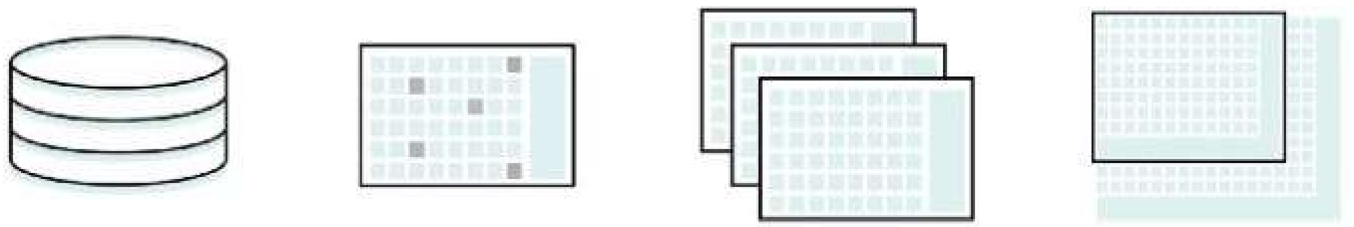
# DATA ANALYTICS

# Different roles of Data analytics



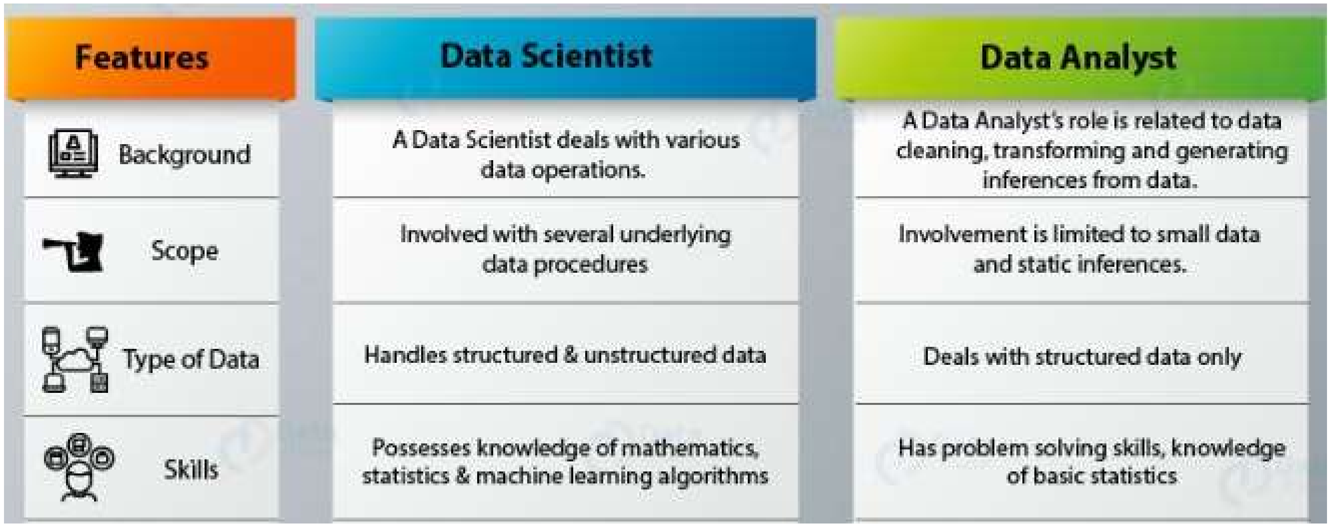
**Managing:**
ensuring the secure storage of data

**Cleansing:**
removing incorrect or biased data

**Aggregating:**
Compiling data from multiple data sources

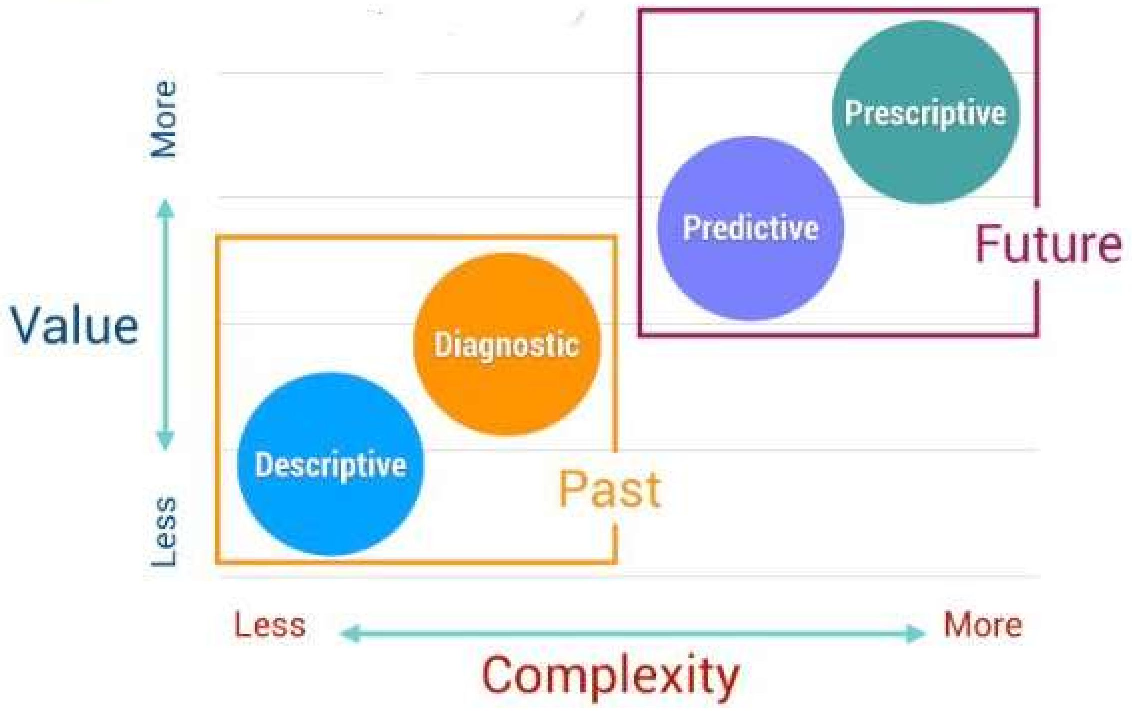
**Abstracting:**
Reducing a data set to its essential characteristics

# Data Science and Data Analytics.

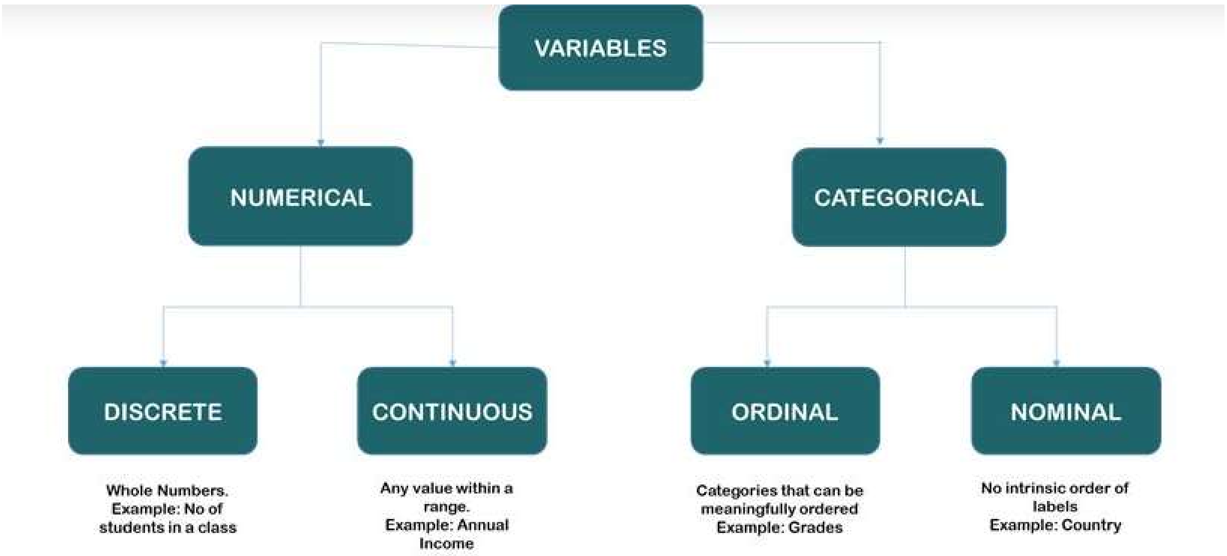


| Features | Data Scientist | Data Analyst |
|---|---|---|
| Background | A Data Scientist deals with various data operations. | A Data Analyst's role is related to data cleaning, transforming and generating inferences from data. |
| Scope | Involved with several underlying data procedures | Involvement is limited to small data and static inferences. |
| Type of Data | Handles structured & unstructured data | Deals with structured data only |
| Skills | Possesses knowledge of mathematics, statistics & machine learning algorithms | Has problem solving skills, knowledge of basic statistics |

# Data Analytics



# Identification of variables



VARIABLES

NUMERICAL

CATEGORICAL

DISCRETE

Whole Numbers.
Example: No of
students in a class

CONTINUOUS

Any value within a
range.
Example: Annual
Income

ORDINAL

Categories that can be
meaningfully ordered
Example: Grades

NOMINAL

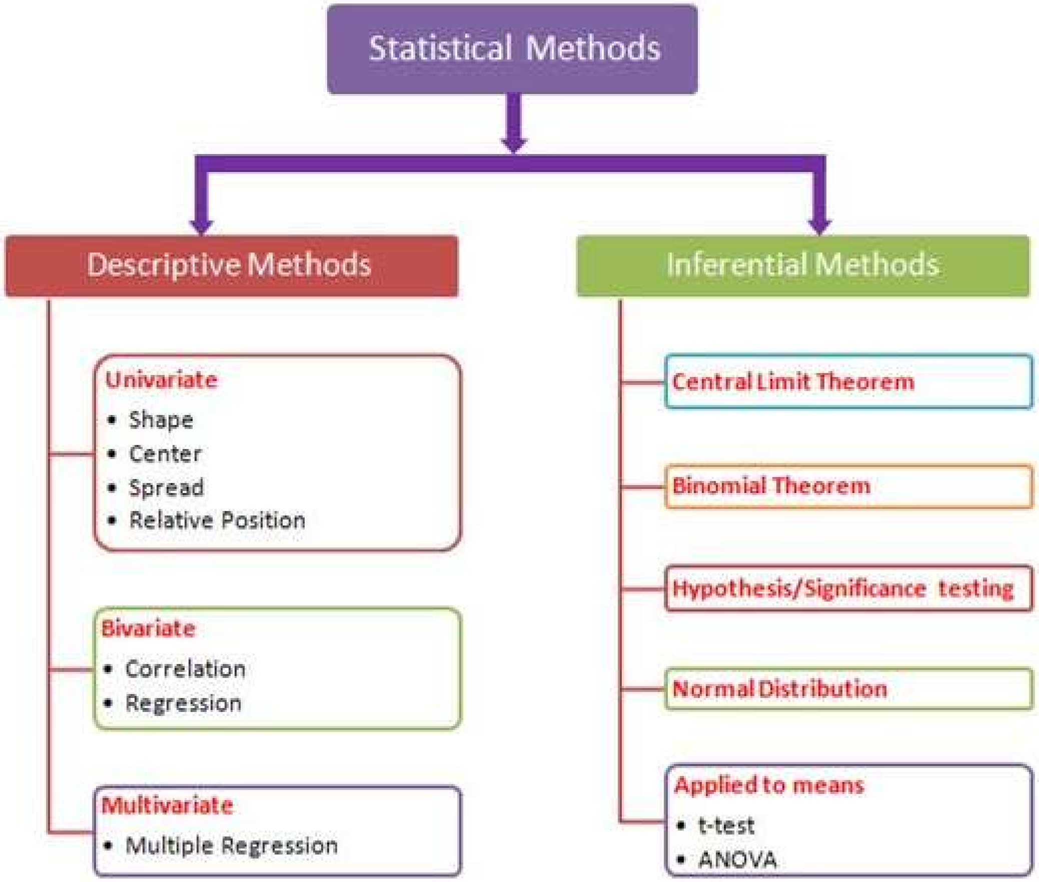
No intrinsic order of
labels
Example: Country

Classification of Variables

# Measurement Scales

- Interval Scale
    - Data classified by ranking.
    - Quantitative classification (time, temperature, etc).
    - Zero point of scale is arbitrary (differences are meaningful).

- Ratio Scale
    - Data classified as the ratio of two numbers.
    - Quantitative classification (height, weight, distance, etc).
    - Zero point of scale is real (data can be added, subtracted, multiplied, and divided).



Statistical Methods

Descriptive Methods

Univariate
- Shape
- Center
- Spread
- Relative Position

Bivariate
- Correlation
- Regression

Multivariate
- Multiple Regression

Inferential Methods

Central Limit Theorem

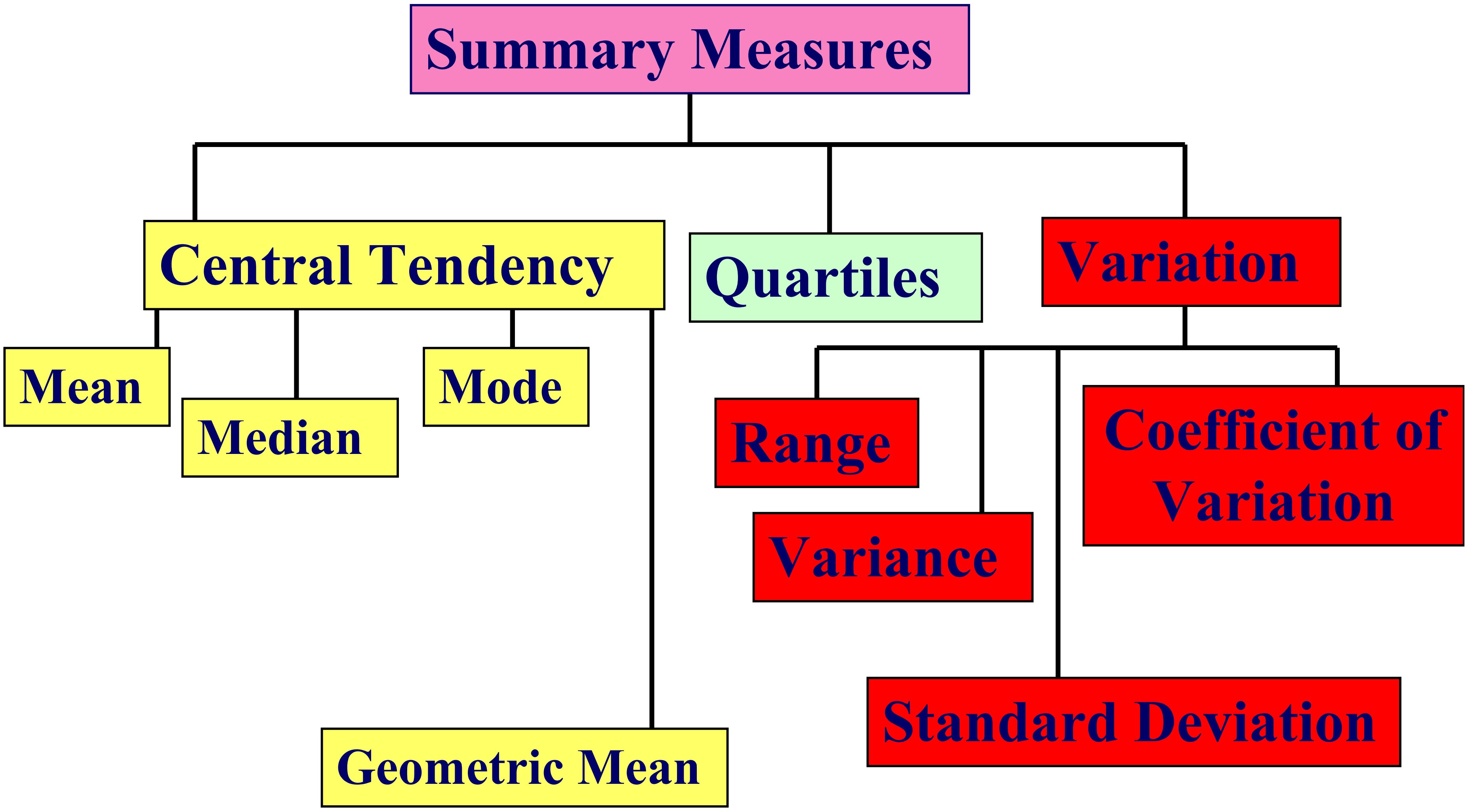
Binomial Theorem

Hypothesis/Significance testing

Normal Distribution

Applied to means
- t-test
- ANOVA

# Summary Measures

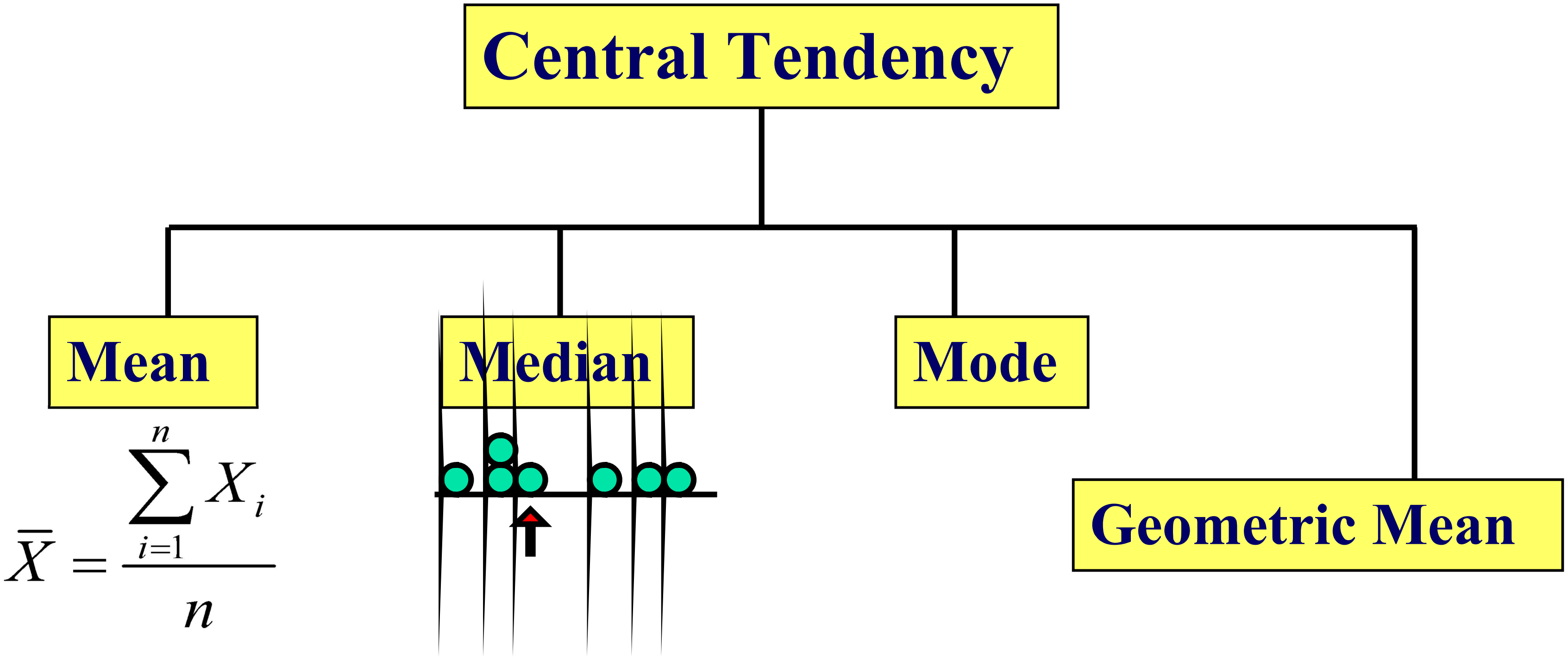


# Central Tendency and Measures of dispersion

# Measures of central tendency

- Yield information about particular places or locations in a group of numbers.
- That is a single number to describe the characteristics of a set of data.

# Measures of Central Tendency



$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

# Arithmetic mean

- Commonly it is called as the mean it is the average of a group of numbers; it is applicable for interval and ratio data.
- Not applicable for nominal and ordinal data.
- It is affected by each value in the data set including extreme values, one of the problem of the with the mean is that it is affected by the extreme values computed by summing all values in the data set and dividing the sum by the number of values in the data set.

# Mean (Arithmetic Mean)

- Mean (Arithmetic Mean) of Data Values
  - Sample mean

Sample Size

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

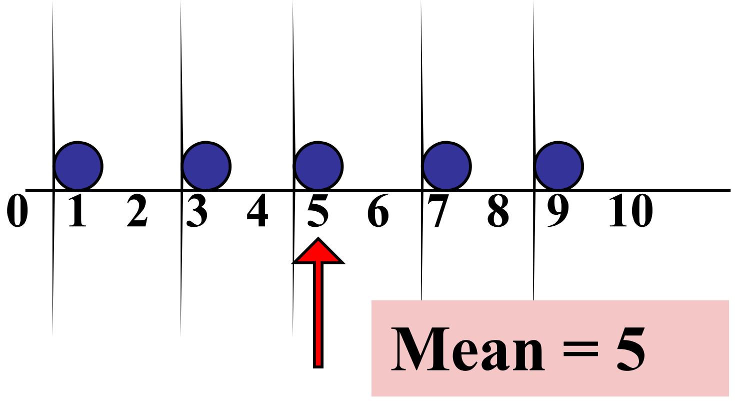
  - Population mean

Population Size

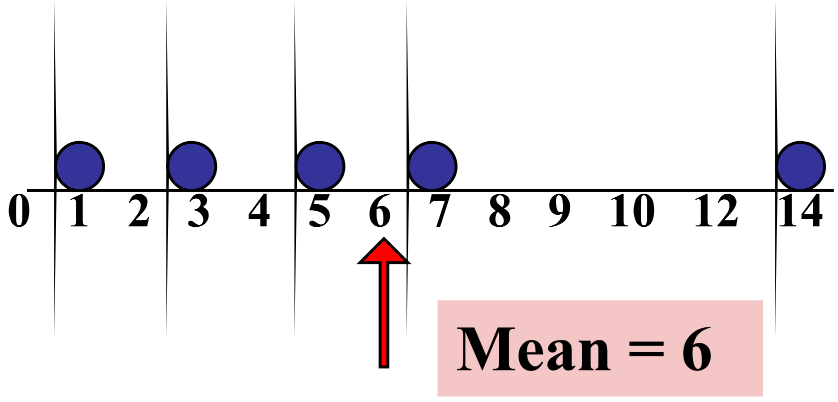
$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

# Mean (Arithmetic Mean)*(continued)*

- The Most Common Measure of Central Tendency
- Affected by Extreme Values (Outliers)



Mean = 5



Mean = 6

# Mean (Arithmetic Mean) From a Frequency Distribution

- Approximating the Arithmetic Mean
  - Used when raw data are not available

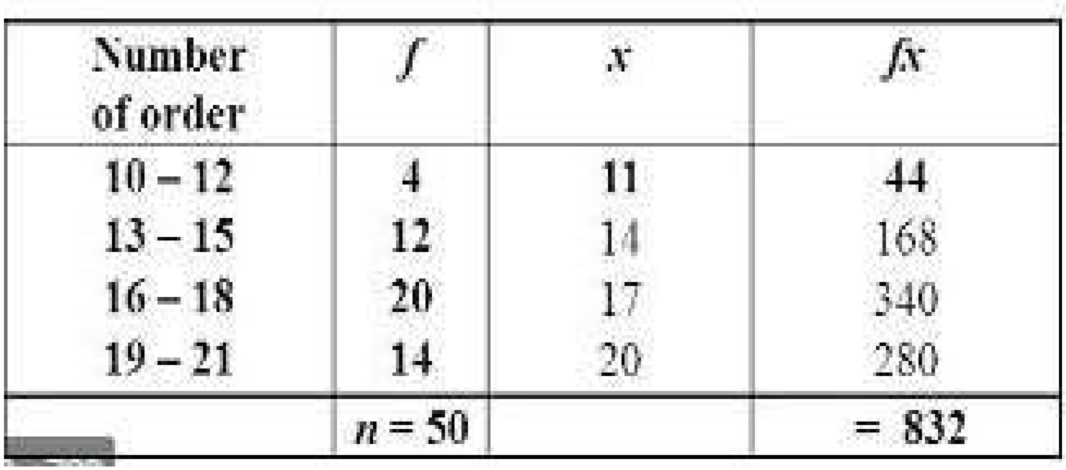
$$\bar{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

$n =$ sample size

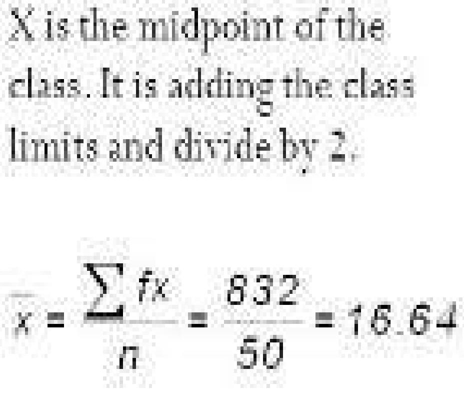
$c =$ number of classes in the frequency distribution

$m_j =$ midpoint of the $j$th class

$f_j =$ frequencies of the $j$th class



| Number of order | $f$ | $x$ | $fx$ |
|---|---|---|---|
| 10 – 12 | 4 | 11 | 44 |
| 13 – 15 | 12 | 14 | 168 |
| 16 – 18 | 20 | 17 | 340 |
| 19 – 21 | 14 | 20 | 280 |
| | $n = 50$ | | = 832 |



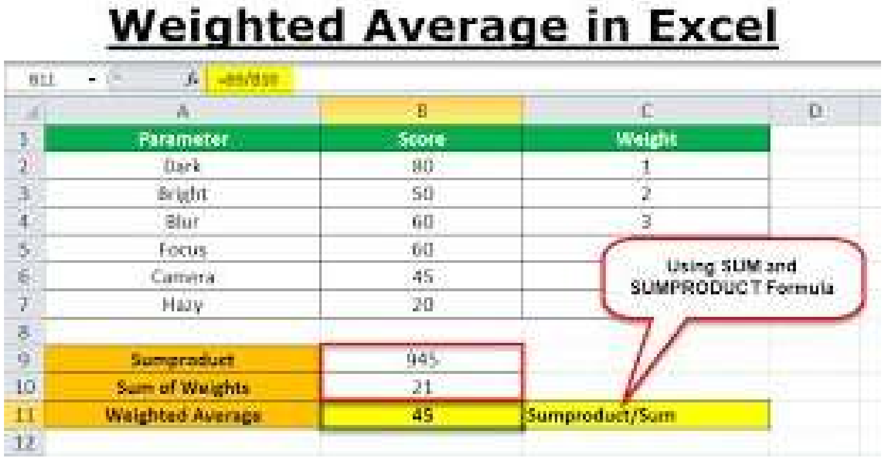
X is the midpoint of the class. It is adding the class limits and divide by 2.

$$\bar{x} = \frac{\sum fx}{n} = \frac{832}{50} = 16.64$$

# Weighted average

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

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 in Excel

B11 =B9/B10

|  | A | B | C | D |
|---|---|---|---|---|
| 1 | Parameter | Score | Weight |  |
| 2 | Dark | 80 | 1 |  |
| 3 | Bright | 50 | 2 |  |
| 4 | Blur | 60 | 3 |  |
| 5 | Focus | 60 |  |  |
| 6 | Camera | 45 |  |  |
| 7 | Hazy | 20 |  |  |
| 8 |  |  |  |  |
| 9 | Sumproduct | 945 |  |  |
| 10 | Sum of Weights | 21 |  |  |
| 11 | Weighted Average | 45 | Sumproduct/Sum |  |
| 12 |  |  |  |  |

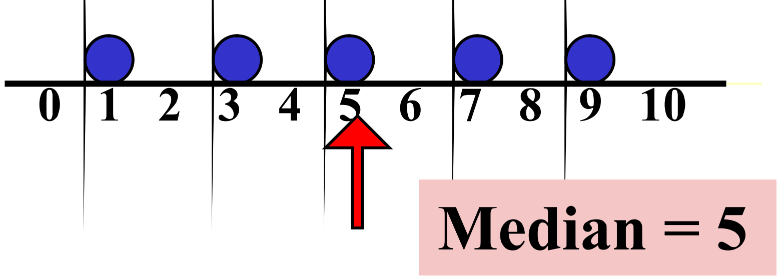
Using SUM and SUMPRODUCT Formula

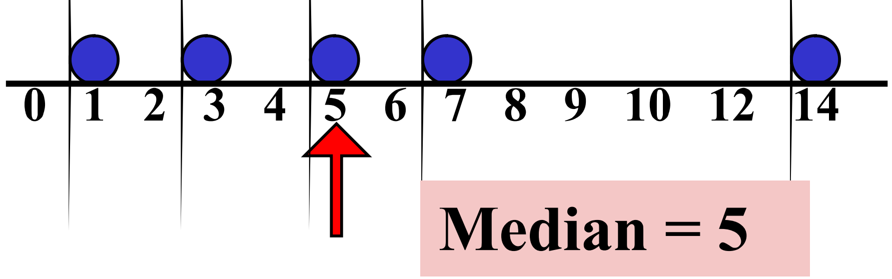
# Median

- Median, the middle value in ordered array of number is called Median. It is applicable for ordinal interval and ratio data.
- It is not applicable for nominal data and one advantage of median is it is unaffected by extremely large and extremely small values.

# Median

- Robust Measure of Central Tendency
- Not Affected by Extreme Values





- In an Ordered Array, the Median is the 'Middle' Number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the 2 middle numbers

# Median of grouped data

$$M_m = l + \left(\frac{\frac{n}{2} - cf}{f}\right)h$$

Where

l = lower limit of median class,

n = number of observations,

cf = cumulative frequency of class precedi

the median class,

f = frequency of median class,

h = class size (assuming class size to be equal)

| Class Interval | Frequency | Cumulative Frequency |
|---|---|---|
| 1–under 3 | 4 | 4 |
| 3–under 5 | 12 | 16 |
| 5–under 7 | 13 | 29 |
| 7–under 9 | 19 | 48 |
| 9–under 11 | 7 | 55 |
| 11–under 13 | 5 | 60 |
|  | 60 |  |

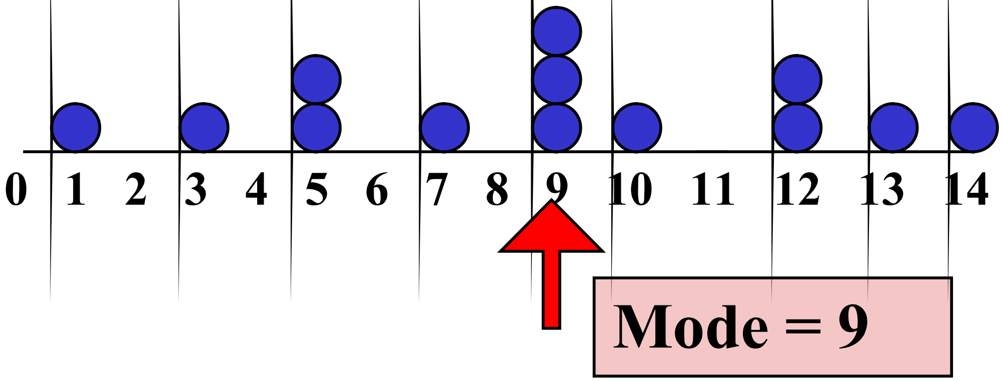
$$\text{median}_{\text{grouped}} = 7 + \left(\frac{\frac{60}{2} - 29}{19}\right)2 = 7 + \left(\frac{1}{19}\right)2 = 7 + .105 = 7.105$$

# Mode

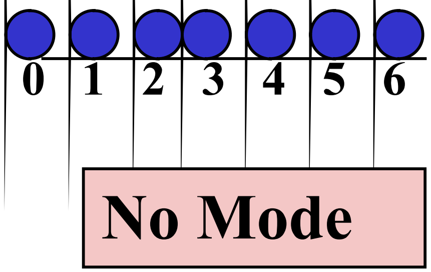
- The most frequently occurring value in a data set is mode applicable to all level of data, measurement nominal, ordinal, interval and ratio.
- Sometimes there is a possibility the data set may be bimodal.
- Bimodal means data sets that have two modes. That means two numbers are repeated same number of time multimodal data sets that contain more than two modes.

# Mode

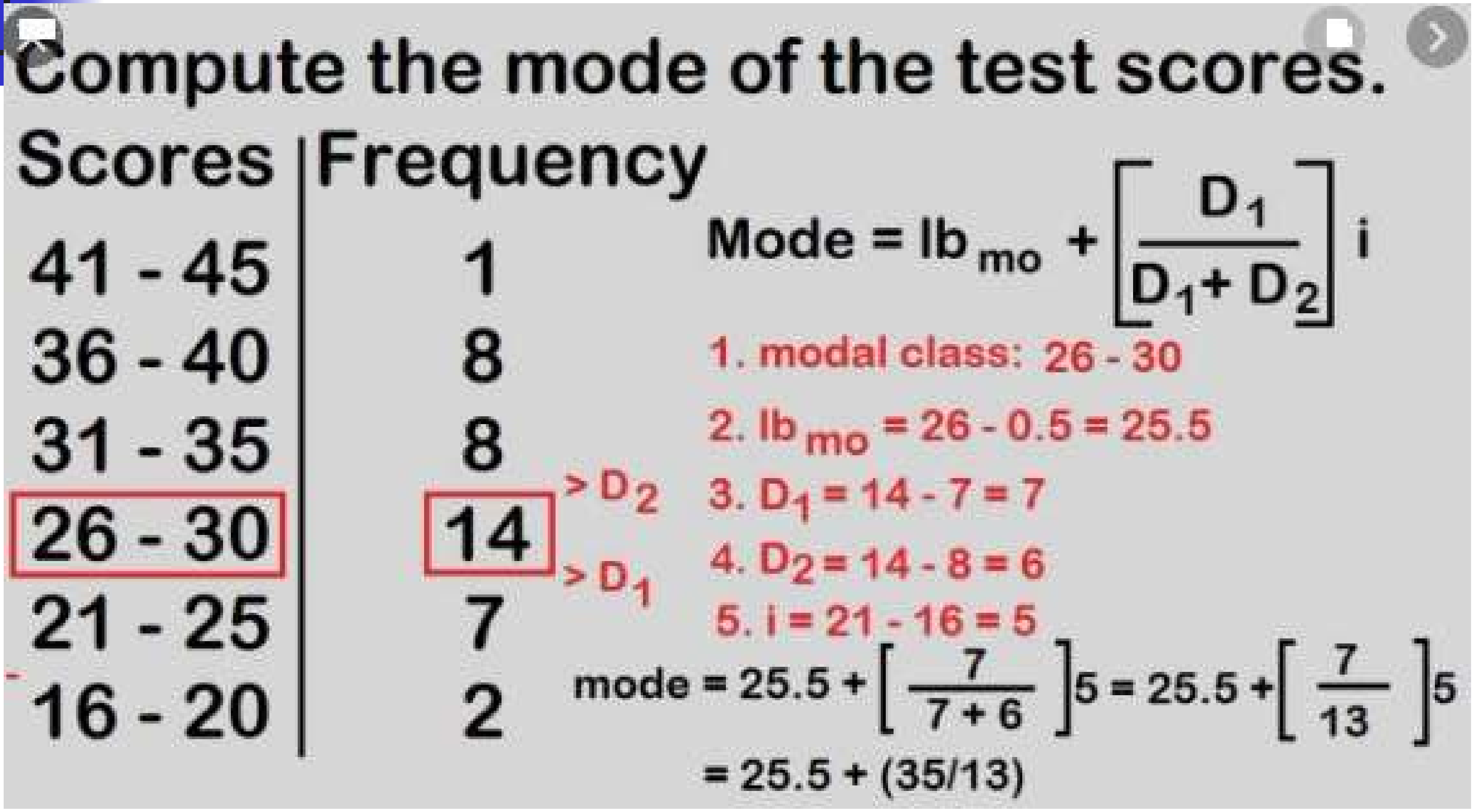
- A Measure of Central Tendency
- Value that Occurs Most Often
- Not Affected by Extreme Values
- There May Not Be a Mode
- There May Be Several Modes
- Used for Either Numerical or Categorical Data



Mode = 9



No Mode



Compute the mode of the test scores.

| Scores | Frequency |
|---|---|
| 41 - 45 | 1 |
| 36 - 40 | 8 |
| 31 - 35 | 8 |
| $\boxed{26 - 30}$ | $\boxed{14}$ |
| 21 - 25 | 7 |
| 16 - 20 | 2 |

$$\text{Mode} = lb_{mo} + \left[\frac{D_1}{D_1 + D_2}\right] i$$

1. modal class: 26 - 30

2. $lb_{mo} = 26 - 0.5 = 25.5$

3. $D_1 = 14 - 7 = 7$

4. $D_2 = 14 - 8 = 6$

5. $i = 21 - 16 = 5$

$$\text{mode} = 25.5 + \left[\frac{7}{7+6}\right]5 = 25.5 + \left[\frac{7}{13}\right]5$$

$$= 25.5 + (35/13)$$

# Geometric Mean

- Useful in the Measure of Rate of Change of a Variable Over Time

$$\bar{X}_G = \left( X_1 \times X_2 \times \cdots \times X_n \right)^{1/n}$$

- Geometric Mean Rate of Return
  - Measures the status of an investment over time

$$\bar{R}_G = \left[ (1+R_1) \times (1+R_2) \times \cdots \times (1+R_n) \right]^{1/n} - 1$$

# Example

An investment of $100,000 declined to $50,000 at the end of year one and rebounded back to $100,000 at end of year two:

$$R_1 = -0.5 \text{ (or} - 50\%) \qquad R_2 = 1 \text{ (or } 100\% \text{ )}$$

Average rate of return:

$$\bar{R} = \frac{(-0.5) + (1)}{2} = 0.25 \text{ (or } 25\%)$$

Geometric rate of return:

$$\bar{R}_G = \left[(1 - 0.5) \times (1 + 1)\right]^{1/2} - 1$$

$$= \left[(0.5) \times (2)\right]^{1/2} - 1 = 1^{1/2} - 1 = 0 \text{ (or } 0\%)$$

# Percentiles

- Divide a group of data into 100 parts it is called percentile.
- For example; somebody say 90th percentile my score is 90th percentile indicates that at most 90% of the data lie below it and at least 10% the data lie above it.
- The median and the 50th percentile have the same value. It is applicable for ordinal, interval and ratio data it is not applicable for nominal data.

- Raw Data: 14, 12, 19, 23, 5, 13, 28, 17
- Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28
- Location of 30th percentile:

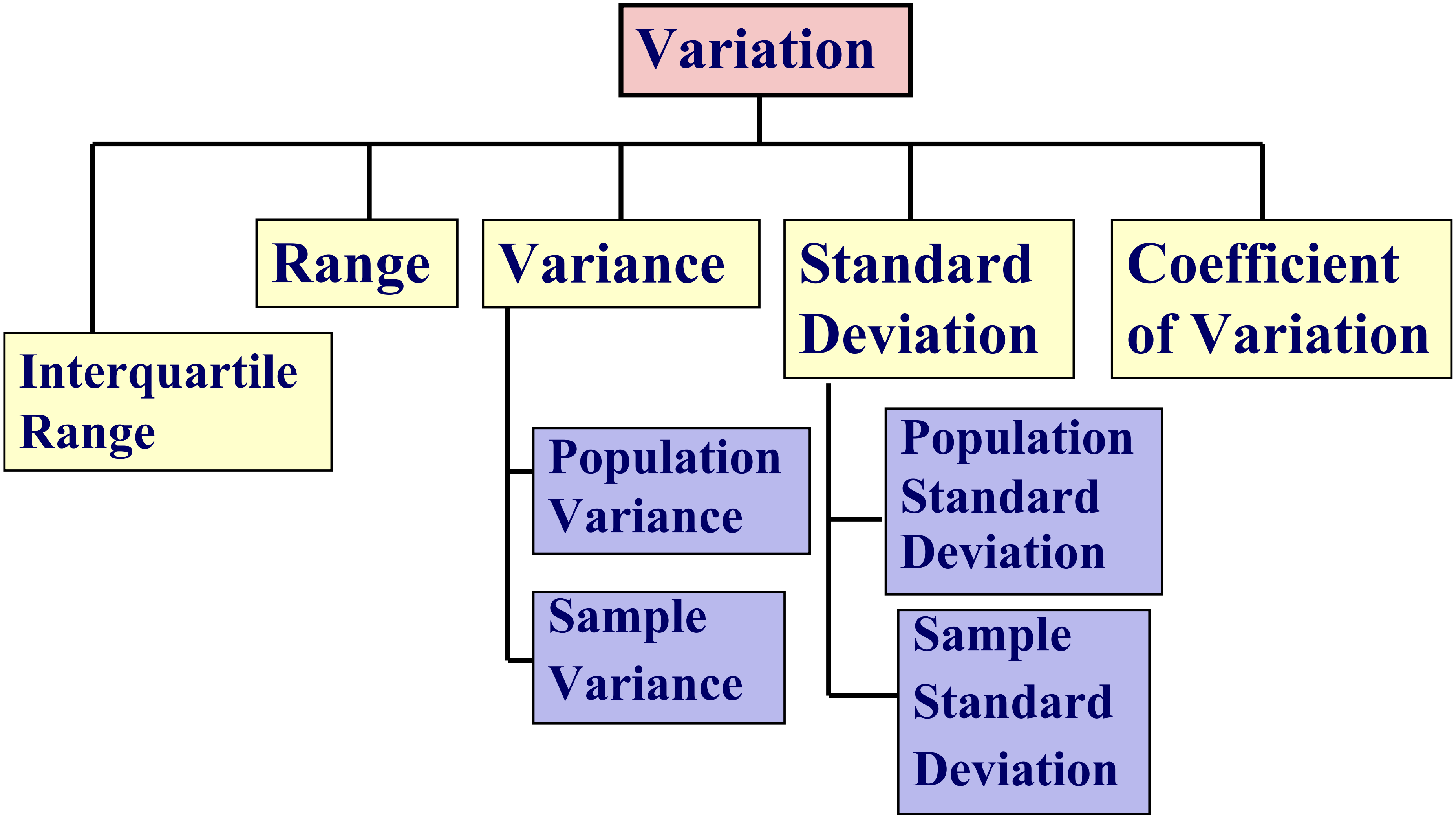
$$i = \frac{30}{100}(8) = 2.4$$

- 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

- Describes the spread or the dispersion of the set of the data.
- The reliability of measure of central tendency is the dispersion because many times, the central tendency will mislead the people.
- So the reliability of that central tendency is calculated by or identified by its corresponding dispersion.

# Measures of Variation

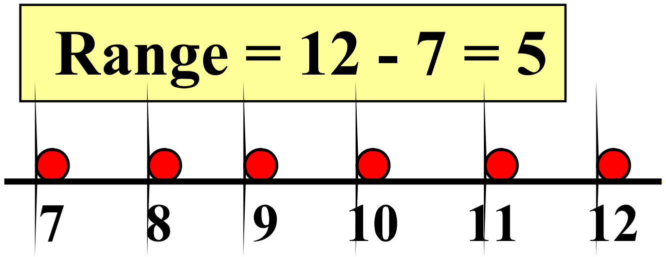


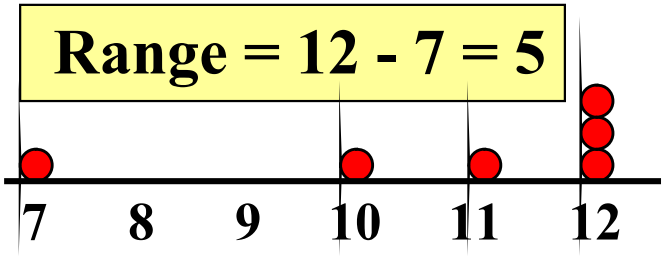
# Range

- Measure of Variation
- Difference between the Largest and the Smallest Observations:

$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$

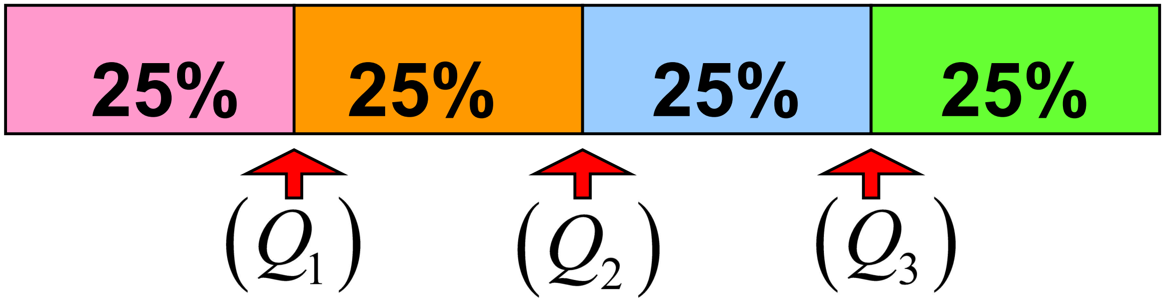
- Ignores How Data are Distributed





# Quartiles

- Split Ordered Data into 4 Quarters



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

- 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
- $Q_1$ 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

# Variance

- Important Measure of Variation
- Shows Variation about the Mean
  - Sample Variance:

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

  - Population Variance:

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

# Standard Deviation

- Most Important Measure of Variation
- Shows Variation about the Mean
- Has the Same Units as the Original Data
  - Sample Standard Deviation:

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

  - Population Standard Deviation:

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

# Standard Deviation From a Frequency Distribution *(continued)*

- Approximating the Standard Deviation
  - Used when the raw data are not available and the only source of data is a frequency distribution
  - $$S = \sqrt{\frac{\sum_{j=1}^{c} \left(m_j - \bar{X}\right)^2 f_j}{n-1}}$$

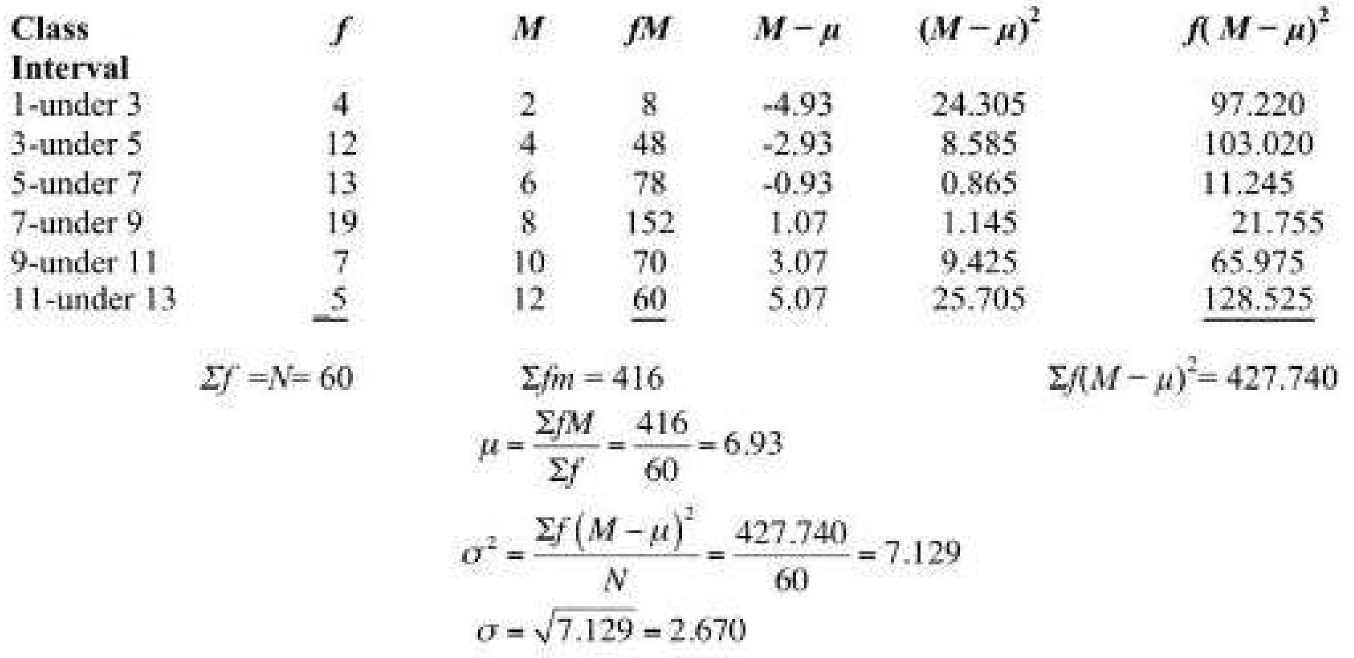
$n =$ sample size

$c =$ number of classes in the frequency distribution

$m_j =$ midpoint of the $j$th class

$f_j =$ frequencies of the $j$th class

# Variance and standard deviation of the grouped data



| Class Interval | $f$ | $M$ | $fM$ | $M-\mu$ | $(M-\mu)^2$ | $f(M-\mu)^2$ |
|---|---|---|---|---|---|---|
| 1-under 3 | 4 | 2 | 8 | -4.93 | 24.305 | 97.220 |
| 3-under 5 | 12 | 4 | 48 | -2.93 | 8.585 | 103.020 |
| 5-under 7 | 13 | 6 | 78 | -0.93 | 0.865 | 11.245 |
| 7-under 9 | 19 | 8 | 152 | 1.07 | 1.145 | 21.755 |
| 9-under 11 | 7 | 10 | 70 | 3.07 | 9.425 | 65.975 |
| 11-under 13 | 5 | 12 | 60 | 5.07 | 25.705 | 128.525 |
| | $\Sigma f = N = 60$ | | $\Sigma fm = 416$ | | | $\Sigma f(M-\mu)^2 = 427.740$ |

$$\mu = \frac{\Sigma fM}{\Sigma f} = \frac{416}{60} = 6.93$$

$$\sigma^2 = \frac{\Sigma f(M-\mu)^2}{N} = \frac{427.740}{60} = 7.129$$

$$\sigma = \sqrt{7.129} = 2.670$$

# Measures of Dispersion: Summary Characteristics

- The **more** the data are spread out, the **greater** the range, variance, and standard deviation.
- The **less** the data are spread out, 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**.

# Mean absolute deviation (MAD)

The mean absolute deviation of a dataset is the average distance between each data point and the mean. It gives us an idea about the variability in a dataset.

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|$$

$m(X)$ = average value of the data set

$n$ = number of data values

$x_i$ = data values in the set

# Coefficient of Variation

- Measure of Relative Variation
- Always in Percentage (%)
- Shows Variation Relative to the Mean
- Used to Compare Two or More Sets of Data Measured in Different Units
- $$CV = \left( \frac{S}{\bar{X}} \right) 100\%$$
- Sensitive to Outliers

# Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $2
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5
- Coefficient of Variation:
  - Stock A:

$$CV = \left(\frac{S}{\overline{\overline{X}}}\right)100\% = \left(\frac{\$2}{\$50}\right)100\% = 4\%$$

  - Stock B:

$$CV = \left(\frac{S}{\overline{\overline{X}}}\right)100\% = \left(\frac{\$5}{\$100}\right)100\% = 5\%$$

# Covariance

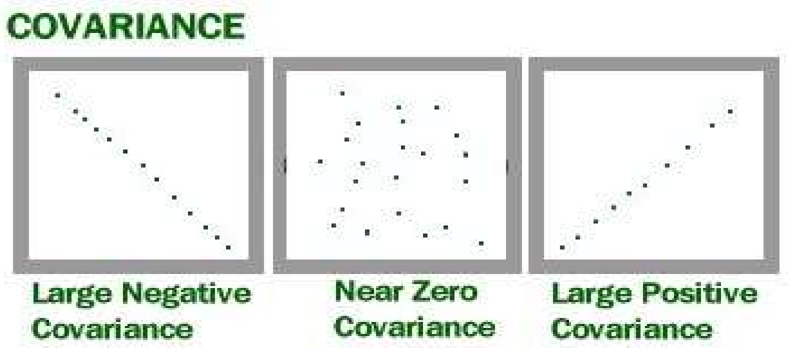
- Covariance is a measure of how much two random variables vary together. It’s similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \overline{x}) * (y_i - \overline{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \overline{x}) * (y_i - \overline{y})}{(N - 1)}$$

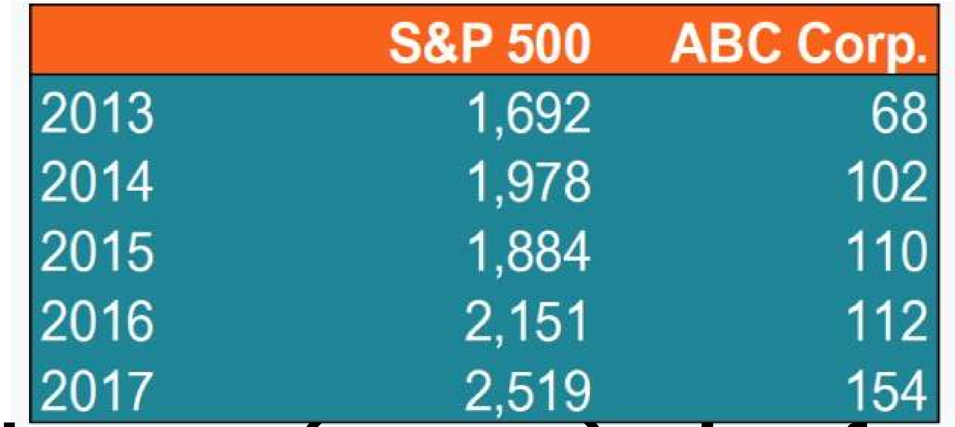


John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the directional relationship between the stock and the S&P 500. John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction.

John can calculate the covariance between the stock of ABC Corp. and S&P 500 by following the steps below

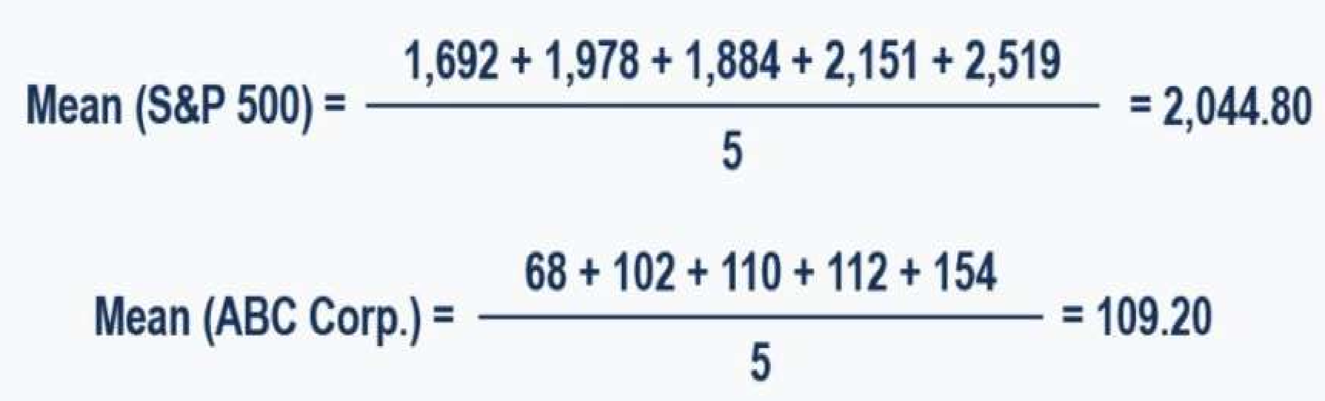
## 1. Obtain the data.

First, John obtains the figures for both ABC Corp. stock and the S&P 500. The prices obtained are summarized in the table below:



| | S&P 500 | ABC Corp. |
|---|---|---|
| 2013 | 1,692 | 68 |
| 2014 | 1,978 | 102 |
| 2015 | 1,884 | 110 |
| 2016 | 2,151 | 112 |
| 2017 | 2,519 | 154 |

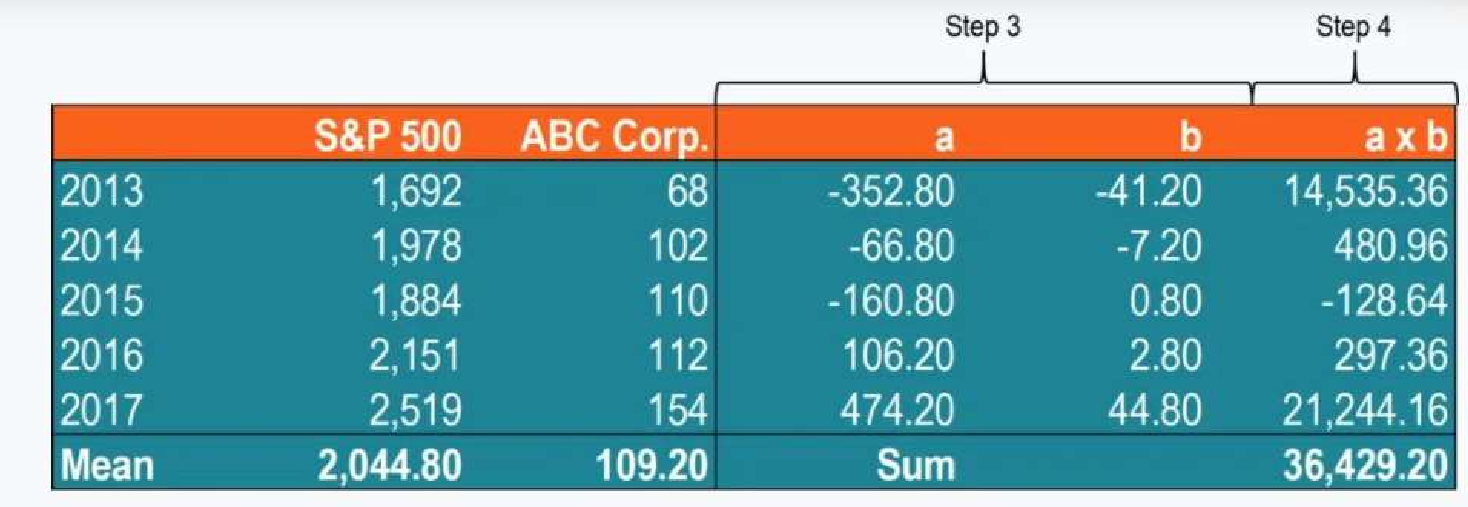
## 2. Calculate the mean (average) prices for each asset.



$$\text{Mean (S\&P 500)} = \frac{1{,}692 + 1{,}978 + 1{,}884 + 2{,}151 + 2{,}519}{5} = 2{,}044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

3. **For each security, find the difference between each value and mean price.**



| | | | Step 3 | | Step 4 |
|---|---|---|---|---|---|
| | S&P 500 | ABC Corp. | a | b | a x b |
| 2013 | 1,692 | 68 | -352.80 | -41.20 | 14,535.36 |
| 2014 | 1,978 | 102 | -66.80 | -7.20 | 480.96 |
| 2015 | 1,884 | 110 | -160.80 | 0.80 | -128.64 |
| 2016 | 2,151 | 112 | 106.20 | 2.80 | 297.36 |
| 2017 | 2,519 | 154 | 474.20 | 44.80 | 21,244.16 |
| **Mean** | **2,044.80** | **109.20** | **Sum** | | **36,429.20** |

4. **Multiply the results obtained in the previous step.**
5. **Using the number calculated in step 4, find the covariance.**

$$\text{Cov(S\&P 500, ABC Corp.)} = \frac{36{,}429.20}{5-1} = 9{,}107.30$$

In such a case, the positive covariance indicates that the price of the stock and the S&P 500 tend to move in the same direction.

# Correlation

- The main problem with interpretation is that the wide range of results that it takes on makes it hard to interpret.
- **Correlation coefficients** are used to measure how strong a relationship is between two variables.
- There are several types of correlation coefficient, but the most popular is Pearson's. **Pearson's correlation** (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

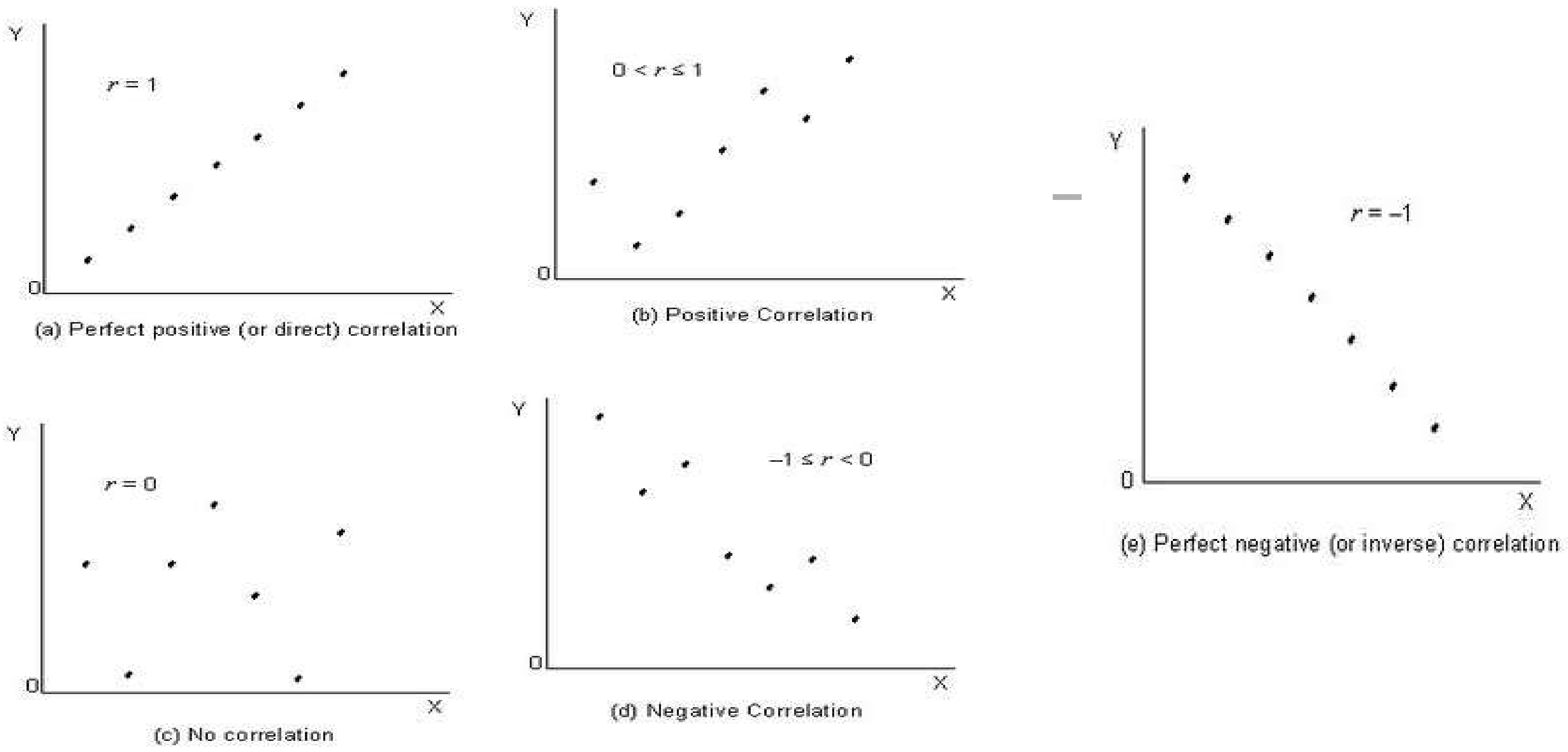
$r$ = correlation coefficient

$x_i$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_i$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable



(a) Perfect positive (or direct) correlation

(b) Positive Correlation

(c) No correlation

(d) Negative Correlation

(e) Perfect negative (or inverse) correlation

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

1 indicates a strong positive relationship.

-1 indicates a strong negative relationship.

A result of zero indicates no relationship at all.

# Measures of shape

- Skewness

  absence of symmetry

- kurtosis

  the peakedness of a distribution.

  There are three layers, there are Leptokurtic, Mesokurtic, Platykurtic.

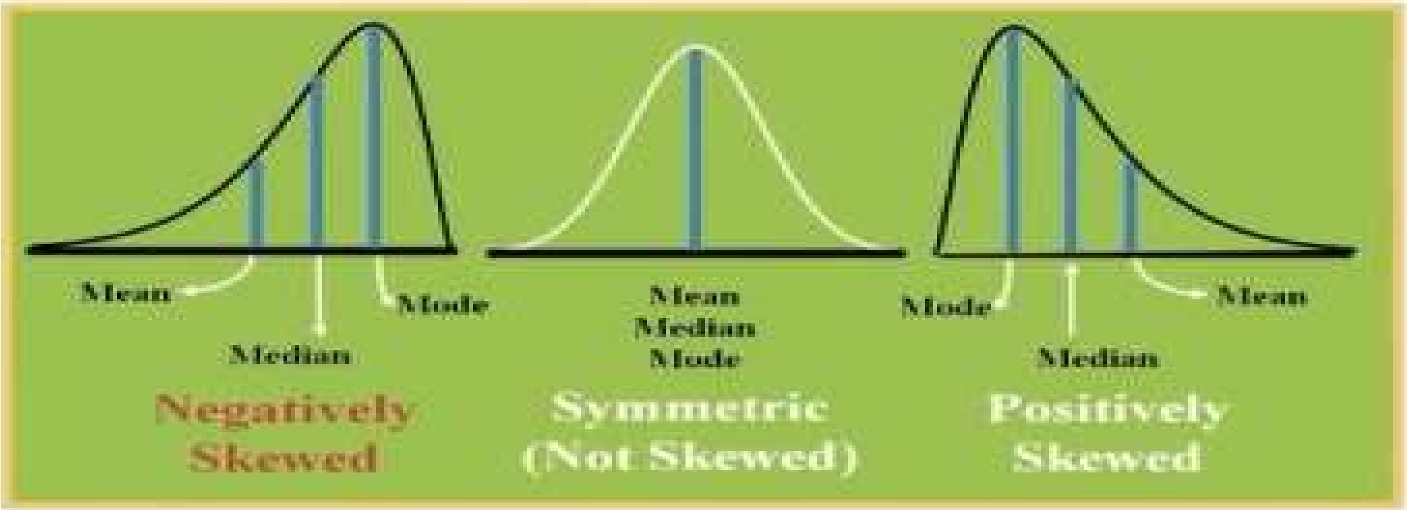
- box and whisker plots

  It is a graphical display of distribution.

  It reveals skewness.

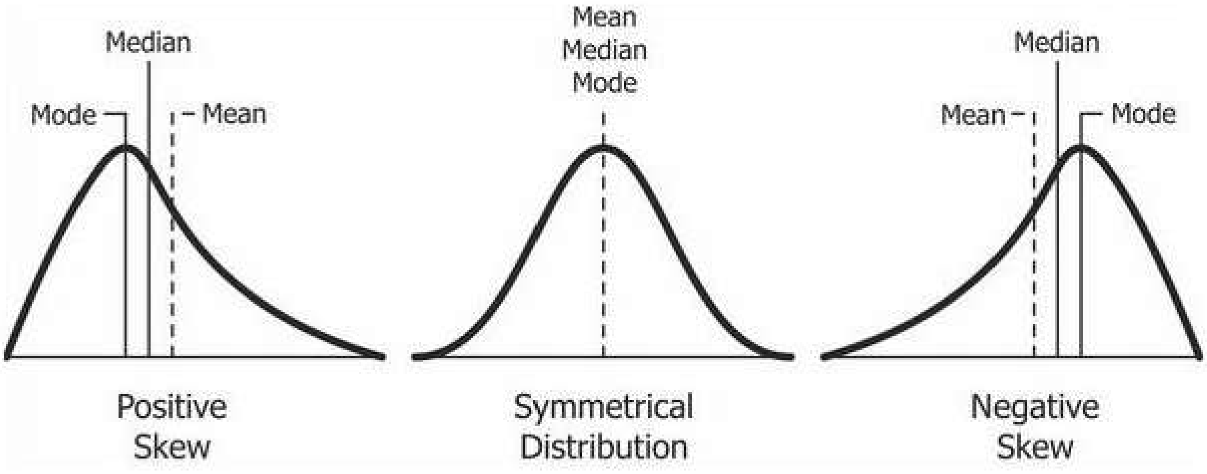
# Skewness

- Skewness, in statistics, is the degree of distortion from the symmetrical bell curve in a probability distribution.
- Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees.



# Measures of Skewness

- If Mean > Mode, the skewness is positive.
- If Mean < Mode, the skewness is negative.
- If Mean = Mode, the skewness is zero.



# Coefficient of skewness

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

$$Sk_2 = \frac{3\bar{X} - Md}{s}$$

**where:**

$Sk_1$ = Pearson's first coefficient of skewness and $Sk_2$ the second Pearson's first coefficient of skewness

$s$ = the standard deviation for the sample

$\bar{X}$ = is the mean value

$Mo$ = the modal (mode) value

$Md$ = is the median value

# Kurtosis (Ku)

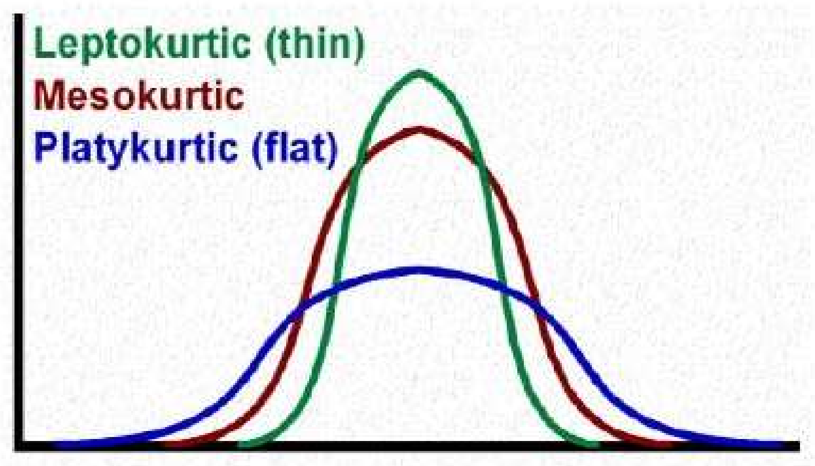
**Measure of relative peakedness of a distribution. It is a shape parameter that characterizes the degree of peakedness.**

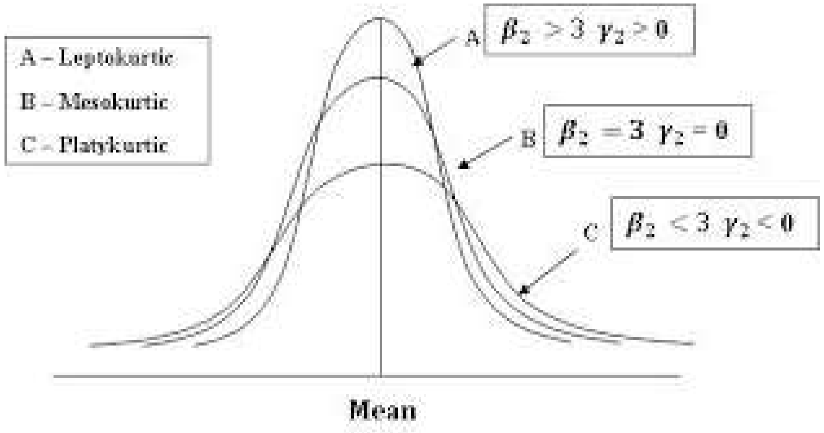
When the peak of a curve becomes relatively high then that curve is called Leptokurtic.

When the curve is flat-topped, then it is called Platykurtic.

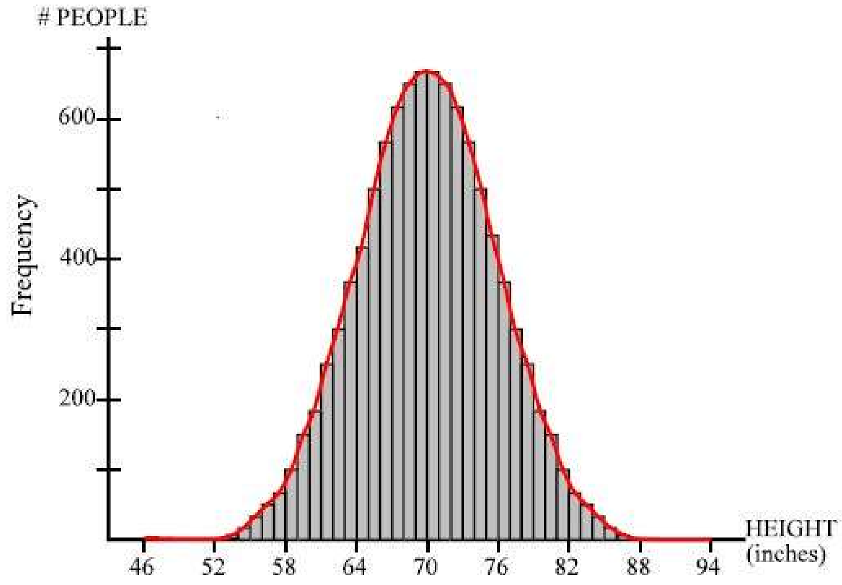
Since normal curve is neither very peaked nor very flat topped, so it is taken as a basis for comparison.

The normal curve is called Mesokurtic.




A – Leptokurtic
B – Mesokurtic
C – Platykurtic
A $\beta_2 > 3$ $\gamma_2 > 0$
B $\beta_2 = 3$ $\gamma_2 = 0$
C $\beta_2 < 3$ $\gamma_2 < 0$
Mean

# Normal distribution



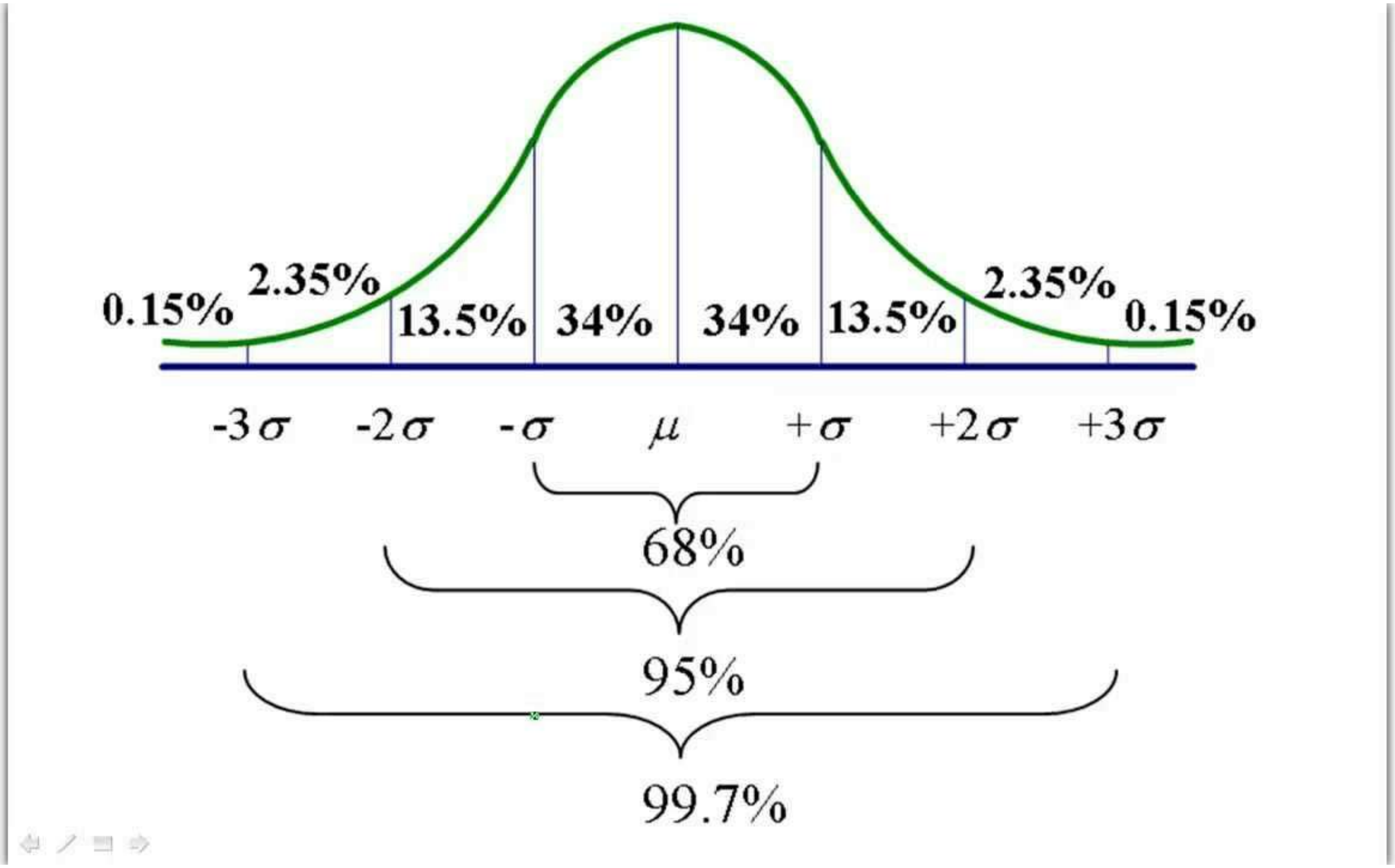
- Skewness = 0
- Kurtosis = 3

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

# The Empirical Rule

- For Data Sets That Are Approximately Bell-shaped:
  - Roughly 68% of the Observations Fall Within 1 Standard Deviation Around the Mean
  - Roughly 95% of the Observations Fall Within 2 Standard Deviations Around the Mean
  - Roughly 99.7% of the Observations Fall Within 3 Standard Deviations Around the Mean

# The Empirical Rule



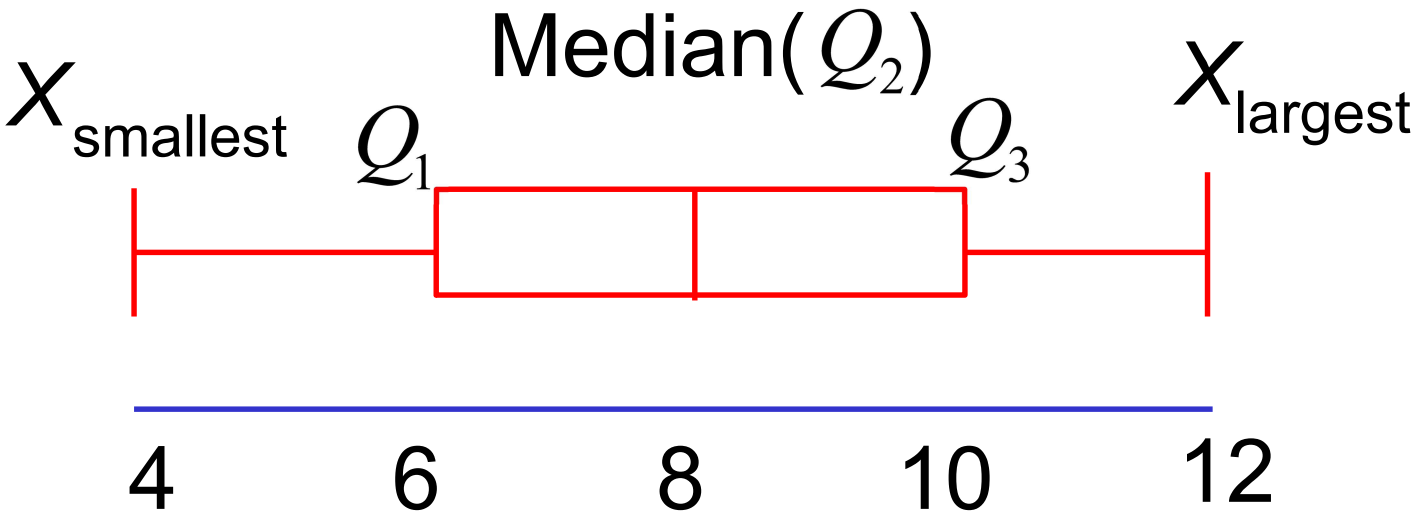
# Example

- The weights of stray dogs at a particular pound average 70 lbs with a standard deviation of 2.5 lbs. Assuming the weights follow a Gaussian distribution:
- What weight is 2 standard deviations below the mean?
- What weight is 1 standard deviation above the mean?
- The middle 68% of dogs weigh how much?

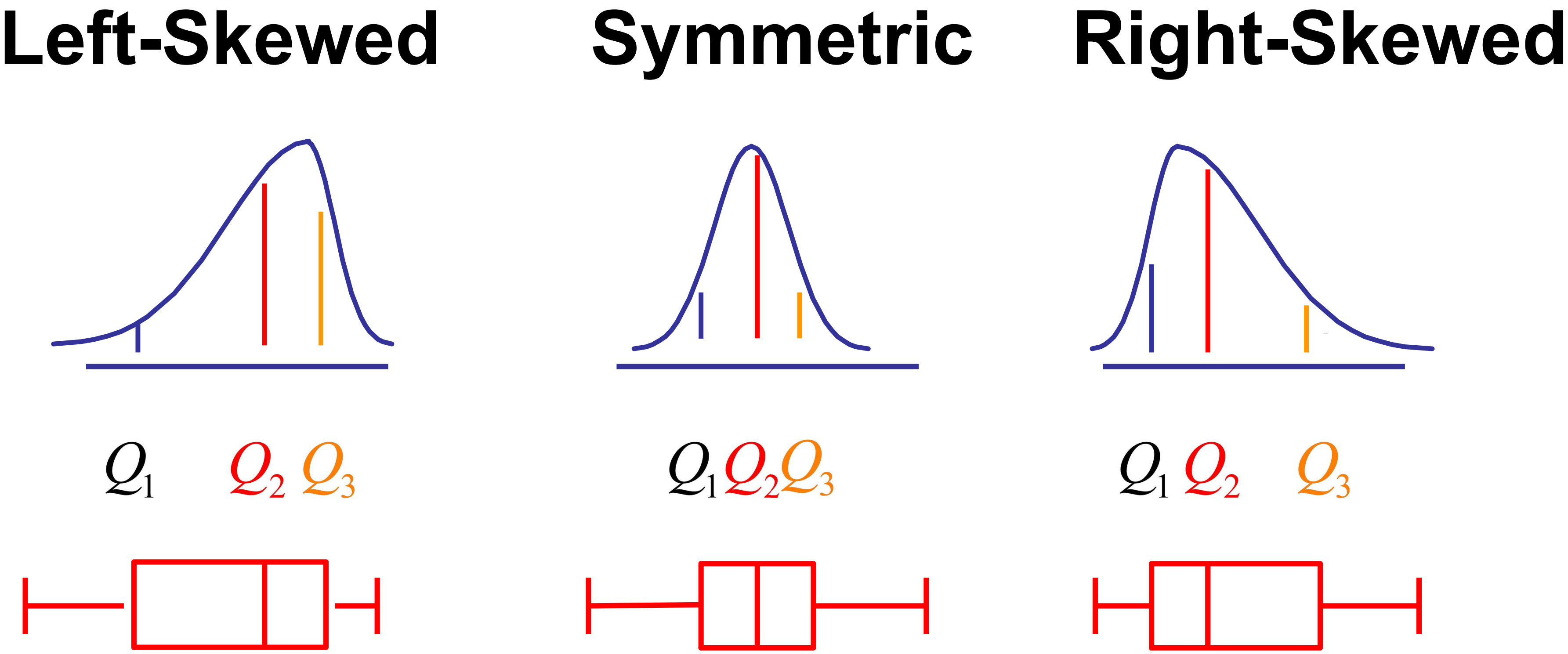
- 2 standard deviations is 2 * 2.5 (5 lbs). So if a dog is 2.5 standard deviations below the mean they weigh 70 lbs – 5 lbs = 65 lbs.
- 1 standard deviation is 2.5 lbs, so a dog 1 standard deviation above the mean would weigh 70 lbs + 2.5 lbs = 72.5 lbs.
- The 68 95 99.7 Rule tells us that 68% of the weights should be within 1 standard deviation either side of the mean. 1 standard deviation above (given in the answer to question 2) is 72.5 lbs; 1 standard deviation below is 70 lbs – 2.5 lbs is 67.5 lbs. Therefore, 68% of dogs weigh between 67.5 and 72.5 lbs.

# Exploratory Data Analysis

- Box-and-Whisker Plot
  - Graphical display of data using 5-number summary



# Distribution Shape & Box-and-Whisker Plot

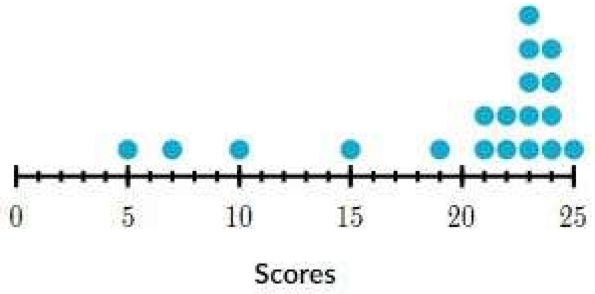


Left-Skewed

Symmetric

Right-Skewed

# Identifying outliers with the 1.5xIQR rule

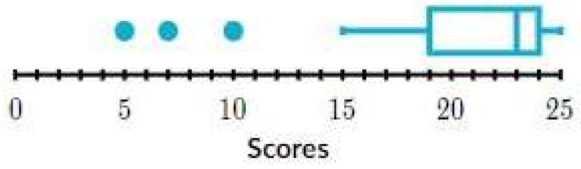


5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

**Find the median, quartiles, and interquartile range**

Step 2) Calculate $1.5 \cdot \mathrm{IQR}$ below the first quartile and check for low outliers.

Step 3) Calculate $1.5 \cdot \mathrm{IQR}$ above the third quartile and check for high outliers.


Scores