# Unit II

- Statistical Inference of Data Statistical Methods for Evaluation- Hypothesis testing, mean, mode, median, random variables (discrete and continuous),expected value, correlation, variance, standard deviation, limit theorem, difference of means
- (Plz Refer PPT,web resources. Also refer Unit III ppts)

# Measures of Central Tendency

(Reference: 2002 Thomson / South-Western)

- Measures of central tendency yield information about “particular places or locations in a group of numbers.”
- Common Measures of Location
  - Mode
  - Median
  - Mean
  - Percentiles
  - Quartiles

# Mode

- The most frequently occurring value in a data set
- Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)

- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes

# Mode -- Example

- The mode is 44.
- There are more 44s than any other value.

| | | | |
|---|---|---|---|
| 35 | 41 | 44 | 45 |
| 37 | 41 | 44 | 46 |
| 37 | 43 | 44 | 46 |
| 39 | 43 | 44 | 46 |
| 40 | 43 | 44 | 46 |
| 40 | 43 | 45 | 48 |

# Median (ΔΙΑΜΕΣΟΣ)

- Middle value in an ordered array of numbers.
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data
- Unaffected by extremely large and extremely small values.

# Median: Computational Procedure

- First Procedure
  - Arrange observations in an ordered array.
  - If number of terms is odd, the median is the middle term of the ordered array.
  - If number of terms is even, the median is the average of the middle two terms.
- Second Procedure
  - The median's position in an ordered array is given by (n+1)/2.

# Median: Example with an Odd Number of Terms

Ordered Array includes:
3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

- There are 17 terms in the ordered array.
- Position of median = (n+1)/2 = (17+1)/2 = 9
- The median is the 9th term, 15.
- If the 22 is replaced by 100, the median remains at 15.
- If the 3 is replaced by -103, the median remains at 15.

# Mean (ΜΕΣΟΣ)

- Is the average of a group of numbers
- Applicable for interval and ratio data, not applicable for nominal or ordinal data
- Affected by each value in the data set, including extreme values
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set

# Population Mean

$$\mu = \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

$$= \frac{24 + 13 + 19 + 26 + 11}{5}$$

$$= \frac{93}{5}$$

$$= 186$$

# Sample Mean

$$\bar{X} = \frac{\sum X}{n} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$$

$$= \frac{57 + 86 + 42 + 38 + 90 + 66}{6}$$

$$= \frac{379}{6}$$

$$= 63167$$

# Quartiles

Measures of central tendency that divide a group of data into four subgroups

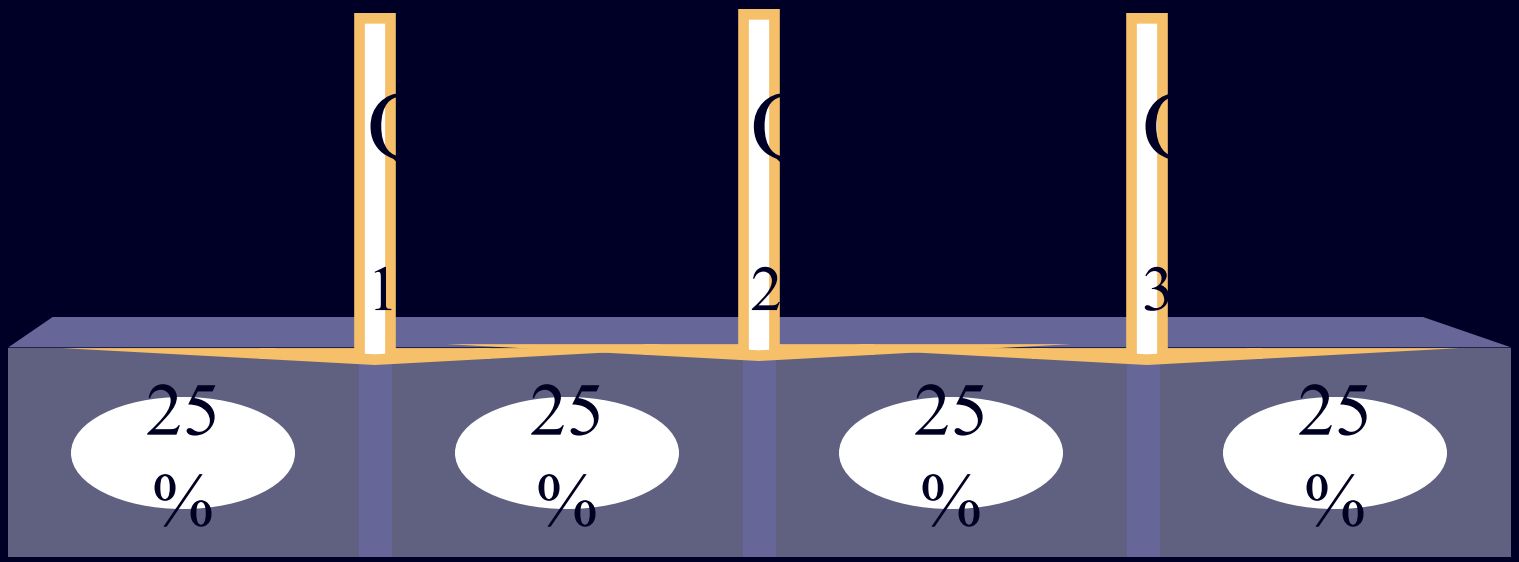
- $Q_1$: 25% of the data set is below the first quartile
- $Q_2$: 50% of the data set is below the second quartile
- $Q_3$: 75% of the data set is below the third quartile

# Quartiles, *continued*

- $Q_1$ is equal to the 25th percentile
- $Q_2$ is located at 50th percentile and equals the median
- $Q_3$ is equal to the 75th percentile

Quartile values are not necessarily members of the data set

# Quartiles



# Quartiles: Example

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129

- $Q_1$:

$$i = \frac{25}{100}(8) = 2 \qquad Q_1 = \frac{109 + 114}{2} = 111.5$$

- $Q_2$:

$$i = \frac{50}{100}(8) = 4 \qquad Q_2 = \frac{116 + 121}{2} = 118.5$$

- $Q_3$:

$$i = \frac{75}{100}(8) = 6 \qquad Q_3 = \frac{122 + 125}{2} = 123.5$$

# Measures of Variability

- Measures of variability describe the spread or the dispersion of a set of data.
- Common Measures of Variability
  - Range
  - Interquartile Range
  - Mean Absolute Deviation
  - Variance
  - Standard Deviation
  - Z scores
  - Coefficient of Variation

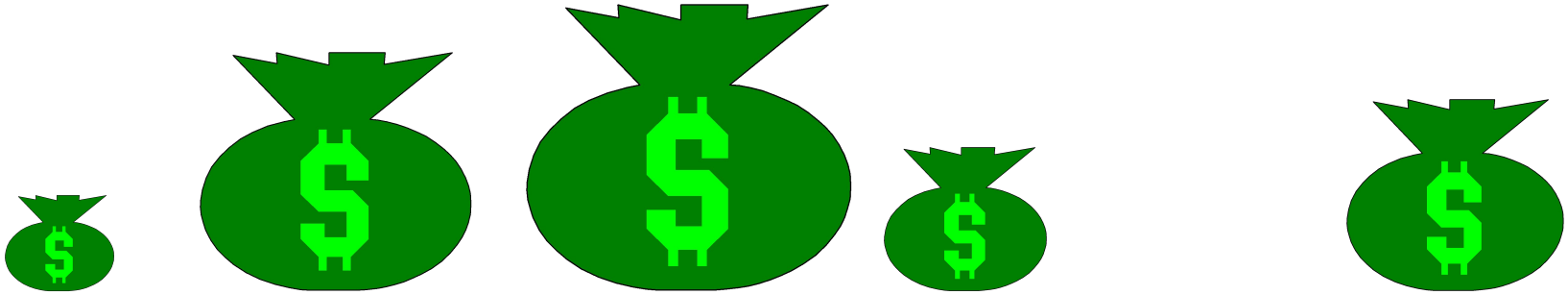
# Variability

**No Variability in Cash Flow**

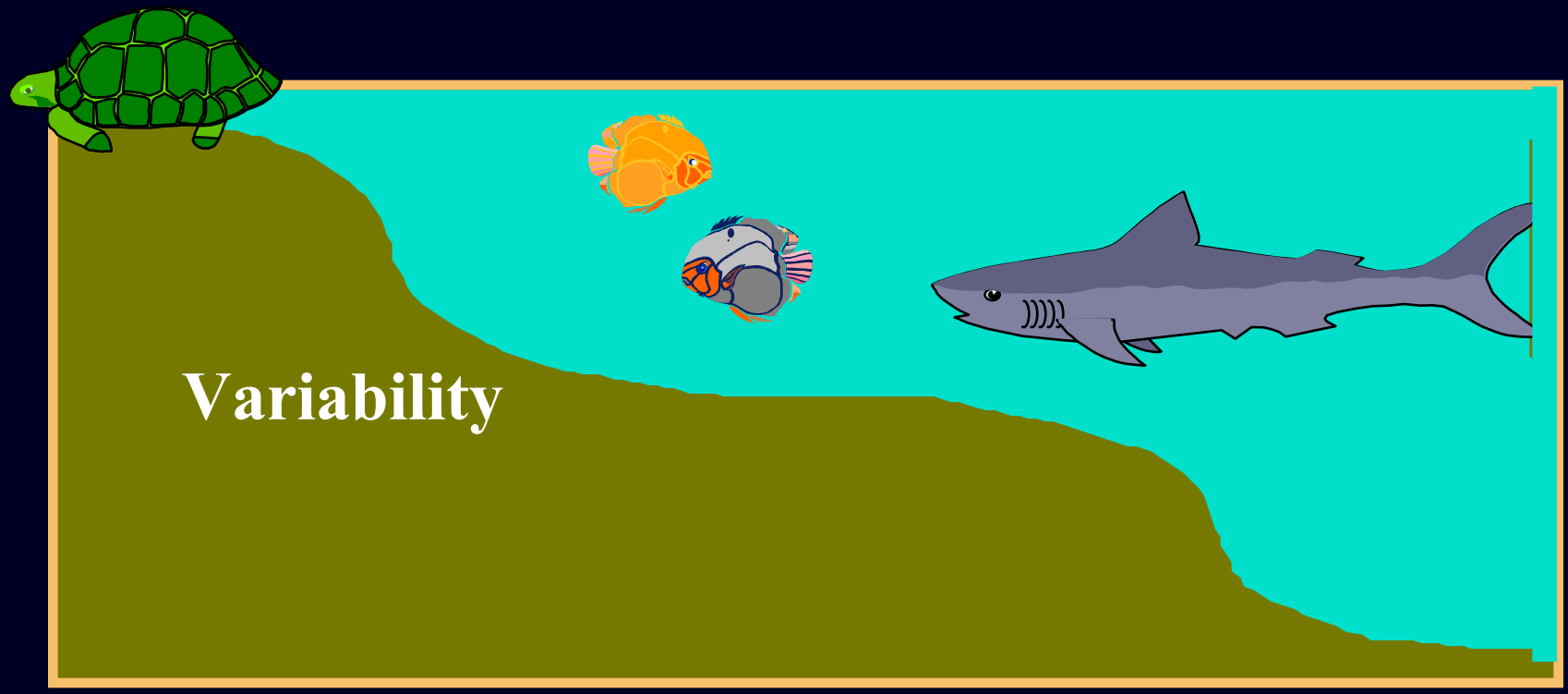
Mean

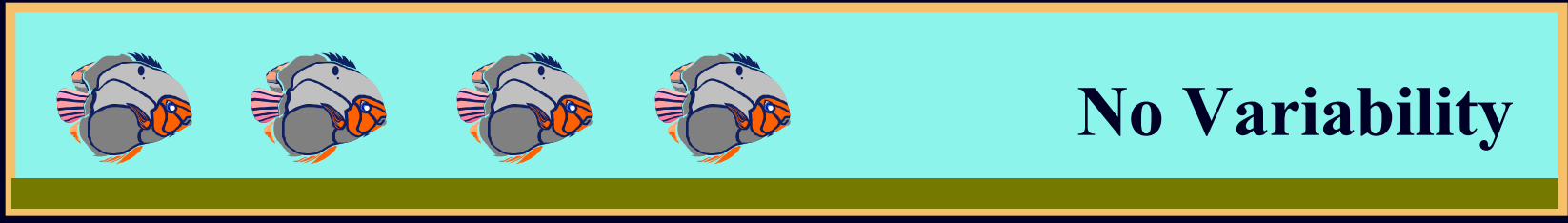
**Variability in Cash Flow**

Mean



# Variability



# Range

- The difference between the largest and the smallest values in a set of data
- Simple to compute
- Ignores all data points except the two extremes
- Example: Range = Larg = 48 - 35 = 13

| | | | |
|---|---|---|---|
| 35 | 41 | 44 | 45 |
| 37 | 41 | 44 | 46 |
| 37 | 43 | 44 | 46 |
| 39 | 43 | 44 | 46 |
| 40 | 43 | 44 | 46 |
| 40 | 43 | 45 | 48 |

# Interquartile Range

- Range of values between the first and third quartiles
- Range of the “middle half”
- Less influenced by extremes

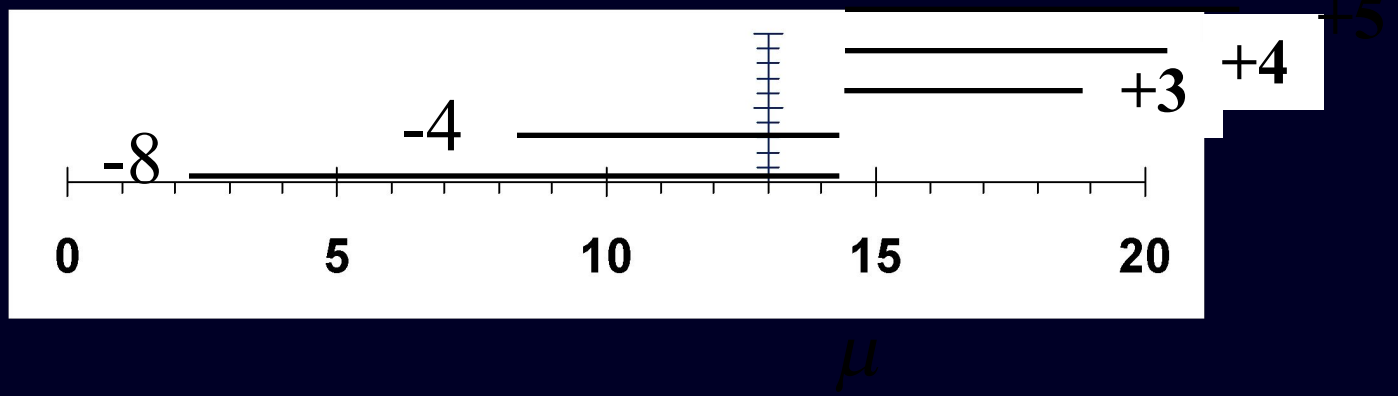
$$\text{Interquartile Range} = Q_3 - Q_1$$

# Deviation from the Mean

- Data set: 5, 9, 16, 17, 18
- Mean:

$$\mu = \frac{\sum X}{N} = \frac{65}{5} = 13$$

- Deviations from the mean: -8, -4, 3, 4, 5



# Mean Absolute Deviation

- Average of the absolute deviations from the mean

| $X$ | $X-\mu$ | $\lvert X-\mu\rvert$ |
|---|---|---|
| 5 | -8 | +8 |
| 9 | -4 | +4 |
| 16 | +3 | +3 |
| 17 | +4 | +4 |
| 18 | +5 | +5 |
| | 0 | 24 |

$$M.A.D. = \frac{\sum |X-\mu|}{N}$$

$$= \frac{24}{5}$$

$$= 4.8$$

# Population Variance

- Average of the squared deviations from the arithmetic mean

| $X$ | $X-\mu$ | $(X-\mu)^2$ |
|---|---|---|
| 5 | -8 | 64 |
| 9 | -4 | 16 |
| 16 | +3 | 9 |
| 17 | +4 | 16 |
| 18 | +5 | 25 |
| | 0 | 130 |

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

# Population Standard Deviation

- Square root of the variance

| $X$ | $X-\mu$ | $(X-\mu)^2$ |
|---|---|---|
| 5 | -8 | 64 |
| 9 | -4 | 16 |
| 16 | +3 | 9 |
| 17 | +4 | 16 |
| 18 | +5 | 25 |
| | 0 | 130 |

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{26.0}$$

$$= 5.1$$

# Empirical Rule

- Data are normally distributed (or approximately normal)

| Distance from the Mean | Percentage of Values Falling Within Distance |
|---|---|
| $\mu \pm 1\sigma$ | 68 |
| $\mu \pm 2\sigma$ | 95 |
| $\mu \pm 3\sigma$ | 99.7 |

# Sample Variance

- Average of the squared deviations from the arithmetic mean

| $X$ | $X-\bar{X}$ | $(X-\bar{X})^2$ |
|---|---|---|
| 2,398 | 625 | 390,625 |
| 1,844 | 71 | 5,041 |
| 1,539 | -234 | 54,756 |
| 1,311 | -462 | 213,444 |
| 7,092 | 0 | 663,866 |

$$S^2 = \frac{\sum (X-\bar{X})^2}{n-1} = \frac{663,866}{3} = 221,288.67$$

# Sample Standard Deviation

- Square root of the sample variance

| $X$ | $X-\overline{X}$ | $(X-\overline{X})^2$ |
|---|---|---|
| 2,398 | 625 | 390,625 |
| 1,844 | 71 | 5,041 |
| 1,539 | -234 | 54,756 |
| 1,311 | -462 | 213,444 |
| 7,092 | 0 | 663,866 |

$$S^2 = \frac{\sum (X-\overline{X})^2}{n-1}$$

$$= \frac{663,866}{3}$$

$$= 221,288.67$$

$$S = \sqrt{S^2}$$

$$= \sqrt{221,288.67}$$

$$= 470.41$$

# Coefficient of Variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

$$C.V. = \frac{\sigma}{\mu}(100)$$

# Coefficient of Variation

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

$$C.V._1 = \frac{\sigma_1}{\mu_1}(100)$$

$$= \frac{4.6}{29}(100)$$

$$= 15.86$$

$$\mu_2 = 84$$

$$\sigma_2 = 10$$

$$C.V._2 = \frac{\sigma_2}{\mu_2}(100)$$

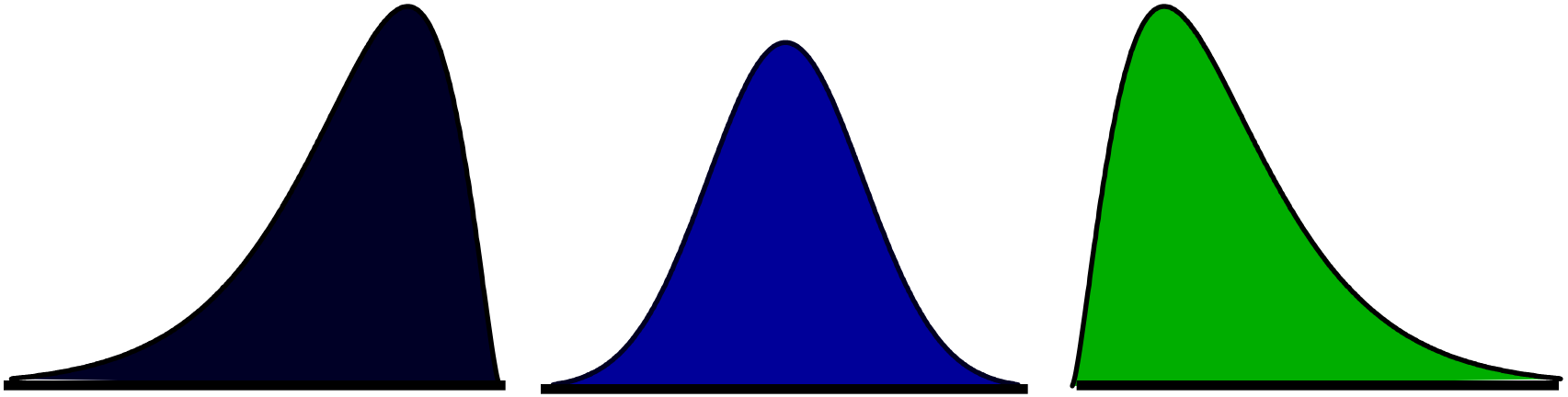
$$= \frac{10}{84}(100)$$

$$= 11.90$$

# Measures of Shape

- **Skewness**
  - Absence of symmetry
  - Extreme values in one side of a distribution
- **Kurtosis**
  - Peakedness of a distribution
- **Box and Whisker Plots**
  - Graphic display of a distribution
  - Reveals skewness

# Skewness

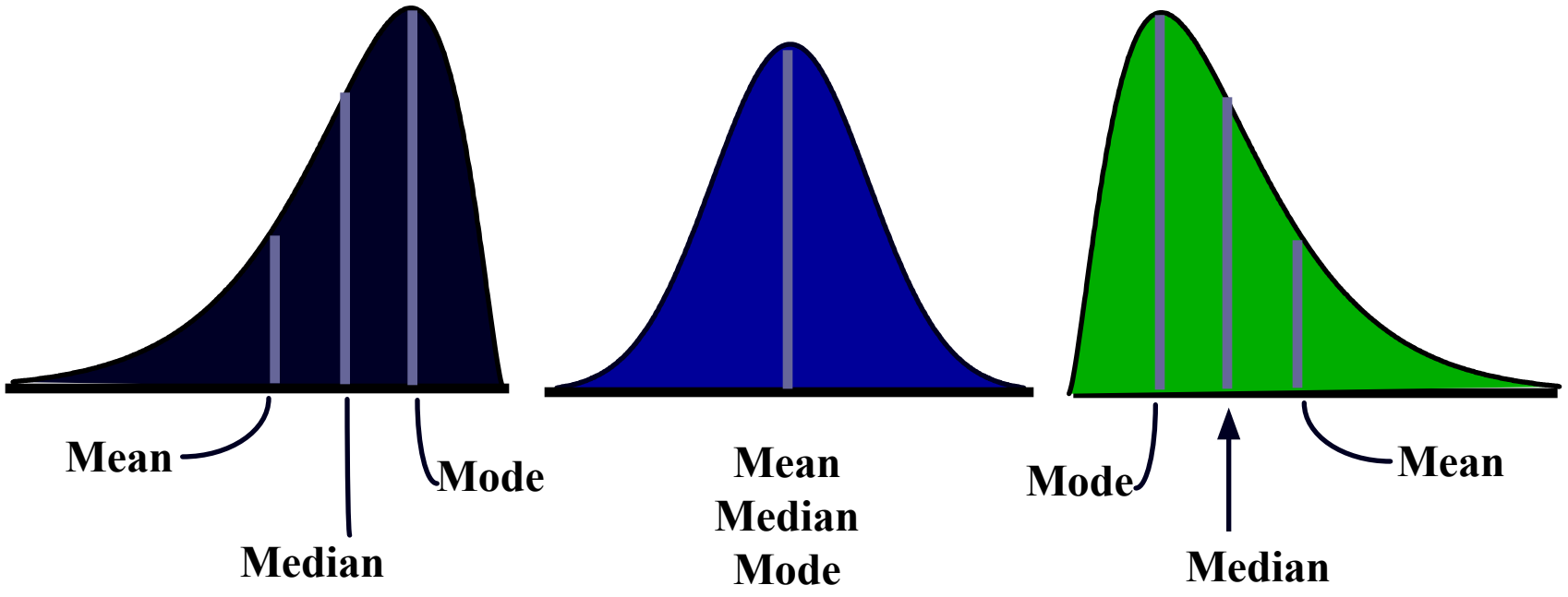


**Negatively Skewed**

**Symmetric (Not Skewed)**

**Positively Skewed**

# Skewness




**Negatively Skewed**

**Symmetric (Not Skewed)**

**Positively Skewed**

# Coefficient of Skewness

- Summary measure for skewness

$$S = \frac{3(\mu - M_d)}{\sigma}$$

- If S < 0, the distribution is negatively skewed (skewed to the left).
- If S = 0, the distribution is symmetric (not skewed).
- If S > 0, the distribution is positively skewed (skewed to the right).

# Coefficient of Skewness

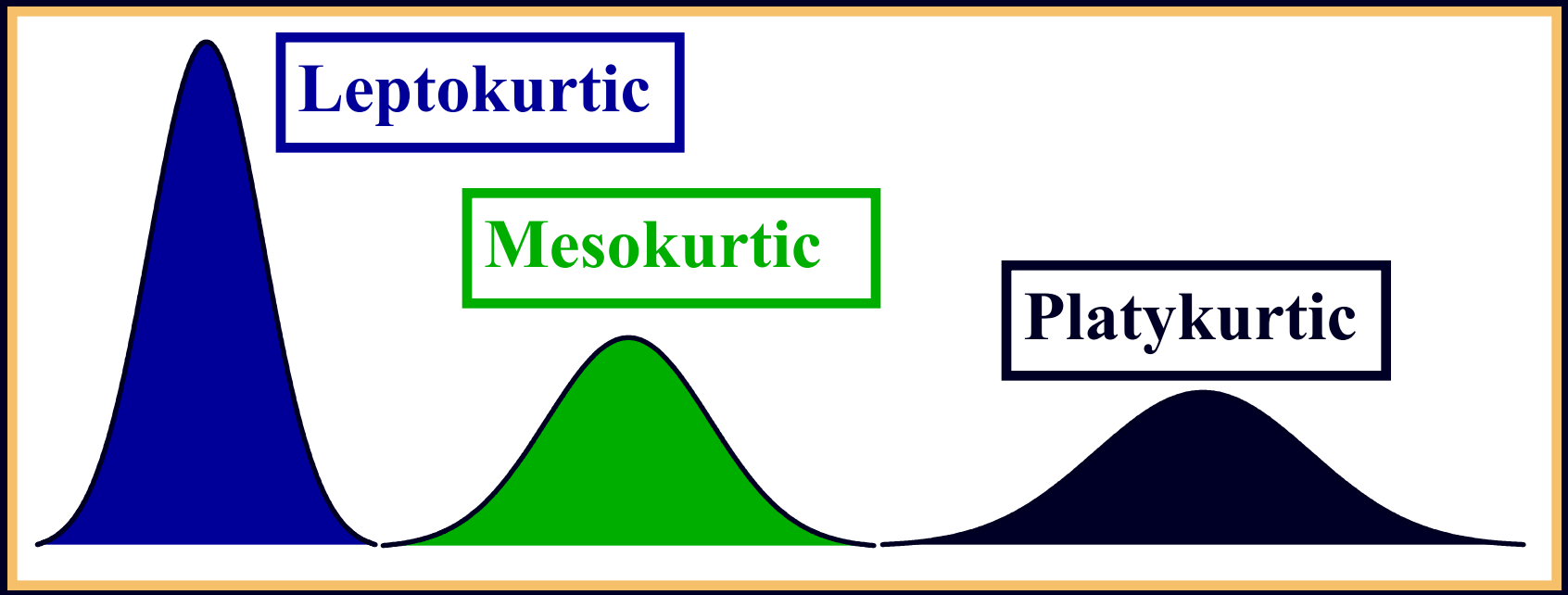
$$\mu_1 = 23$$
$$M_{d1} = 26$$
$$\sigma_1 = 12.3$$
$$S_1 = \frac{3(\mu_1 - M_{d1})}{\sigma_1} = \frac{3(23-26)}{12.3} = -0.73$$

$$\mu_2 = 26$$
$$M_{d2} = 26$$
$$\sigma_2 = 12.3$$
$$S_2 = \frac{3(\mu_2 - M_{d2})}{\sigma_2} = \frac{3(26-26)}{12.3} = 0$$

$$\mu_3 = 29$$
$$M_{d3} = 26$$
$$\sigma_3 = 12.3$$
$$S_3 = \frac{3(\mu_3 - M_{d3})}{\sigma_3} = \frac{3(29-26)}{12.3} = +0.73$$

# Kurtosis

- Peakedness of a distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal in shape
  - Platykurtic: flat and spread out



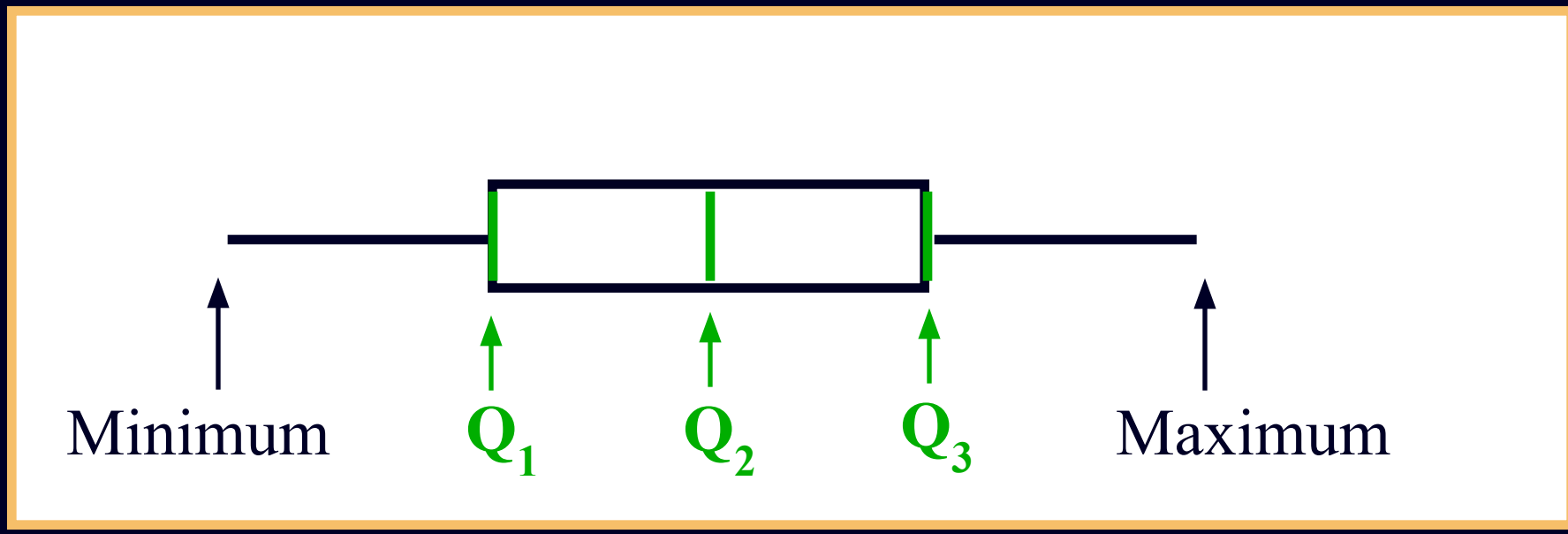
# Box and Whisker Plot

- **Five specific values are used:**
  - **Median, $Q_2$**
  - **First quartile, $Q_1$**
  - **Third quartile, $Q_3$**
  - **Minimum value in the data set**
  - **Maximum value in the data set**

# Box and Whisker Plot, *continued*

- **Inner Fences**
  - **IQR = $Q_3$ - $Q_1$**
  - **Lower inner fence = $Q_1$ - 1.5 IQR**
  - **Upper inner fence = $Q_3$ + 1.5 IQR**
- **Outer Fences**
  - **Lower outer fence = $Q_1$ - 3.0 IQR**
  - **Upper outer fence = $Q_3$ + 3.0 IQR**

# Box and Whisker Plot



# Skewness: Box and Whisker Plots, and Coefficient of Skewness