Fundamentals of Statistics and Probability

Learning Objective

Statistical analysis

Importance of probability theory in the design and development of machine learning algorithms.

Introduction to Statistics

At the end of the module, you will be able,

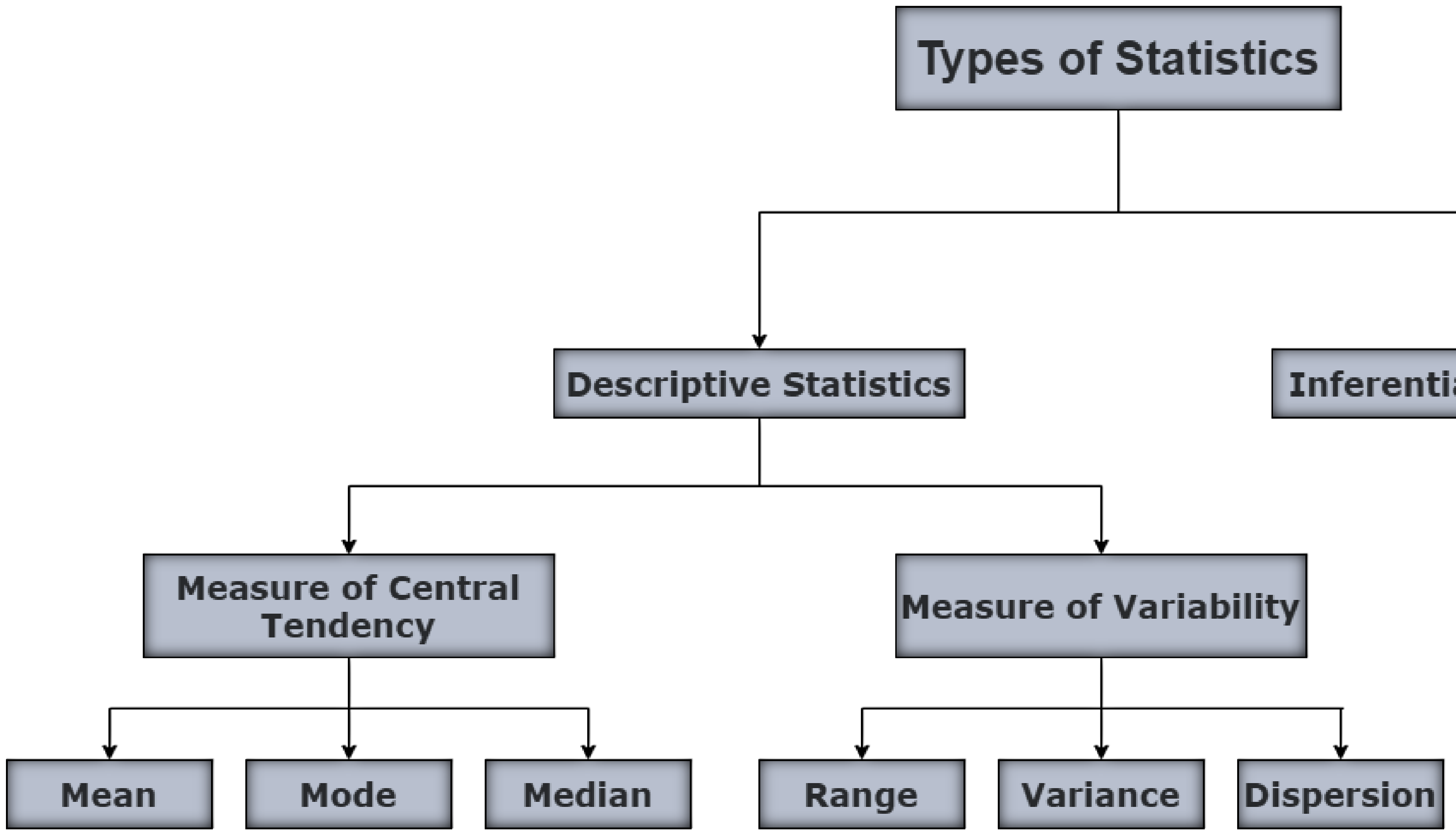
- Discuss the term Statistics and its types
- Demonstrate Descriptive statistics and measures used in it.
- Demonstrate Inferential statistics and measures used in it

Introduction to Statistics

Data and Statistics

- Statistics is the science of learning from Data
- Data is essentially numbers (or text/symbols) which represent some information.
- It helps to think of data as 'values' of quantitative and qualitative variables
- What are the variable types:
  - Numerical or Quantitative: (Continuous and Discrete)
  - Categorical or Qualitative: (Always discrete)
    - Nominal
    - Ordinal

Types of Statistics



Descriptive Statistics

- Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table
- There are two categories
  - Measure of central tendency
  - Measure of Variability/Dispersion

Measures of Central Tendency

- Measures of central tendency yield information about "particular places or locations in a group of numbers."
- A single number to describe the characteristics of a set of data

Summary statistics

Central tendency or measures of location

- Arithmetic mean
- Weighted mean
- Median
- Mode
- Percentile

Dispersion

- Skewness
- Kurtosis
- Range
- Interquartile range
- Variance
- Standard score
- Coefficient of variation

Arithmetic Mean

- Commonly called 'the mean'
- It 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

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

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

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

$$= 18.6$$

Sample Mean

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

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

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

$$= 63.167$$

Mean of Grouped Data

- Weighted average Of class midpoints
- Class frequencies are the weights

10 - 20 5

$$\mu = \frac{\sum fM}{\sum f} = \frac{\sum fM}{N} = \frac{f_1M_1 + f_2M_2 + f_3M_3 + \cdots + f_iM_i}{f_1 + f_2 + f_3 + \cdots + f_i}$$

Calculation of Grouped Data Mean

Class Interval	Frequency(f)	Class Midpoint(M)	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
	50		2150

$$\mu = \frac{\sum fM}{\sum f} = \frac{2150}{50} = 43.0$$

weighted average

- Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers.
- The average you need is the weighted average.

Formula :

$$Weighted\ \ Average = \frac{\sum xw}{\sum w}$$

- where x is a data value and w is the weight assigned to that data value. The sum is taken over all data values.

weighted average- Example

Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average Of your scores. If the minimum average for an A is 90, will you earn an A?

weighted average- Example

Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average Of your scores. If the minimum average for an A is 90, will you earn an A?

$$Weighted\ Average = \frac{(83)(0.40)+(95)(0.60)}{0.40+0.60}$$

$$= \frac{32+57}{1} = 90.2$$

You will earn an A!

Median

- 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 the observations in an ordered array
  - If there is an odd number of terms, the median is the middle term of the ordered array
  - If there is an even number of terms, 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

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 is 15.
- If the 3 is replaced by −250, the median is 15.

Median: Example with an Even Number of Terms

Ordered Array

3,4,5,7,8,9,11 ,14, 15, 16, 16 ,17, 19, 19, 20, 21

- There are 16 terms in the ordered array
- Position Of median = (n+1)/2 = (16+1)/2 = 8.5
- The median is between the 8th and 9th terms, 14.5
- If the 21 is replaced by 100, the median is 14.5
- If the 3 is replaced by -88, the median is 14.5

Median of Grouped data

$$Median = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

Where :

L = the lower limit of the median class

cf_p = cumulative frequency of class preceding the median class

f_{med} = frequency of the median class

W = width of the median class

N = total of frequencies

Median of Grouped data-Example

Class Interval	Frequency	Cumulative Frequency
20-under 30	6	6
30-under 40	18	24
40-under 50	11	35
50-under 60	11	46
60-under 70	3	49
70-under 80	1	50
	N = 50	

$$Md = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

$$= 40 + \frac{\frac{50}{2} - 24}{11}(10)$$

$$= 40.909$$

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

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

Mode of Grouped Data

- Midpoint Of the modal class
- Modal class has the greatest frequency

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

$$Mode = L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right) w =$$

$$30 + \left(\frac{12}{12 + 7} \right) 10 = 36.31$$

Percentiles

- Measures of central tendency that divide a group of data into 100 parts
- Example: 90th percentile indicates that at most 90% Of the data lie below it, and at least 10% Of the data lie above it
- The median and the 50th percentile have the same value
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data

Percentiles: Computational Procedure

- Organize the data into an ascending ordered array
- Calculate the p th percentile location: $i = \frac{P}{100}(n)$
- Determine the percentile's location and its value.
- If i is a whole number, the percentile is the average Of the values at the i and (i+l) positions
- If i is not a whole number, the percentile is at the (i+l) position in the ordered array

Percentiles:Example

- **Percentiles:**

  Raw Data: 14, 12, 19, 23, 5, 13, 28, 17

  Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28

- **Location of 30th percentile:**

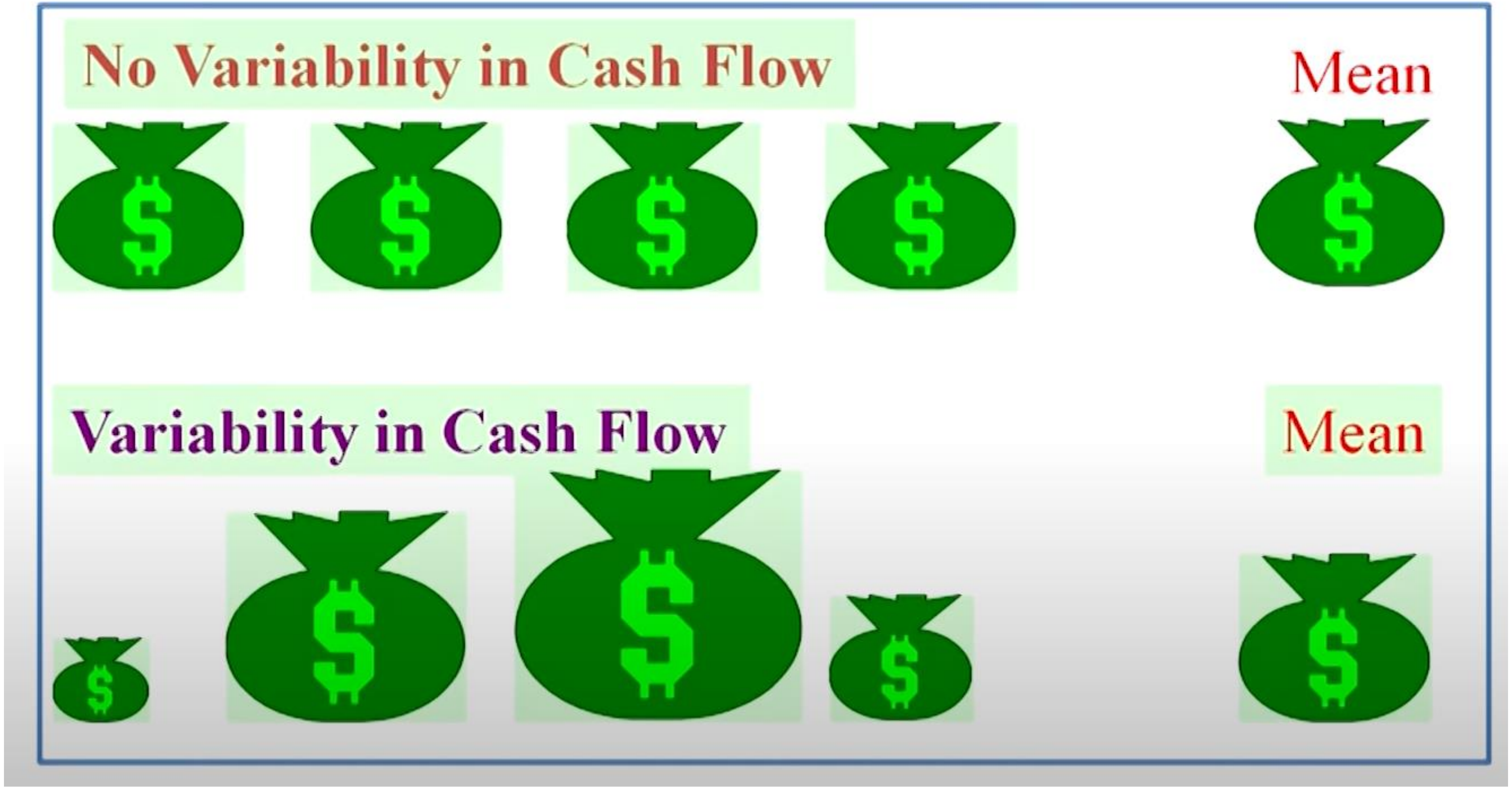
$$i = \frac{30}{100}(8) = 2.4$$

- **The location index, i, is not a whole number; i+1 = 2.4+1=3.4;**
- **the whole number portion is 3; the 30th percentile is at the 3rd**
- **location of the array; the 30th percentile is 13.**

Dispersion

- Measures Of variability describe the spread or the dispersion of a set of data
- Reliability of measure of central tendency
- To compare dispersion Of various samples

Variability



Measures of Variability or dispersion

Common Measures of Variability

- Range
- Inter-quartile range
- Mean Absolute Deviation
- Variance
- Standard Deviation
- Z scores
- Coefficient of Variation

Range — Ungrouped data

- 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:

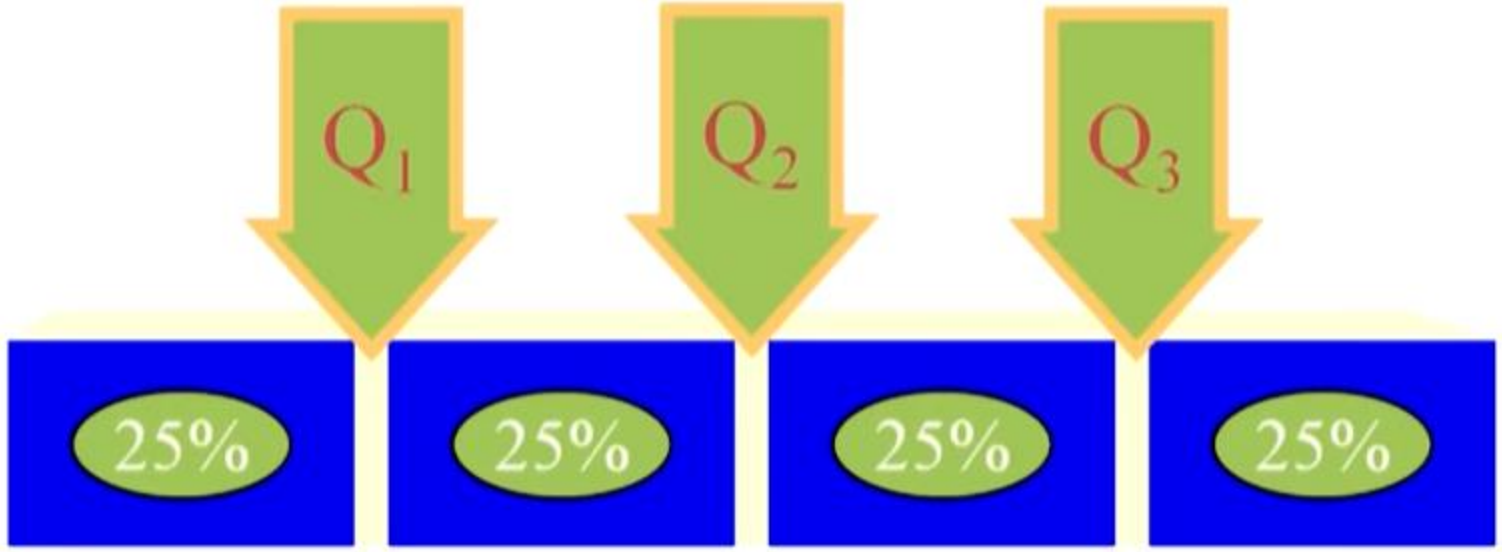
Range = Largest — Smallest = 48 - 35 = 13

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

Quartiles

- Measures of central tendency that divide a group of data into four subgroups
- QI: 25% of the data set is below the first quartile
- Q2: 50% of the data set is below the second quartile
- Q3: 75% of the data set is below the third quartile
- QI is equal to the 25th percentile
- Q2 is located at 50th percentile and equals the median
- Q3 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:

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

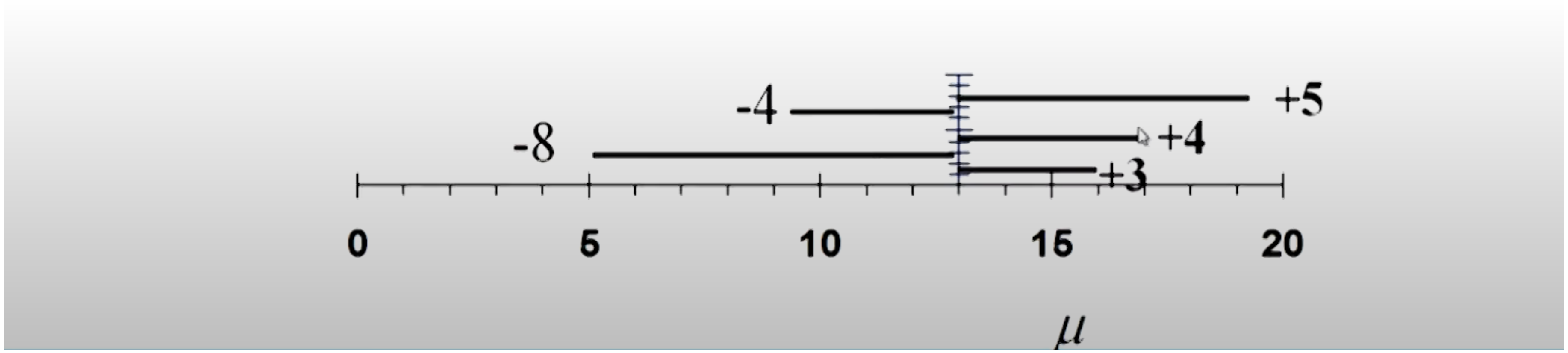
Interquartile Range

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

Interquartile Range = Q3-Q1

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 **absolute** deviation 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

$$M.A.D. = \frac{\sum \lvert X - \mu \rvert}{N} = \frac{24}{5} = 4.8$$

Population Variance

- Average of Squared deviation 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

- Squared 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

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} = \frac{130}{5} = 26.0$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{26.0} = 5.1$$

Population Standard Deviation

- Squared 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

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

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

$$= 26.0$$

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

$$= \sqrt{26.0}$$

$$= 5.1$$

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

$$S^2 = \frac{\sum (X-\bar{X})^2}{n-1}$$
$$= \frac{663,866}{3}$$
$$= 221,288.67$$
$$S = \sqrt{S^2}$$
$$= \sqrt{221,288.67}$$
$$= 470.41$$

Measures of Dispersion

Data set: 3,4,3,1,2,3,9,5,6,7,4,8

- Range (Max-Min) (9-1 = 8)
- Inter Quartile Range: 3rd quartile -1st quartile (75th Percentile-25thPercentile) (6.5 - 3 = 3.5)
- Sample Standard deviation

$$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{1}{12-1}\sum((3-4.58)^2 + (4-4.58)^2 \ldots)$$

Measures of Dispersion

Questions that go with Standard deviation

- **Why do we use the square function on the deviations? What are its implications?**
- **Why do we work on standard deviation and not the variance?**
- **Why do we average by dividing by N-1 and not N?**

Uses of Standard Deviation

- Indicator of financial risk
- Quality Control
  - construction of quality control charts
  - process capability studies
- Comparing populations
  - household incomes in two cities
  - employee absenteeism at two plants

Standard Deviation as an Indicator of financial risk

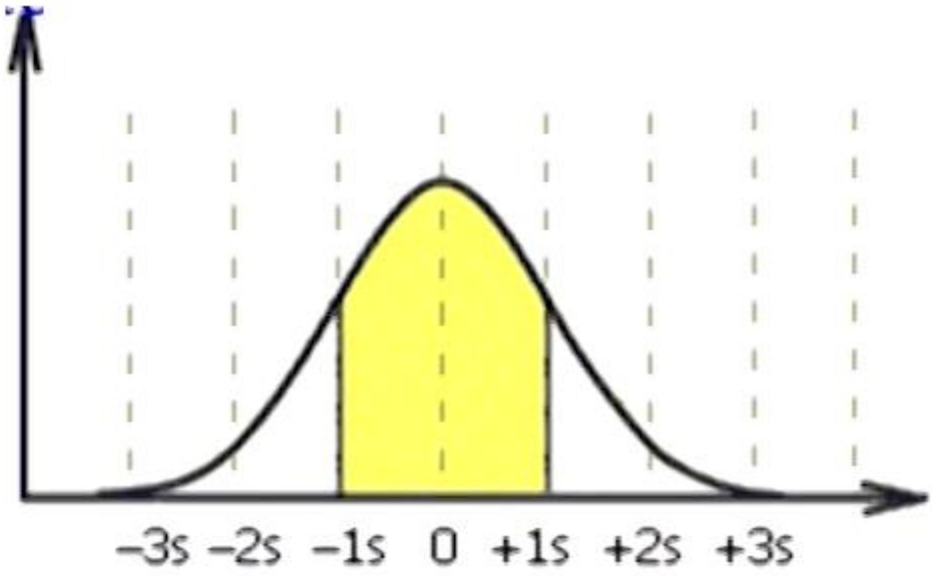
Financial Security	Annualized Rate of Return μ	σ
A	15%	3%
B	15%	7%

Central Tendency and Dispersion

- **important property of a normal distribution**
- **Various kurtosis**
- **box and whisker plots**

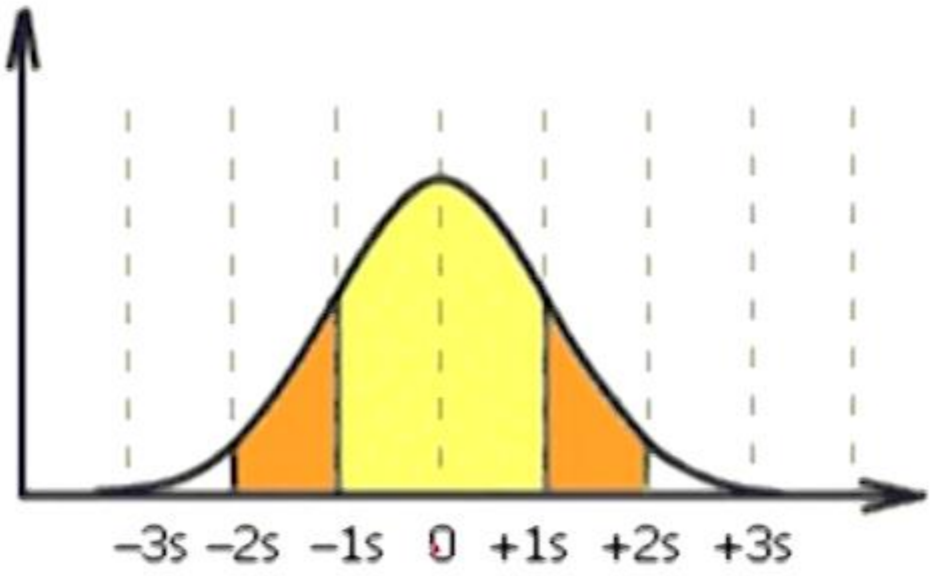
The Empirical Rule (Normal Distribution)

- **Approximately 68% of all Observations fall within one standard deviation Of the mean.**



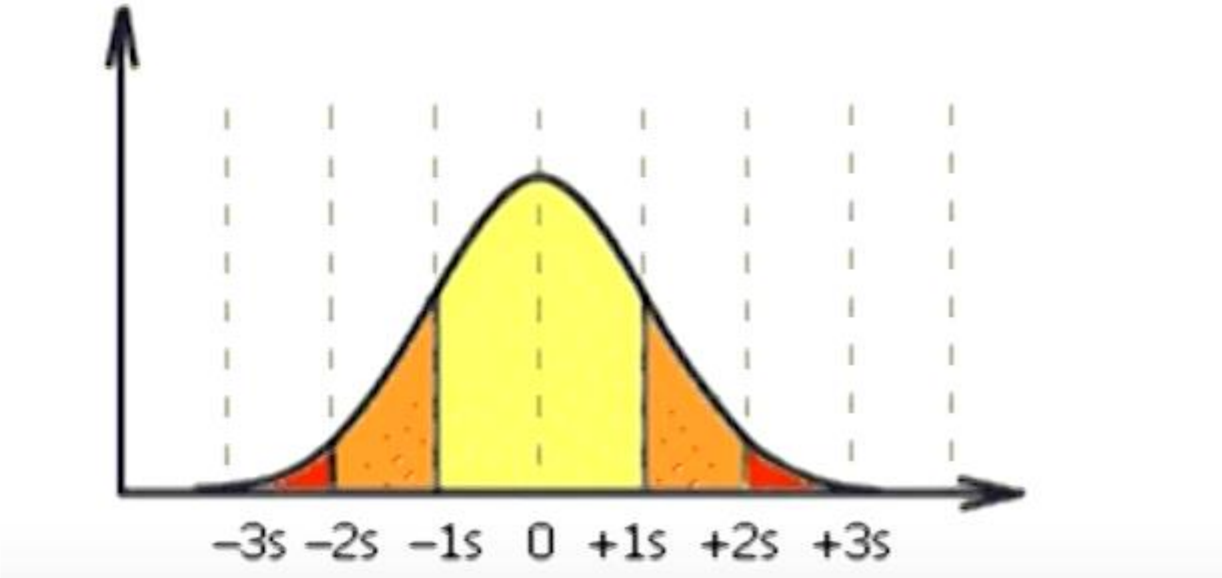
The Empirical Rule (Normal Distribution)

- **Approximately 95% Of all observations fall within two standard deviations of the mean.**

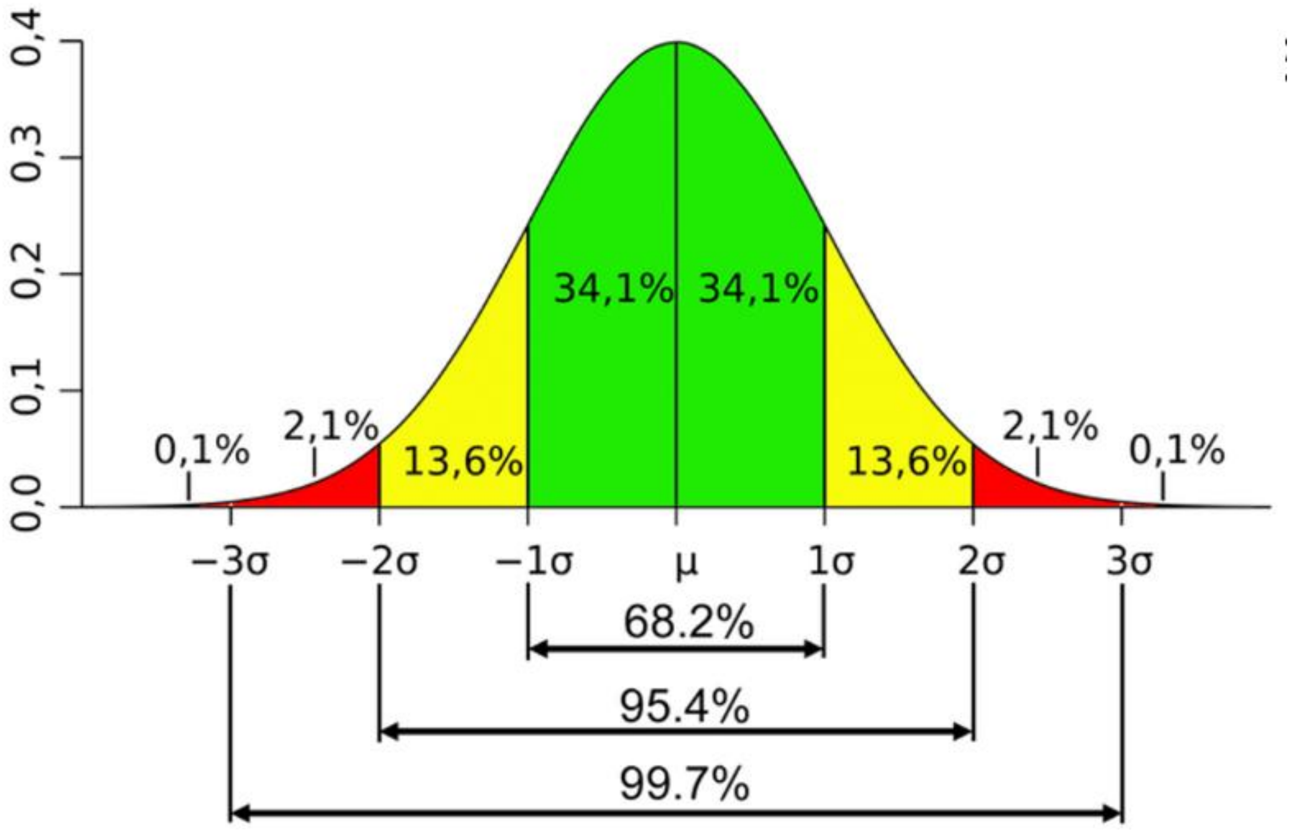


The Empirical Rule (Normal Distribution)

- Approximately 99.7% of all observations fall within three standard deviations of the mean.



The Empirical Rule (Normal Distribution)



The Empirical Rule (Normal Distribution)

- **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

Chebysheff's Theorem (Not often used because interval is very wide.)

- A more general interpretation of the standard deviation is derived from Chebysheff's Theorem, which applies to all shapes of histograms (not just bell shaped).
- The proportion of observations in any sample that lie within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2} \quad for \quad k > 1$$

For k=2 (say), the theorem states that at least 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

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

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

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

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

$$= 15.86$$

$$\mu_2 = 84$$

$$\sigma_2 = 10$$

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

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

$$= 11.90$$

Variance and Standard Deviation of Grouped Data

Population

$$\sigma^2 = \frac{\sum f\left(M-\mu\right)^2}{N}$$

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

Sample

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

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

Population Variance and Standard Deviation of Grouped Data(mu=43)

Class Interval	f	M	fM	$M-\mu$	$(M-\mu)^2$	$f(M-\mu)^2$
20-under 30	6	25	150	-18	324	1944
30-under 40	18	35	630	-8	64	1152
40-under 50	11	45	495	2	4	44
50-under 60	11	55	605	12	144	1584
60-under 70	3	65	195	22	484	1452
70-under 80	1	75	75	32	1024	1024
	50		2150			7200

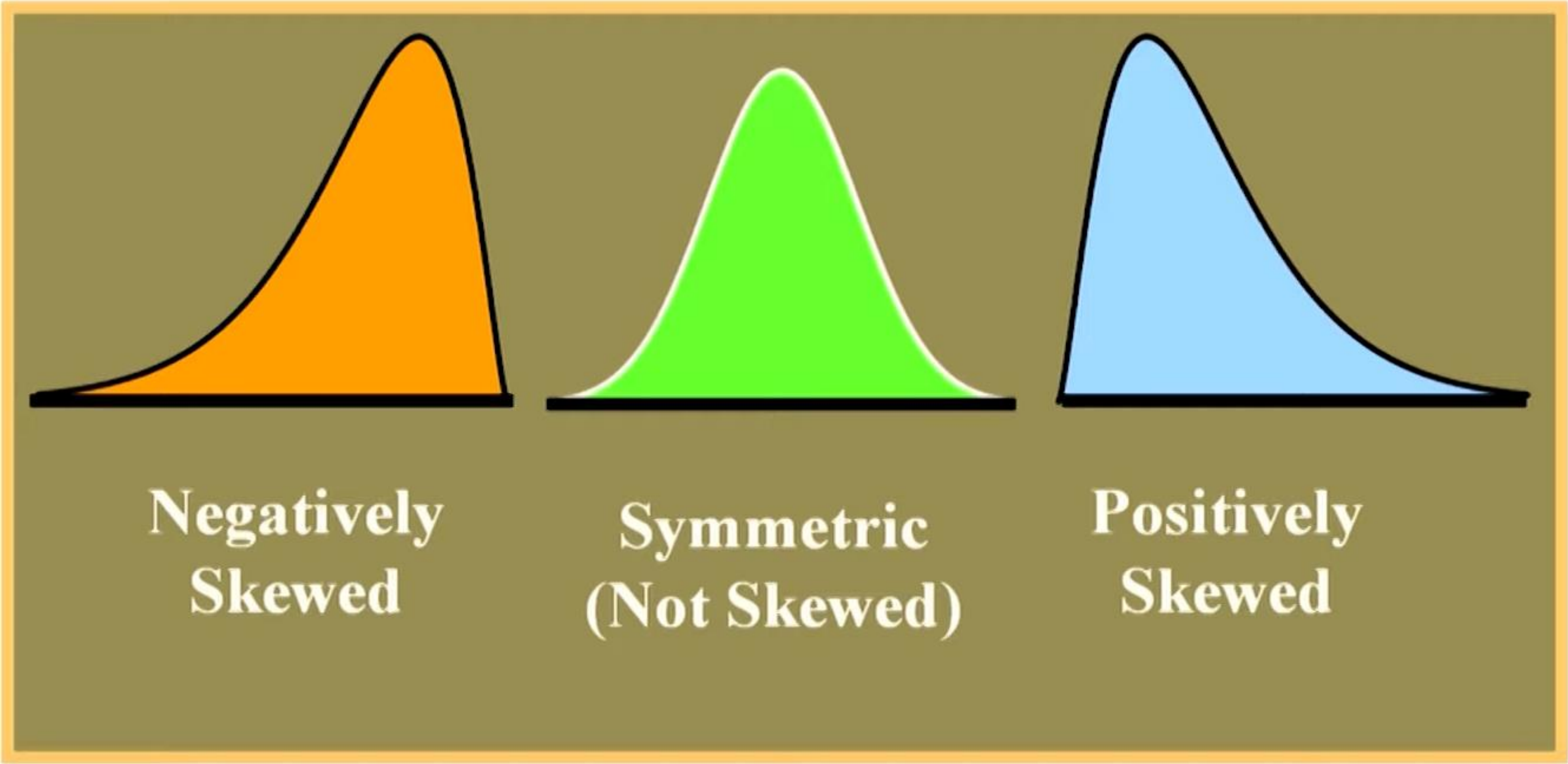
$$\sigma^2 = \frac{\sum f(M-\mu)^2}{N} = \frac{7200}{50} = 144$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$$

Measures of Shape

- **Skewness**
  - Absence of symmetry
  - Extreme values in one side of a distribution
- **Kurtosis - Peakedness of a distribution**
  - Leptokurtic: high and thin
  - Mesokurtic: normal shape
  - Platykurtic: flat and spread out
- **Box and Whisker Plots**
  - Graphic display of a distribution
  - Reveals skewness

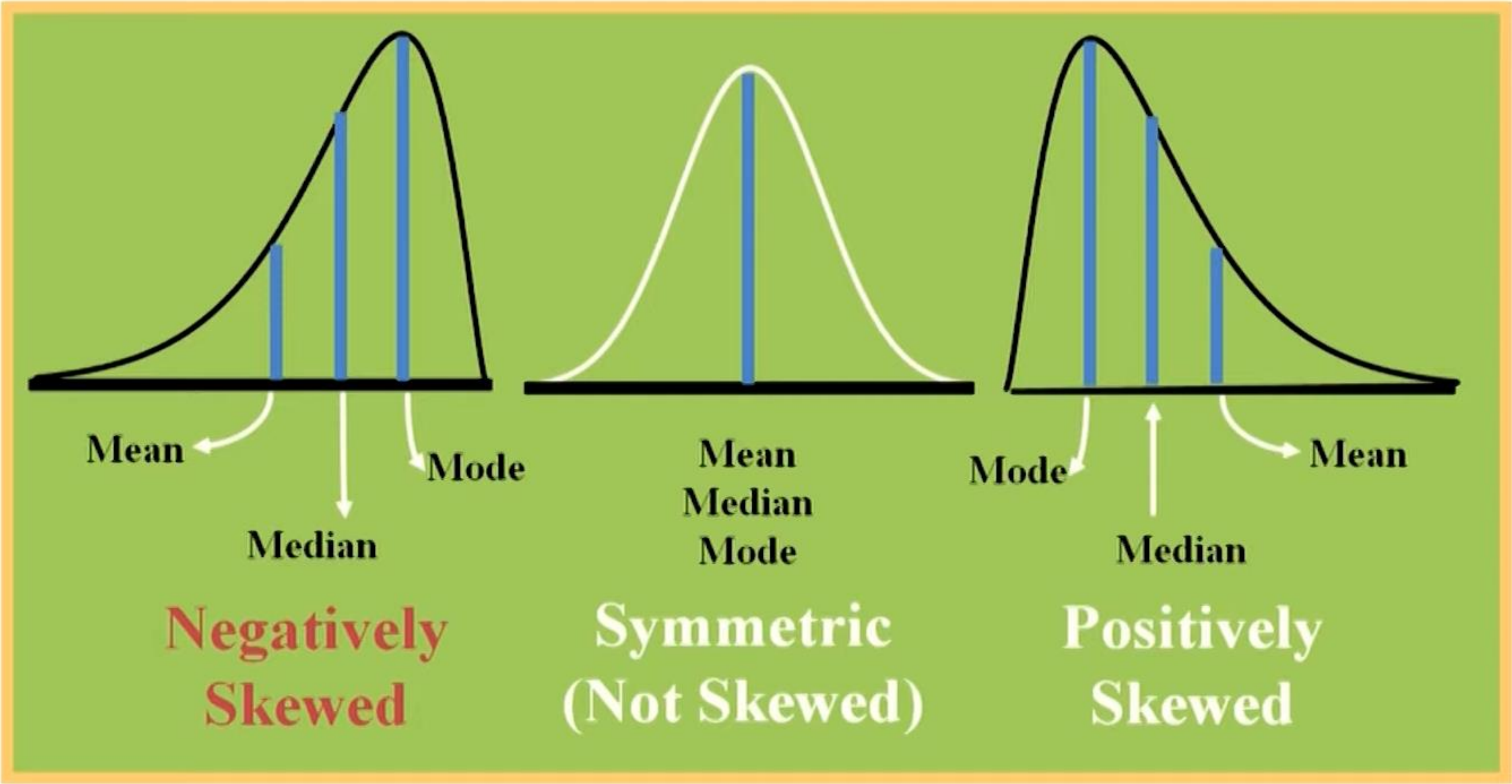
skewness


Negatively
Skewed
Symmetric
(Not Skewed)
Positively
Skewed

skewness

- The skewness of a distribution is measured by comparing the relative positions of the mean, median and mode.
- Distribution is symmetrical
  - Mean = Median = Mode
- Distribution skewed right
  - Median lies between mode and mean, and mode is less than mean
- Distribution skewed left
  - Median lies between mode and mean, and mode is greater than mean

skewness


Mean
Median
Mode
Mean
Median
Mode
Mode
Median
Mean
Negatively Skewed
Symmetric (Not Skewed)
Positively Skewed

Coefficient of Skewness

➢ **Summary measure for skewness**

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

- 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)

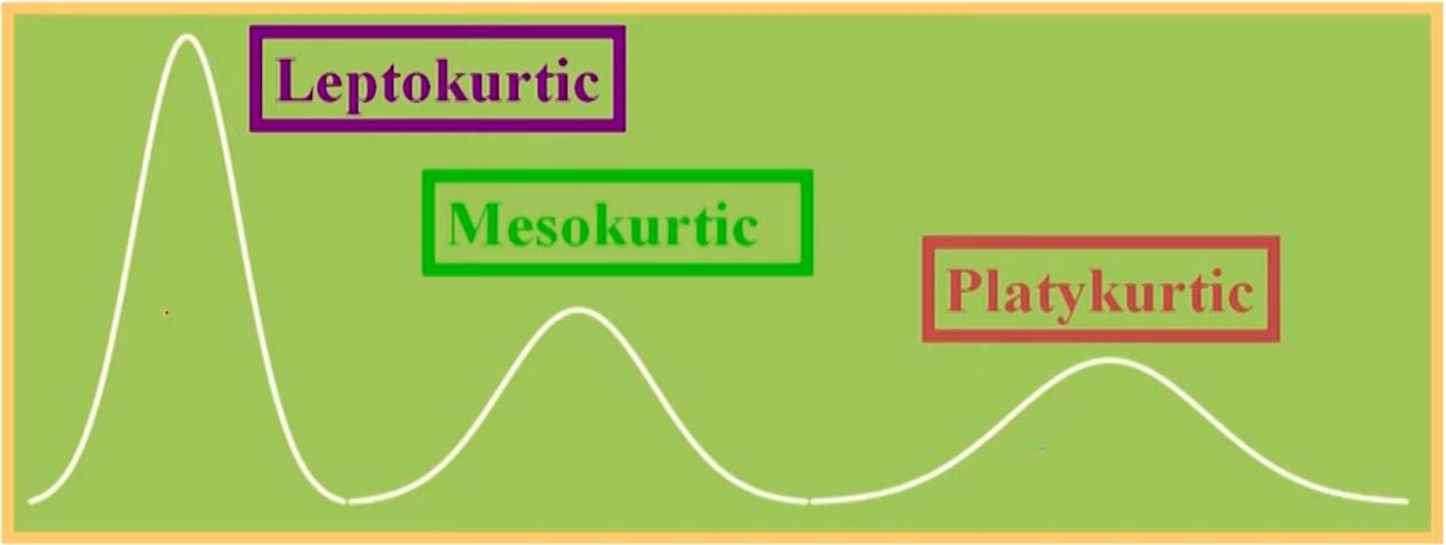
Coefficient of Skewness

$\mu_1 = 23$	$\mu_2 = 26$	$\mu_3 = 29$
$M_{d_1} = 26$	$M_{d_2} = 26$	$M_{d_3} = 26$
$\sigma_1 = 12.3$	$\sigma_2 = 12.3$	$\sigma_3 = 12.3$
$S_1 = \dfrac{3(\mu_1 - M_{d_1})}{\sigma_1}$	$S_2 = \dfrac{3(\mu_2 - M_{d_2})}{\sigma_2}$	$S_3 = \dfrac{3(\mu_3 - M_{d_3})}{\sigma_3}$
$= \dfrac{3(23-26)}{12.3}$	$= \dfrac{3(26-26)}{12.3}$	$= \dfrac{3(29-26)}{12.3}$
$= -0.73$	$= 0$	$= +0.73$

Kurtosis

Peakedness Of a distribution

- Leptokurtic: high and thin
- Mesokurtic: normal in shape
- Platykurtic: flat and spread out

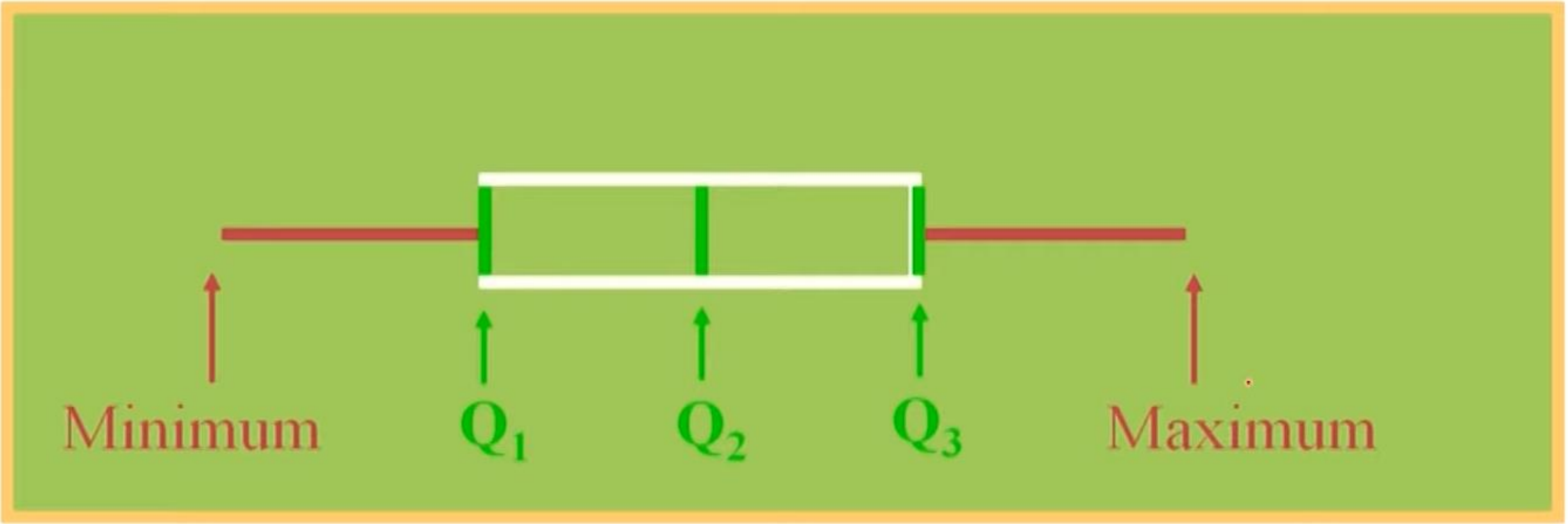


Box and Whisker Plot

Five specific values are used:

- Median, Q2
- First quartile, Q1
- Third quartile, Q3
- Minimum value in the data set
- Maximum value in the data set

Box and Whisker Plot



Skewness: Box and Whisker Plots, and Coefficient of Skewness