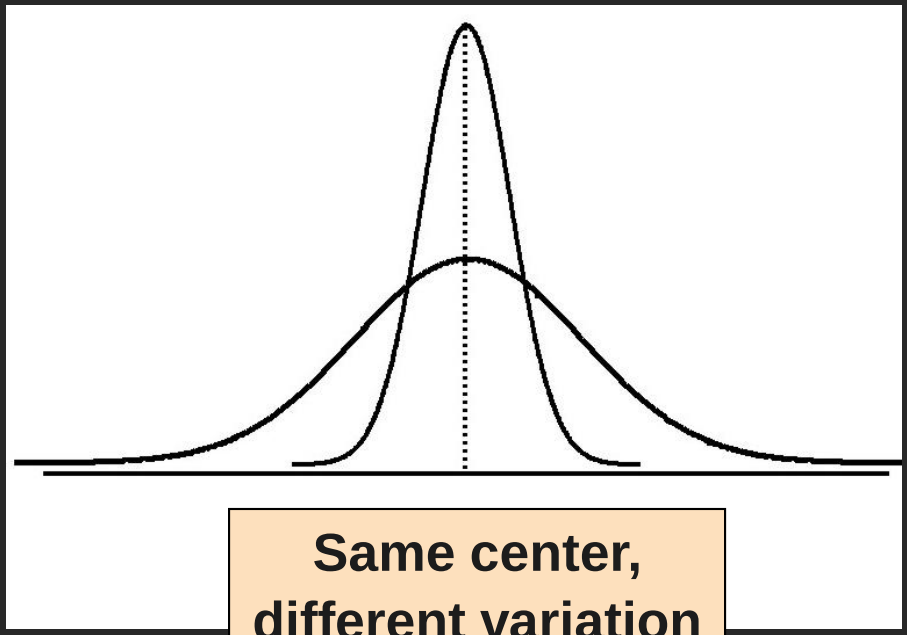
| Skewness Statistic | < 0 | 0 | >0 |
|---|---|---|---|

# Measures of Variation

DCOVA




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



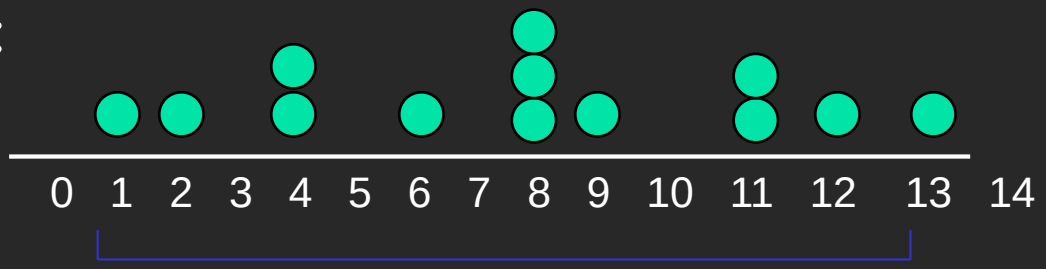
# Measures of Variation: The Range

DCOVA

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

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

Example:




**Range = 13 - 1 = 12**

# Measures of Variation: Why The Range Can Be Misleading

DCOVA

- Does not account for how the 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**

# Measures of Variation: The Sample Variance

DCOVA

- 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

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

# Measures of Variation: The Sample Standard Deviation

DCOVA

- 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:

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

# Measures of Variation: The Standard Deviation

DCOVA

Steps for Computing Standard Deviation

1. Compute the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by n-1 to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

# Measures of Variation: Sample Standard Deviation: Calculation Example

DCOVA

**Sample Data ($X_i$) :**

| 10 | 12 | 14 | 15 | 17 | 18 | 18 | 24 |
|---|---|---|---|---|---|---|---|

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

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

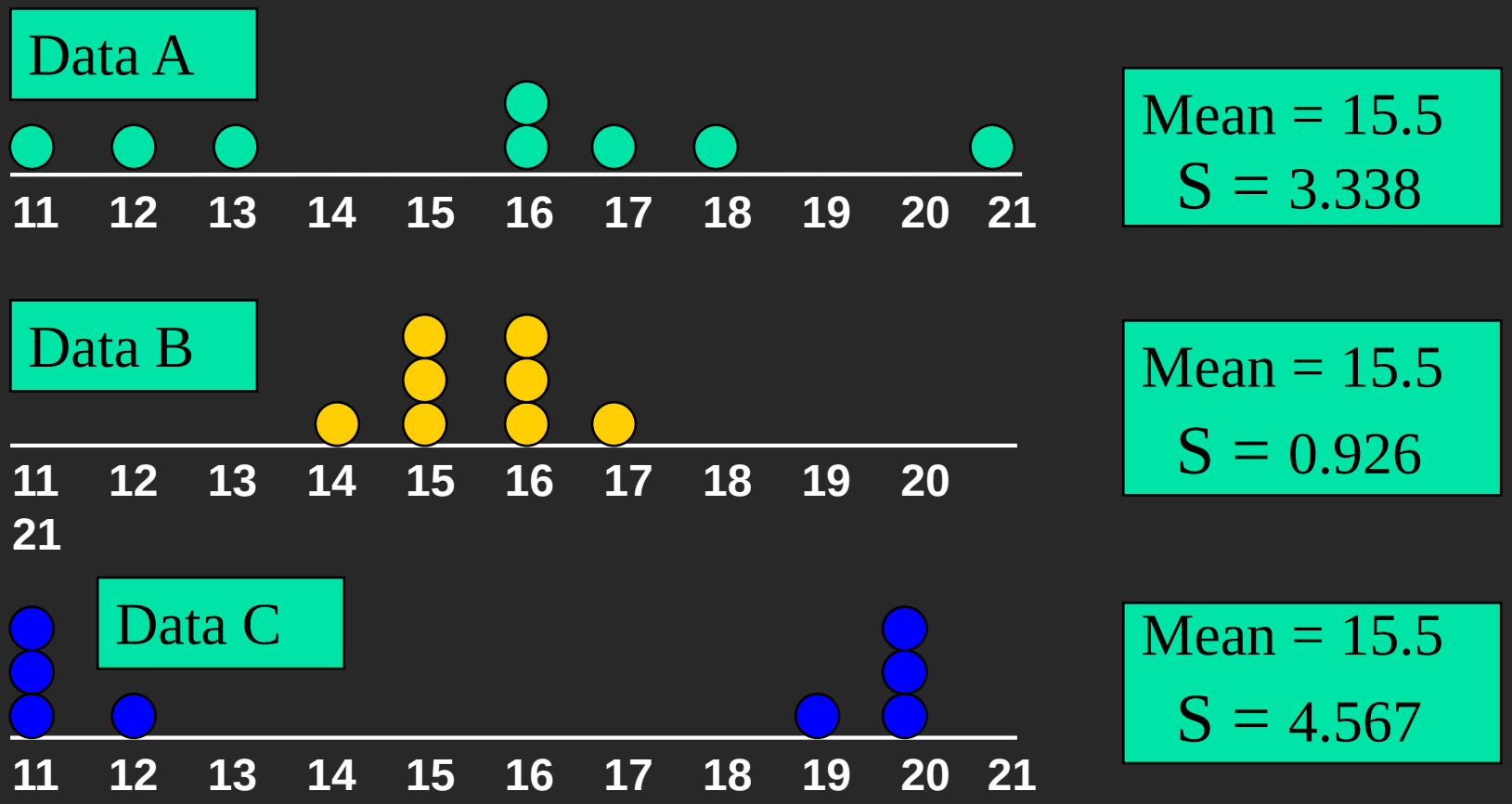
$$= \sqrt{\frac{(10-16)^2+(12-16)^2+(14-16)^2+\cdots+(24-16)^2}{8-1}}$$

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

A measure of the "average" scatter around the mean

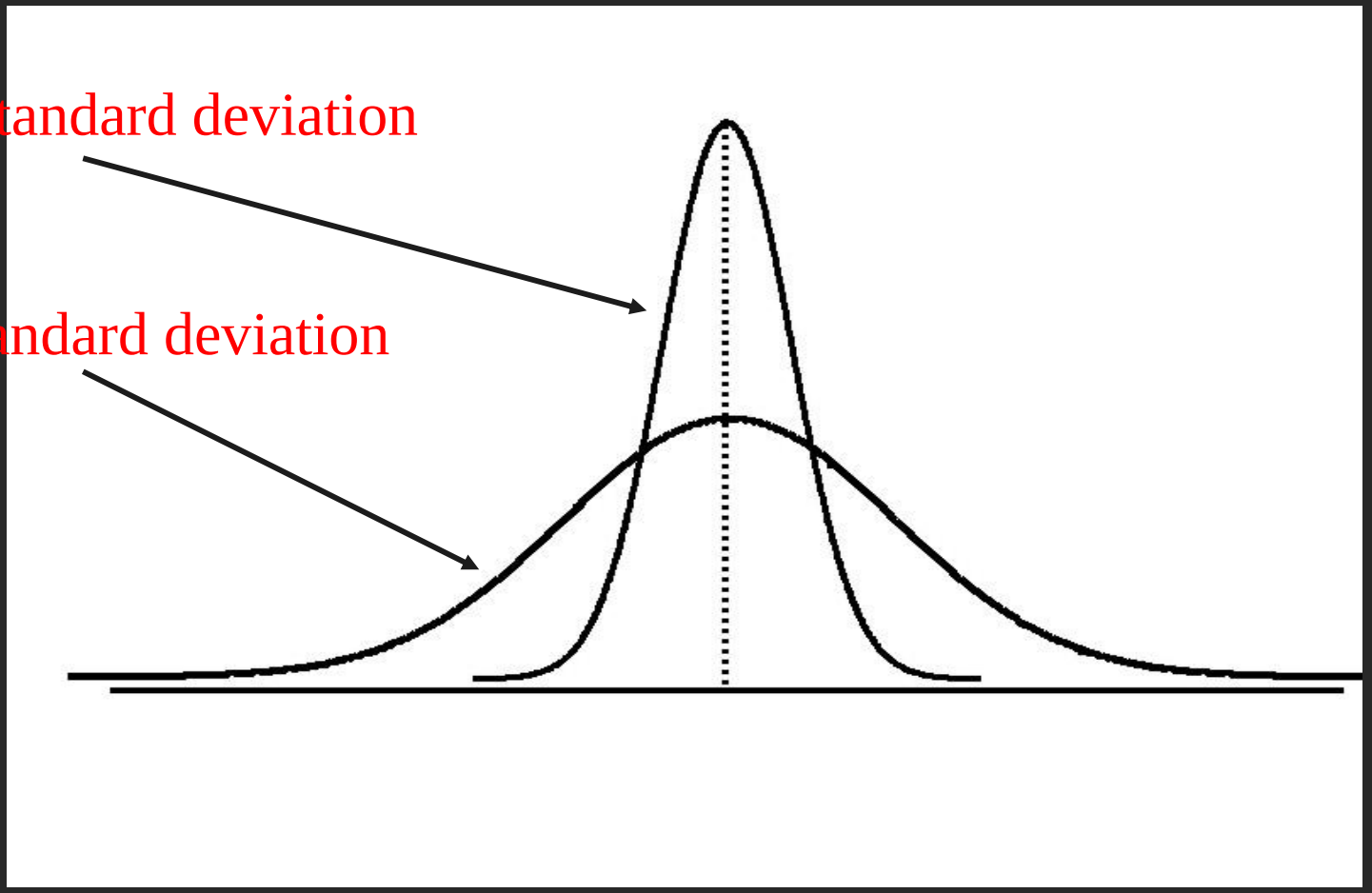
# Measures of Variation: Comparing Standard Deviations

DCOVA



# Measures of Variation: Comparing Standard Deviations

DCOVA



# Numerical Descriptive Measures For A Population: The Variance $\sigma^2$

DCOVA

- Average of squared deviations of values from the mean

  - Population variance:

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

Where $\mu$ = population mean

$N$ = population size

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

# Numerical Descriptive Measures For A Population: The Standard Deviation σ

DCOVA

- 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:

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

# Sample statistics versus population parameters

DCOVA

| Measure | Population Parameter | Sample Statistic |
|---|---|---|
| Mean | $\mu$ | $\overline{X}$ |
| Variance | $\sigma^2$ | $S^2$ |
| Standard Deviation | $\sigma$ | $S$ |

# Measures of Variation: Summary Characteristics

DCOVA

- The more the data are spread out, the greater the range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

# Measures of Variation: The Coefficient of Variation

DCOVA

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare the variability of two or more sets of data measured in different units

$$CV = \left( \frac{S}{\bar{X}} \right) \cdot 100\%$$

# Measures of Variation: Comparing Coefficients of Variation

DCOVA

- 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

# Measures of Variation: Comparing Coefficients of Variation (con't)

DCOVA

- Stock A:
  - Average price last year = \$50
  - Standard deviation = \$5

$$CV_A = \left(\frac{S}{\overline{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock C:
  - Average price last year = \$8
  - Standard deviation = \$2

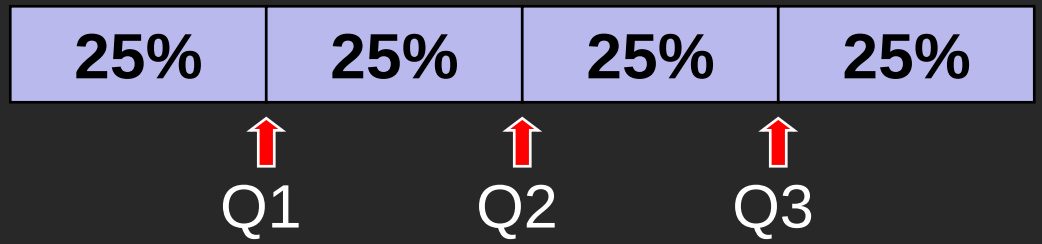
$$CV_C = \left(\frac{S}{\overline{X}}\right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

Stock C has a much smaller standard deviation but a much higher coefficient of variation

# Quartile Measures

DCOVA

- 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% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than the third quartile

# Quartile Measures: Locating Quartiles

DCOVA

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

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

Second quartile position: $Q_2 = (n+1)/2$ ranked value

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

where **n** is the number of observed values

# Quartile Measures: Calculation Rules

DCOVA

- When calculating the ranked position use the following rules
  - If the result is a whole number then it is the ranked position to use
  - If the result is a fractional half (e.g. 2.5, 7.5, 8.5, etc.) then average the two corresponding data values.
  - If the result is not a whole number or a fractional half then round the result to the nearest integer to find the ranked position.

# Quartile Measures: Locating Quartiles

DCOVA

**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, is a measure of central tendency

# Quartile Measures
# Calculating The Quartiles: Example

DCOVA

**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 **$Q_1 = (12+13)/2 = 12.5$**

$Q_2$ 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 **$Q_3 = (18+21)/2 = 19.5$**

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

# Quartile Measures: The Interquartile Range (IQR)

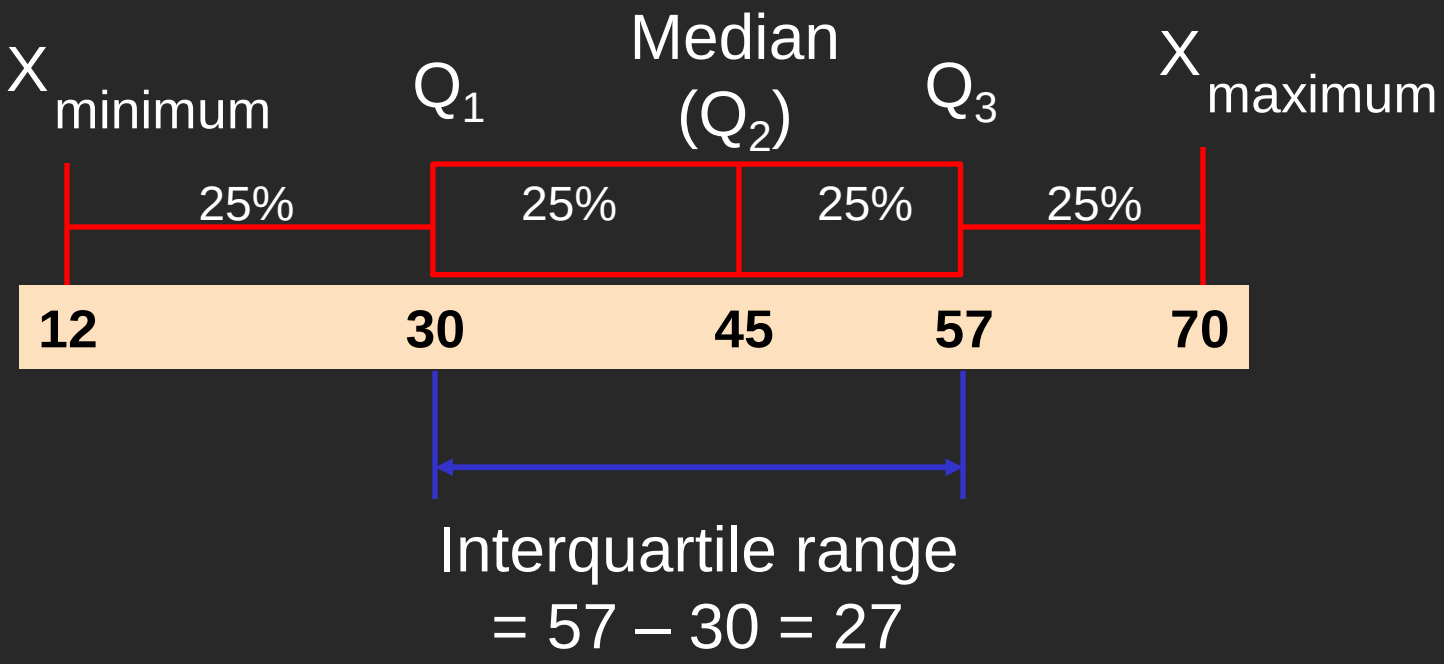
DCOVA

- The IQR is $Q_3 - Q_1$ and measures the spread in the middle 50% of the data
- The IQR is also called the midspread because it covers the middle 50% of the data
- The IQR is a measure of variability that is not influenced by outliers or extreme values
- Measures like $Q_1$, $Q_3$, and IQR that are not influenced by outliers are called resistant measures

# Calculating The Interquartile Range

DCOVA

Example:



# The Five Number Summary

DCOVA

The five numbers that help describe the center, spread and shape of data are:

- $X_{smallest}$
- First Quartile ($Q_1$)
- Median ($Q_2$)
- Third Quartile ($Q_3$)
- $X_{largest}$

# Relationships among the five-number summary and distribution shape

DCOVA

| Left-Skewed | Symmetric | Right-Skewed |
|---|---|---|
| $\text{Median} - X_{\text{smallest}}$ | $\text{Median} - X_{\text{smallest}}$ | $\text{Median} - X_{\text{smallest}}$ |
| $>$ | $\approx$ | $<$ |
| $X_{\text{largest}} - \text{Median}$ | $X_{\text{largest}} - \text{Median}$ | $X_{\text{largest}} - \text{Median}$ |
| $Q_1 - X_{\text{smallest}}$ | $Q_1 - X_{\text{smallest}}$ | $Q_1 - X_{\text{smallest}}$ |
| $>$ | $\approx$ | $<$ |
| $X_{\text{largest}} - Q_3$ | $X_{\text{largest}} - Q_3$ | $X_{\text{largest}} - Q_3$ |
| $\text{Median} - Q_1$ | $\text{Median} - Q_1$ | $\text{Median} - Q_1$ |
| $>$ | $\approx$ | $<$ |
| $Q_3 - \text{Median}$ | $Q_3 - \text{Median}$ | $Q_3 - \text{Median}$ |

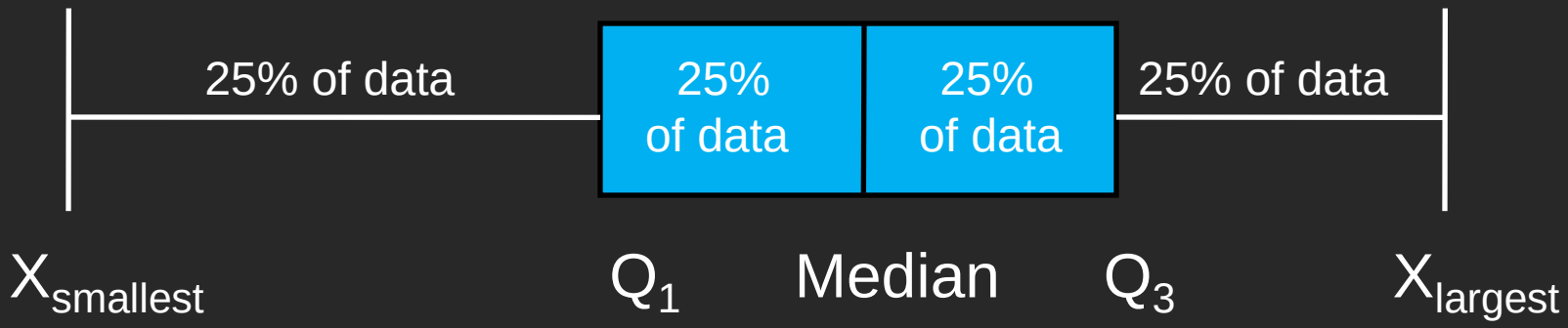
# Five Number Summary and The Boxplot

DCOVA

- The Boxplot: A Graphical display of the data based on the five-number summary:

$X_{smallest}$ -- $Q_1$ -- Median -- $Q_3$ -- $X_{largest}$

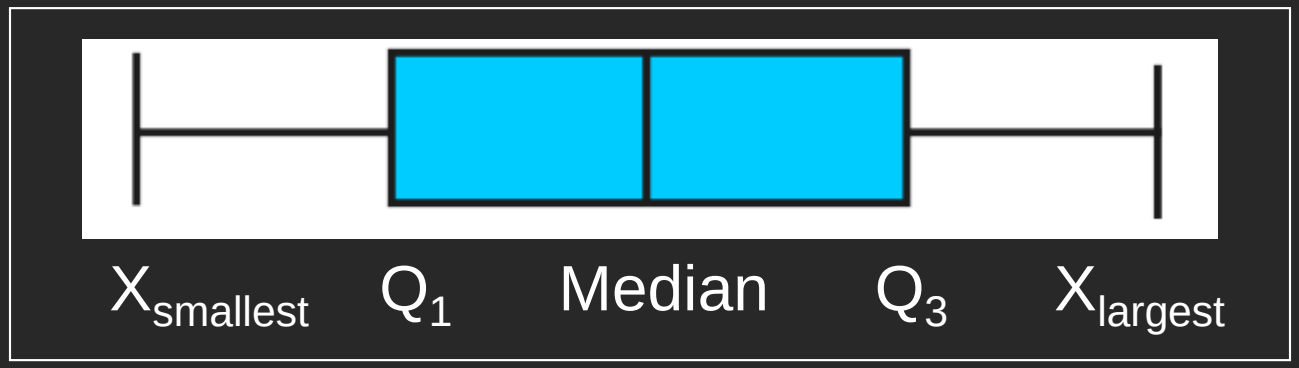
Example:



# Five Number Summary: Shape of Boxplots

DCOVA

- If data are symmetric around the median then the box and central line are centered between the endpoints

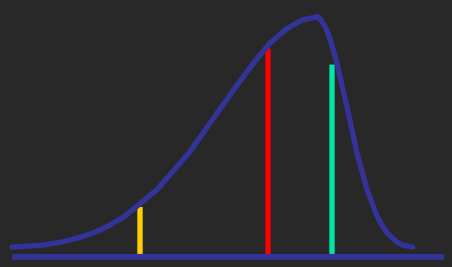



- A Boxplot can be shown in either a vertical or horizontal orientation

# Distribution Shape and The Boxplot

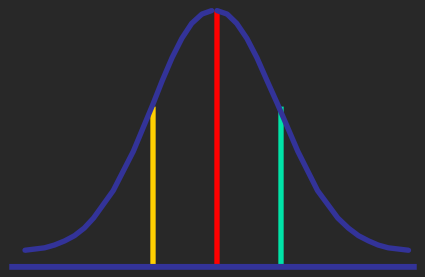
DCOVA

Left-Skewed



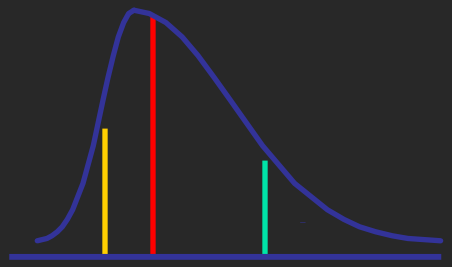
$Q_1$ $Q_2$ $Q_3$

Symmetric



$Q_1$ $Q_2$ $Q_3$

Right-Skewed

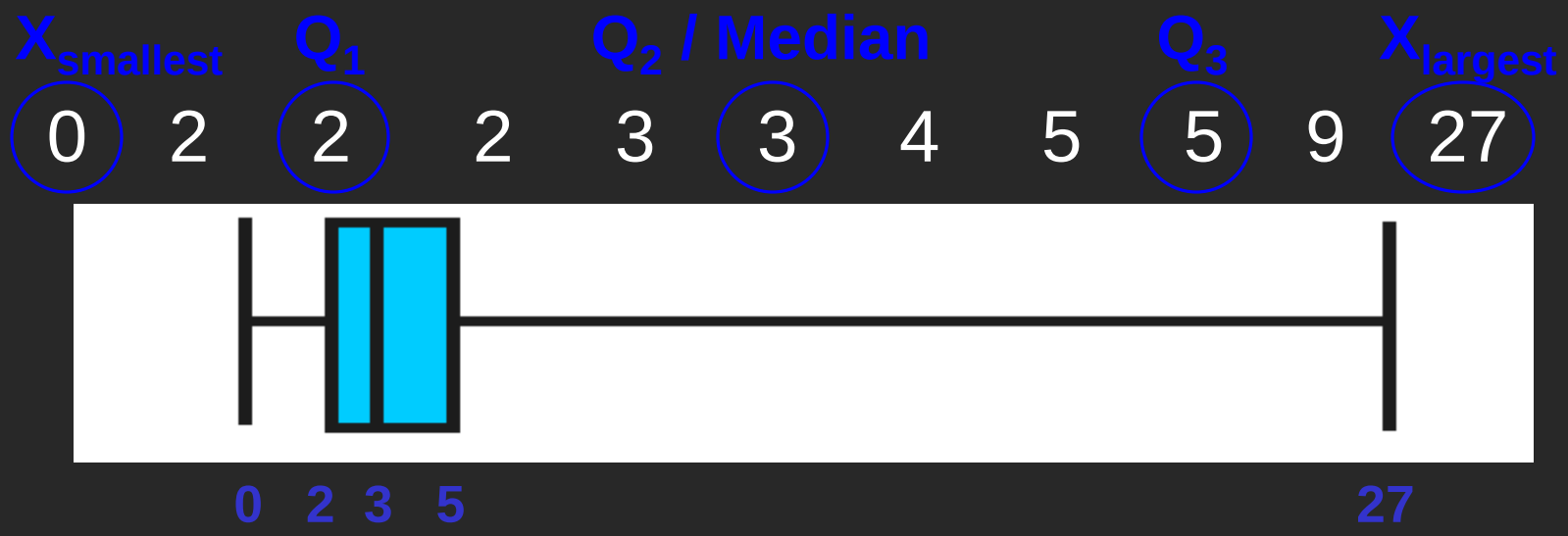


$Q_1$ $Q_2$ $Q_3$

# Boxplot Example

DCOVA

- Below is a Boxplot for the following data:




- The data are right skewed, as the plot depicts

# Locating Extreme Outliers: Z-Score

**(It is studied with Chapter 6)**

DCOVA

- To compute the Z-**score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

# Locating Extreme Outliers: Z-Score

DCOVA

$$Z = \frac{X - \overline{X}}{S}$$

where X represents the data value

$\overline{X}$ is the sample mean

S is the sample standard deviation

# Locating Extreme Outliers: Z-Score

DCOVA

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

$$Z = \frac{X - \overline{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

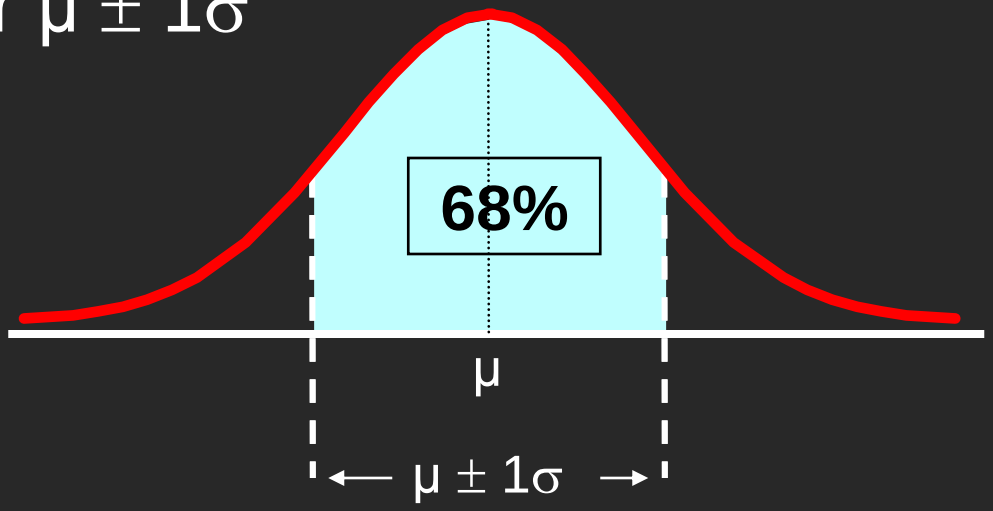
A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

# The Empirical Rule

**(It is studied with Chapter 6)**

DCOVA

- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$

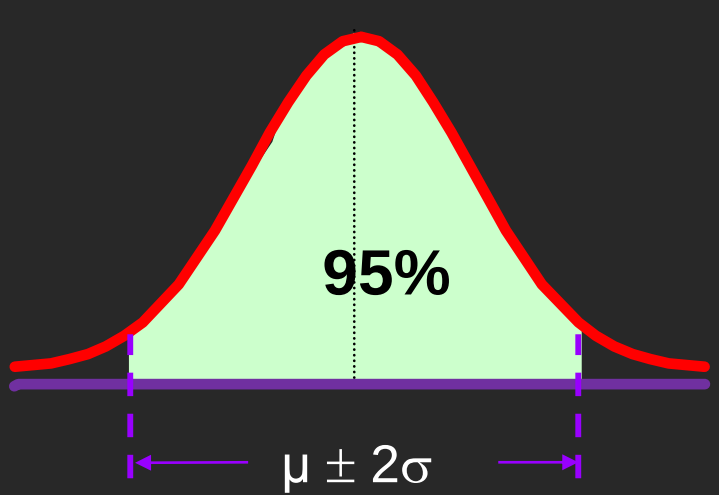


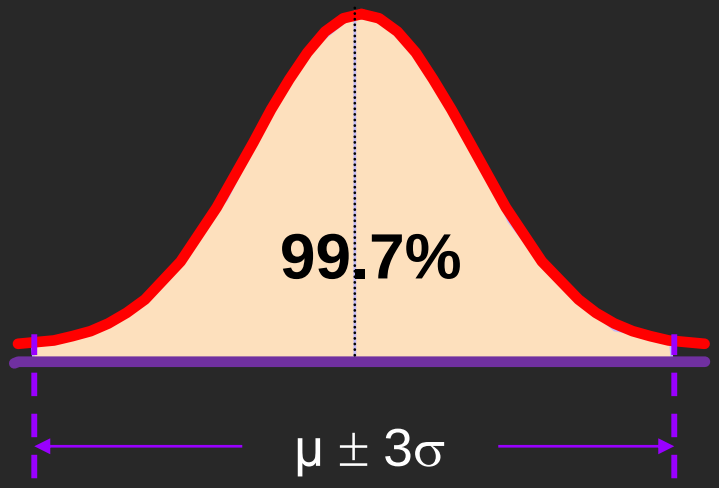
# The Empirical Rule

DCOVA

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or μ ± 2σ

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or μ ± 3σ



# Using the Empirical Rule

DCOVA

- Suppose that the variable Math SAT scores is bell-shaped with a mean of 500 and a standard deviation of 90. Then,
  - Approximately 68% of all test takers scored between 410 and 590, (500 ± 90).
  - Approximately 95% of all test takers scored between 320 and 680, (500 ± 180).
  - Approximately 99.7% of all test takers scored between 230 and 770, (500 ± 270).

# Chapter Summary

**In this chapter we have discussed:**

- Describing the properties of central tendency, variation, and shape in numerical data
- Constructing and interpreting a boxplot
- Computing descriptive summary measures for a population
- Calculating the covariance and the coefficient of correlation