Introduction to Business Analytics

Module 2

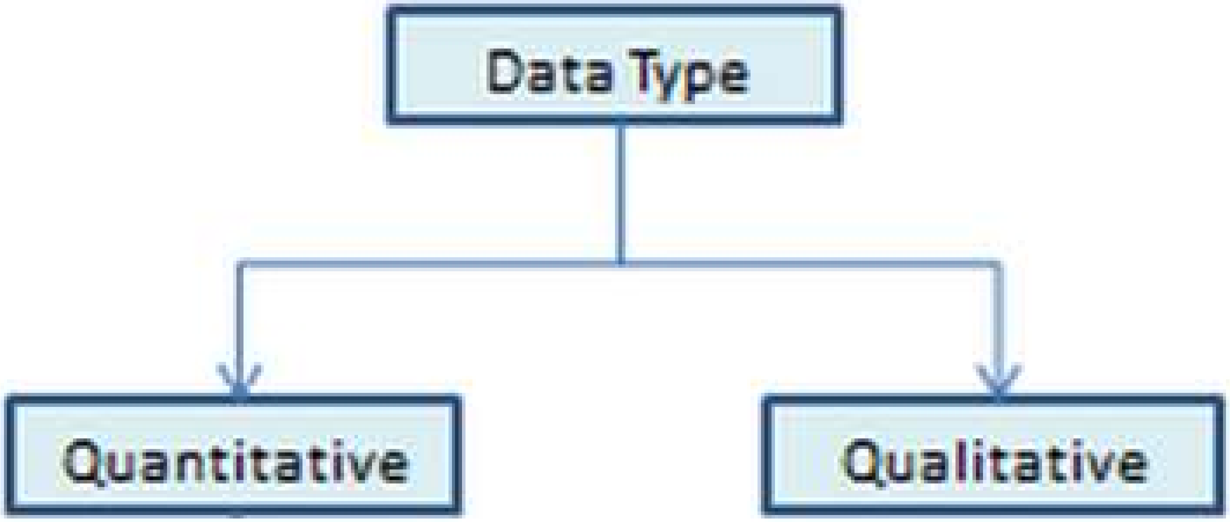
Dr. Mubarak Ali

Classification of Data

Classification of Data

First Classification

- Different methods or tools to analyze different types of data



Quantitative and Qualitative Data

Qualitative Data or Categorical Data: Color of any item, Taste of a coffee, Gender of a student.

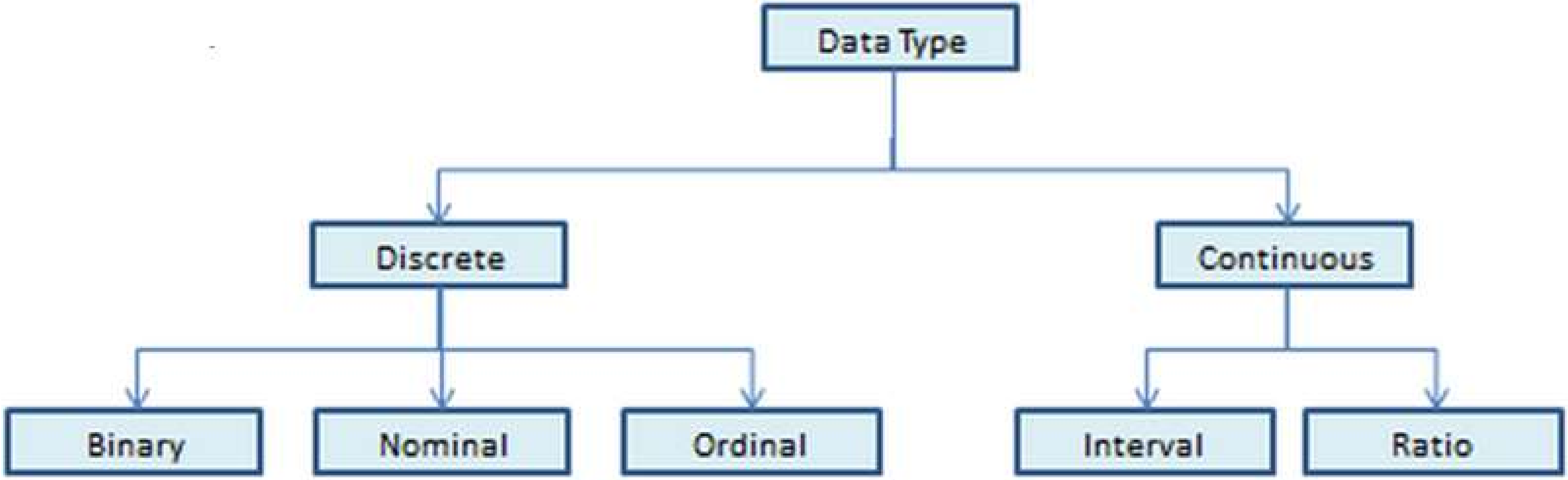
Quantitative data: can be measured

No of items of green color = 4, Cost of the coffee = Rs. 10, Temperature of coffee = 76°C, Weight of a student = 65 kg

Classification of Data

Second Classification

- Different methods or tools to analyze different types of data



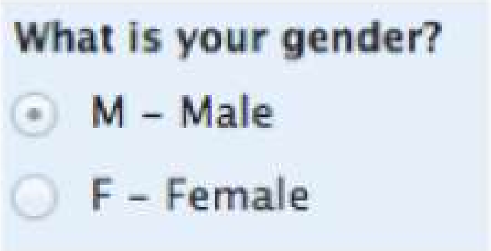
Discrete Data

Discrete data: All values on the real number line are not possible – only certain values are possible. Eg. Grades of subject

Discrete Data is classified to three types

1) **Binary**
2) **Nominal**
3) **Ordinal**

Binary data or Dichotomous data : Only takes on two possible values. Eg. On or Off, True or False, Yes or No.

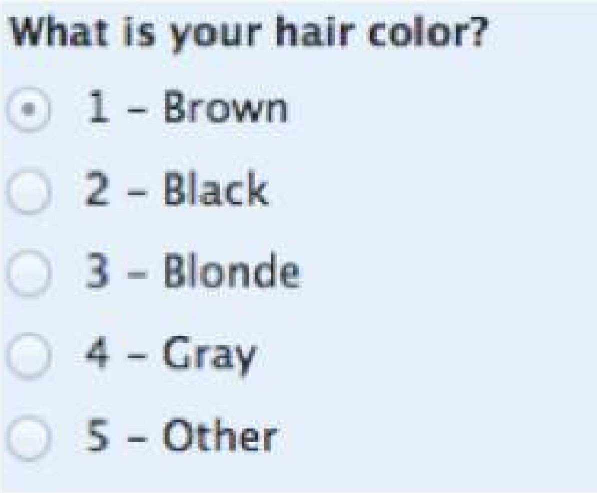


What is your gender?

● M – Male

○ F – Female

Nominal data : Take more than 2 values but these values are not ordered – there is no natural ordering or comparison of these values. Eg. Nationality, Occupation, Religion etc.



What is your hair color?

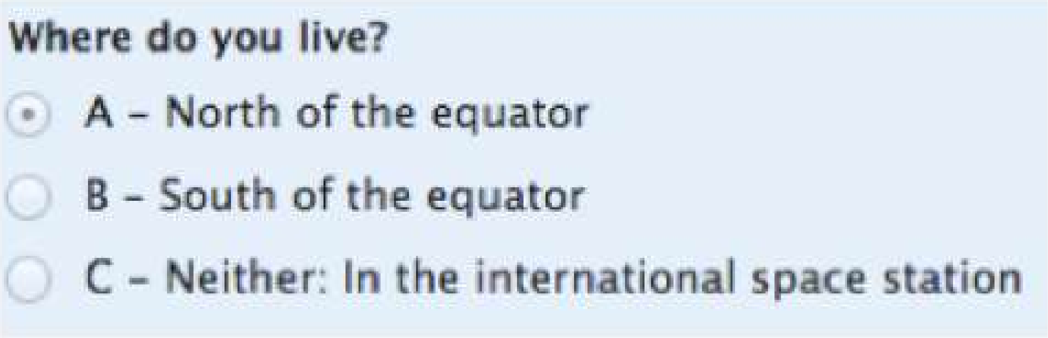
● 1 – Brown

○ 2 – Black

○ 3 – Blonde

○ 4 – Gray

○ 5 – Other



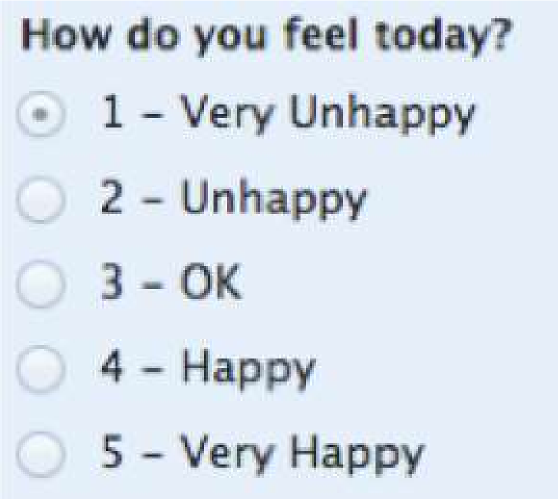
Where do you live?

● A – North of the equator

○ B – South of the equator

○ C – Neither: In the international space station

Ordinal data: Takes more than 2 values but these are naturally ordered. Eg. grades in an exam, results of the running race.



How do you feel today?

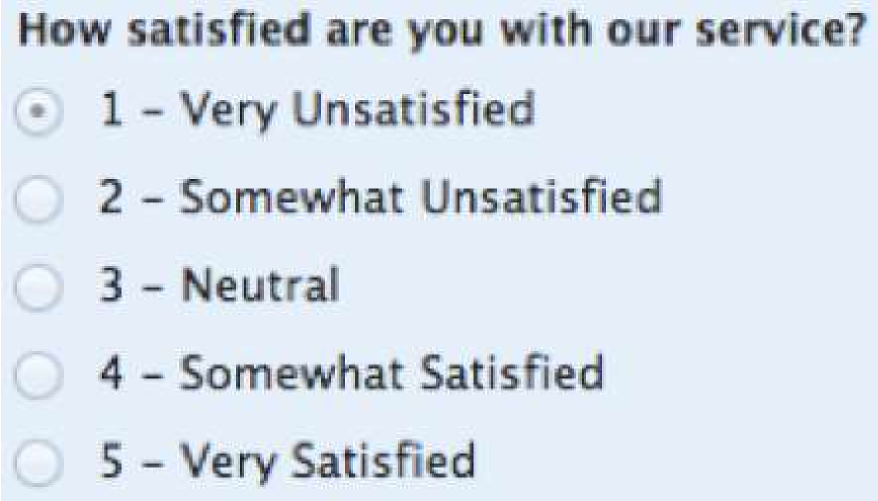
● 1 – Very Unhappy

○ 2 – Unhappy

○ 3 – OK

○ 4 – Happy

○ 5 – Very Happy



How satisfied are you with our service?

● 1 – Very Unsatisfied

○ 2 – Somewhat Unsatisfied

○ 3 – Neutral

○ 4 – Somewhat Satisfied

○ 5 – Very Satisfied

Continuous Data

Continuous data: Any value is theoretically possible (2.7398). Eg. Temperature, Pressure, Humidity, Length

Interval data: Measurement where the difference between two values is meaningful. difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees. (pH, time).

Ratio data: When its variable equals 0.0, there is none of that value. Also has the properties of the interval data. Variables like height, weight etc. you can look at the ratio of two measurements. A weight of 4 grams is twice a weight of 2 grams.

- *A temperature of 100 degrees C is not twice as hot as 50 degrees C.*
- *A pH of 3 is not twice as acidic as a pH of 6.*

Introduction to Business Analytics

Module 2

Dr. Mubarak Ali M

Statistics

- Deals with the collection, presentation, analysis and use of **data**
- To make decisions, solve problems,
- Design products and improving existing products
- and designing, developing improving production processes or business processes

Statistical Thinking

- All of us do statistical thinking everyday
- Example:

  Fuel Mileage of your bike

  Time taken to reach your institute

  Expected arrival time of a train

Data

Data is any factual information or measurement that is collected and used for making a decision, reasoning, or any calculation

Definitions

Population: Set of all units of interest

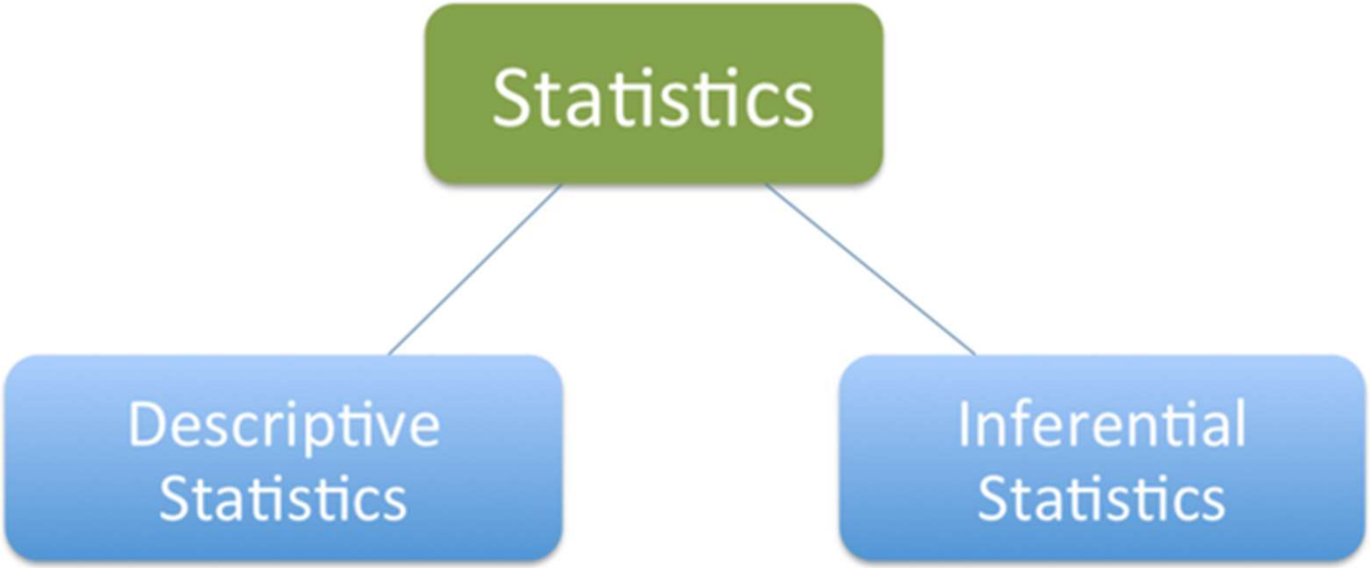
Sample: Subset of population

Random sample: Sample collected in such a way that every unit in population is equally likely to be selected to ensure good overall representation of population

Random sampling: The activity/procedure to extract random samples from a population.

Statistical Inference: Estimate key parameters of a population based on a sample drawn from the population.

Categories of Statistics



Statistics

Descriptive Statistics

Inferential Statistics

Presenting, organizing and summarizing data

Tabular Methods
Graphical Methods
Numerical Methods

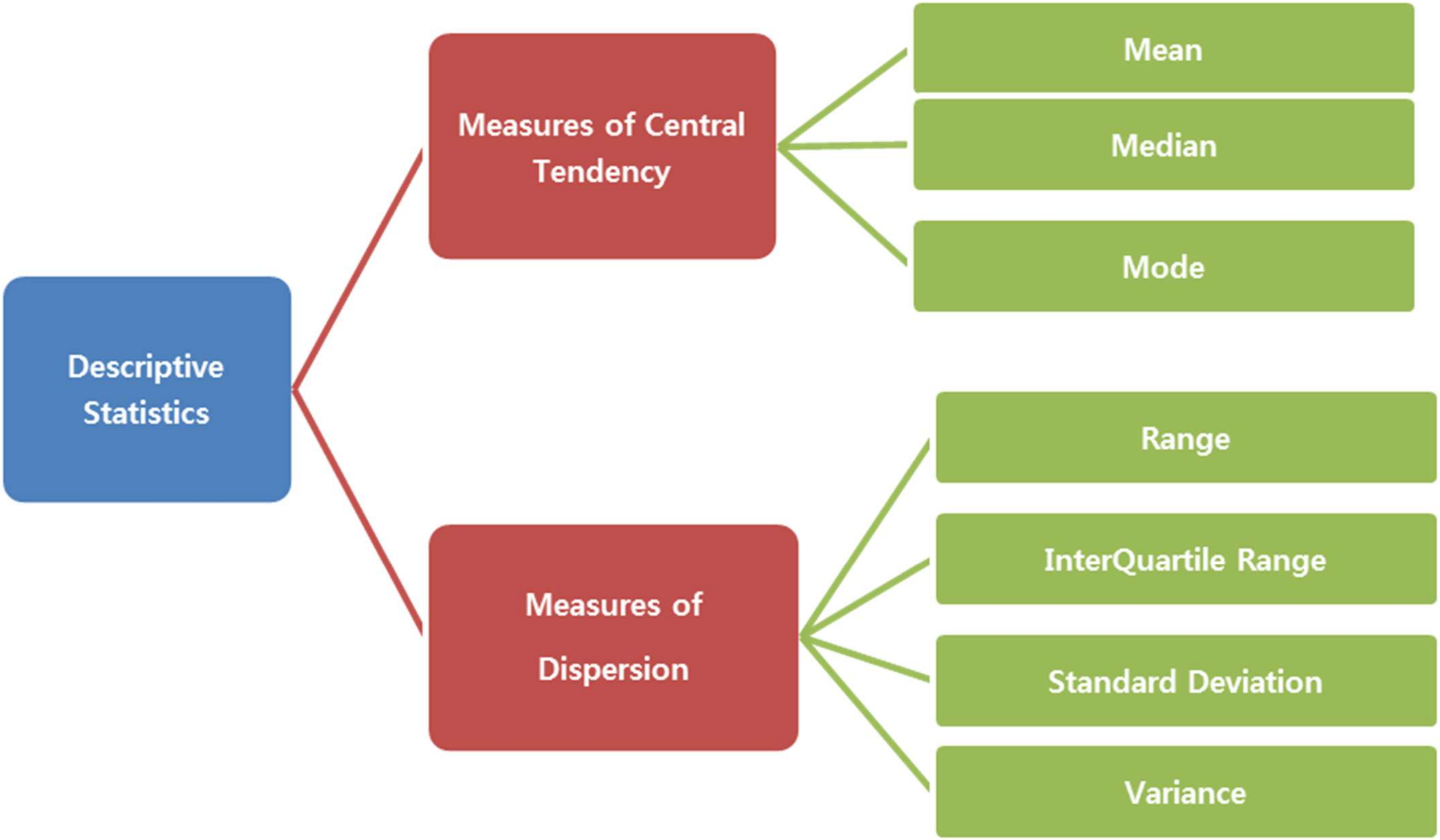
Drawing conclusions about a population based on data observed in a sample

Confidence Intervals,
Significance Tests

Descriptive Analytics

Numerical methods

Numerical Methods of Descriptive Analytics



Measures of Central Tendency

1. Mean: It is the arithmetic average of all values:

Mean = Sum of all values /Total number of values

2. Median: It is the central value of the data points, when arranged in ascending or descending order

With an odd number of observations, the median is the middle value. An even number of observations has no single middle value so in this case, we take average of the values for the middle two observations

3. Mode: It is the most frequently occurring value .

Based on the nature of data, any one of these is used.

Measures of Dispersion

1. Range: It represents the gap between the highest and lowest value in the group.

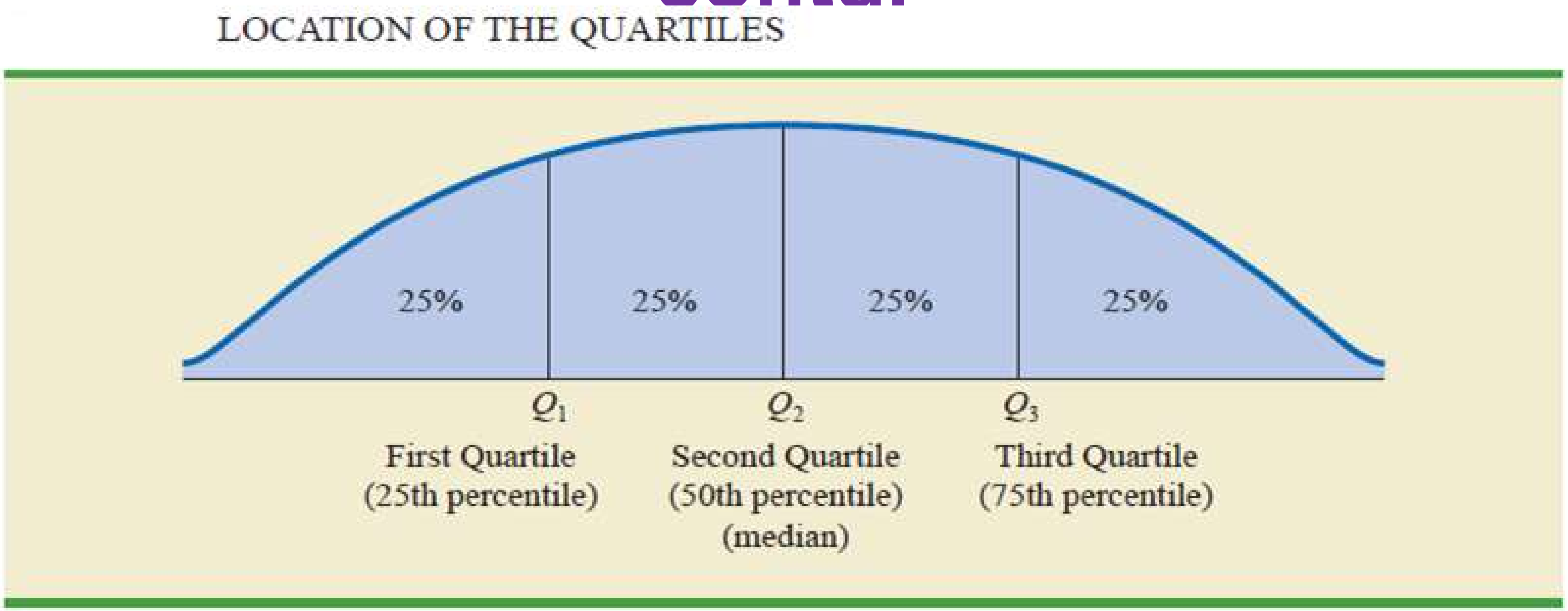
Range = Maximum value – Minimum Value

2. InterQuartile Range:

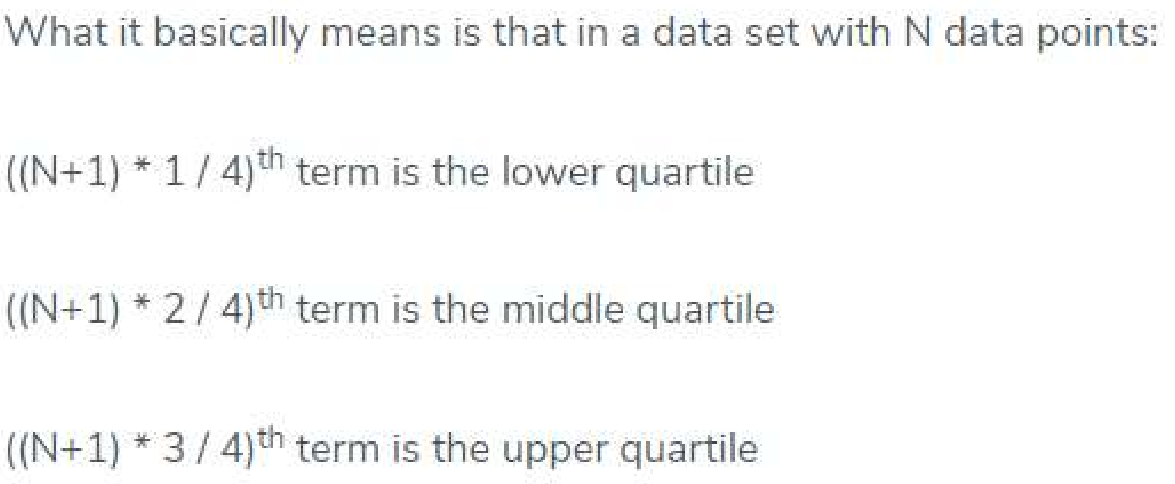
A **percentile** provides information about how the data are spread over the interval from the smallest value to the largest value

Quartiles are just specific percentiles; thus, the steps for computing percentiles can be applied directly in the computation of quartiles.

contd.


LOCATION OF THE QUARTILES

InterQuartile Range (IQR):-A measure of variability that overcomes the dependency on extreme values

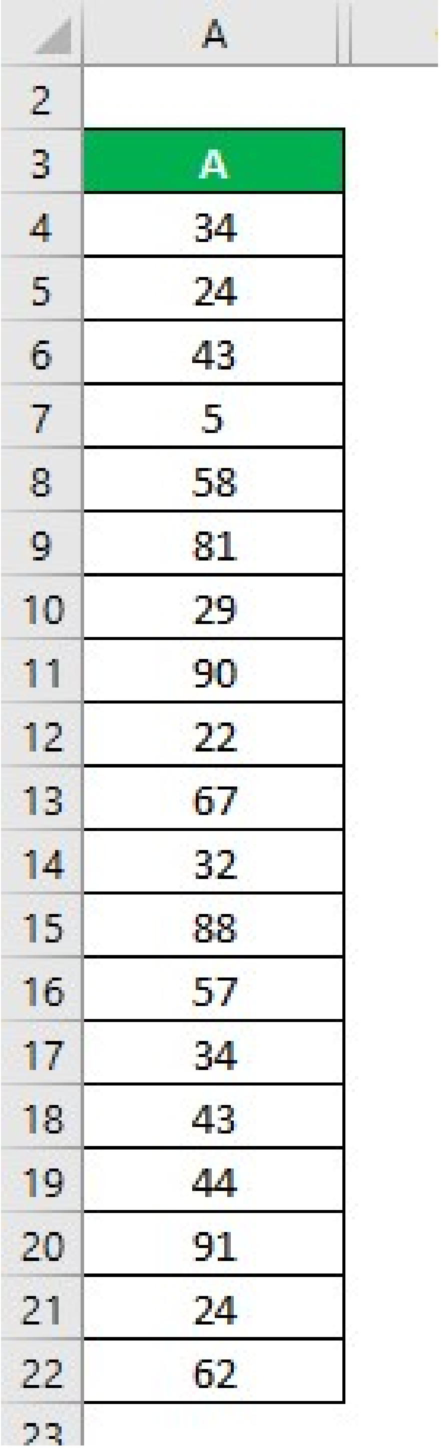

What it basically means is that in a data set with N data points:

((N+1) * 1 / 4)th term is the lower quartile

((N+1) * 2 / 4)th term is the middle quartile

((N+1) * 3 / 4)th term is the upper quartile

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$$



	A
2	
3	A
4	34
5	24
6	43
7	5
8	58
9	81
10	29
11	90
12	22
13	67
14	32
15	88
16	57
17	34
18	43
19	44
20	91
21	24
22	62
23	

Contd.

Find IQR of the Data Set A

First arrange the data in ascending order.

Lower Quartile (Q1) = 29

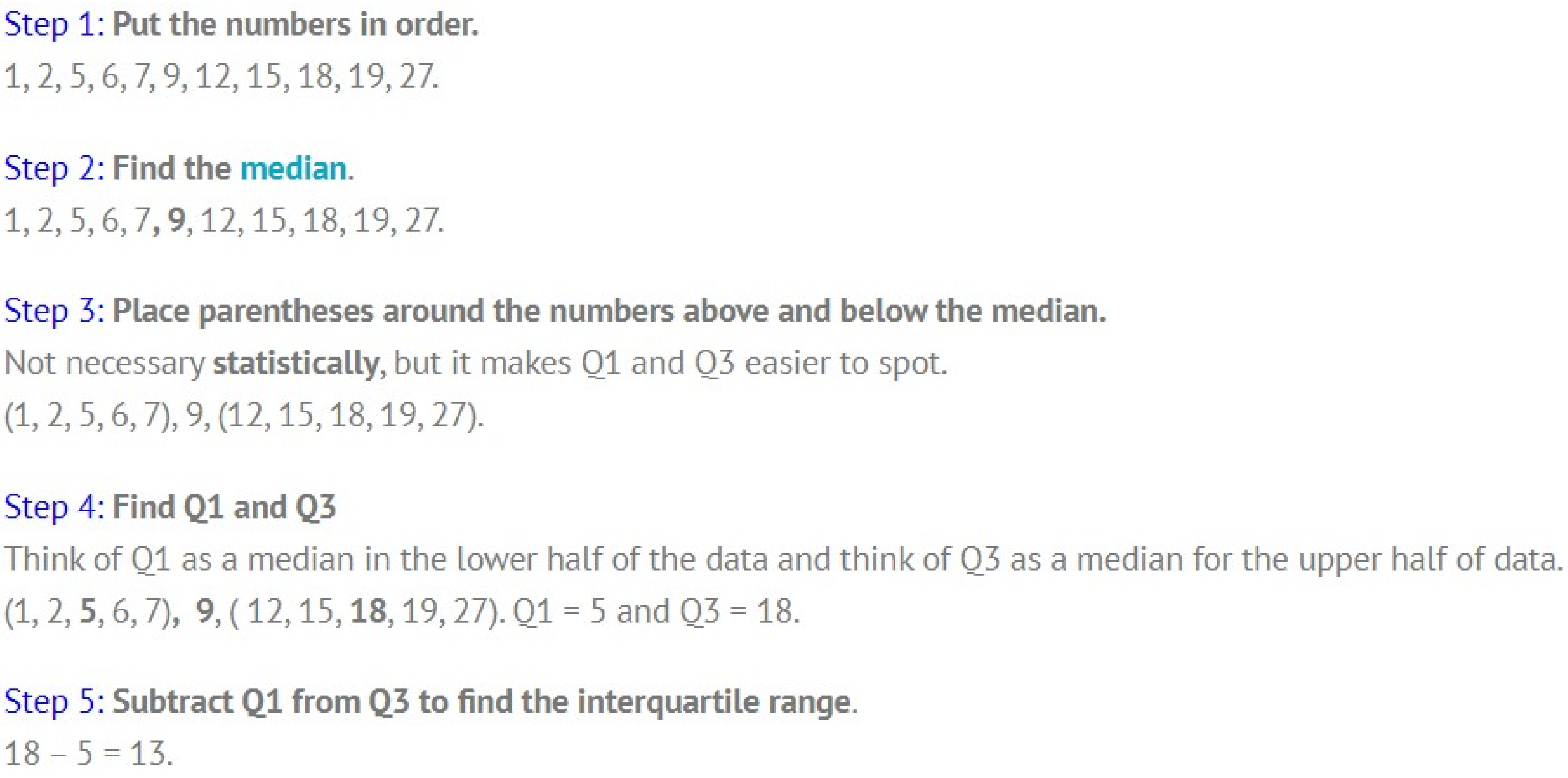
Middle Quartile (Q2) = 43

Upper Quartile (Q3)= 67

Interquartile Range = 67-29 = 38

Alternate method of finding IQR

If you have ODD set of numbers



Step 1: **Put the numbers in order.**

1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.

Step 2: **Find the median.**

1, 2, 5, 6, 7**, 9**, 12, 15, 18, 19, 27.

Step 3: **Place parentheses around the numbers above and below the median.**

Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.

(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).

Step 4: **Find Q1 and Q3**

Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.

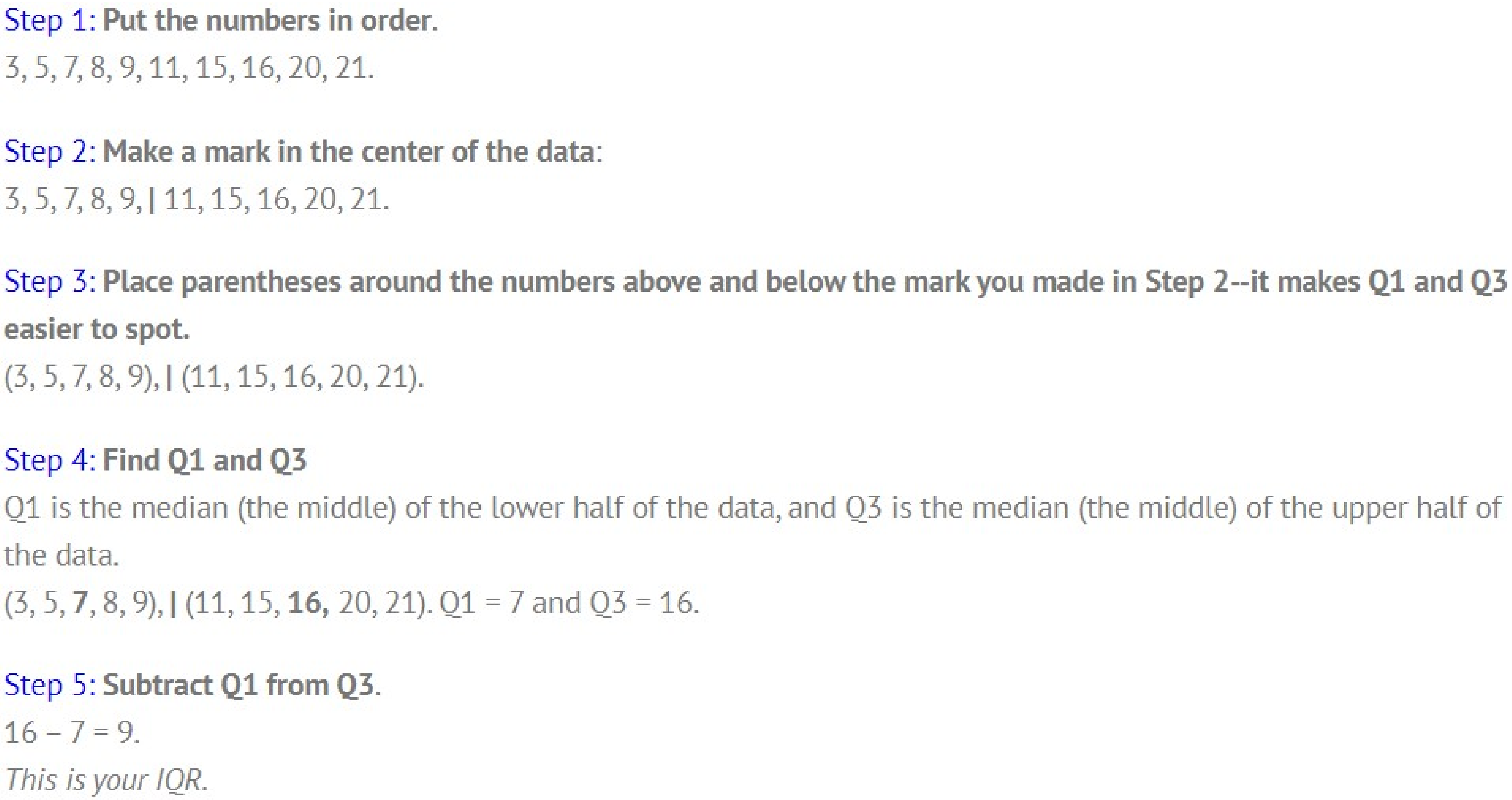
(1, 2, **5**, 6, 7)**,** **9**, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.

Step 5: **Subtract Q1 from Q3 to find the interquartile range.**

18 – 5 = 13.

Alternate method of finding IQR

If you have EVEN set of numbers



Step 1: **Put the numbers in order.**

3, 5, 7, 8, 9, 11, 15, 16, 20, 21.

Step 2: **Make a mark in the center of the data:**

3, 5, 7, 8, 9, **|** 11, 15, 16, 20, 21.

Step 3: **Place parentheses around the numbers above and below the mark you made in Step 2--it makes Q1 and Q3 easier to spot.**

(3, 5, 7, 8, 9), **|** (11, 15, 16, 20, 21).

Step 4: **Find Q1 and Q3**

Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.

(3, 5, **7**, 8, 9), **|** (11, 15, **16,** 20, 21). Q1 = 7 and Q3 = 16.

Step 5: **Subtract Q1 from Q3.**

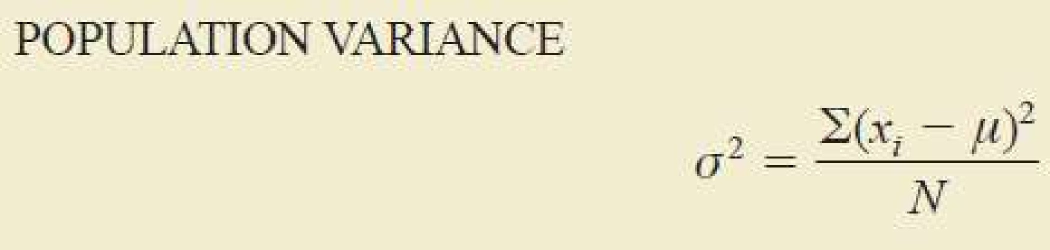
16 – 7 = 9.

This is your IQR.

Variance

3. Variance:- The **variance** is a measure of variability that utilizes all the data.

- The difference between each x_i and the mean μ is called a *deviation about the mean.*
- If the data are for a population, the average of the squared deviations is called the *population variance*.

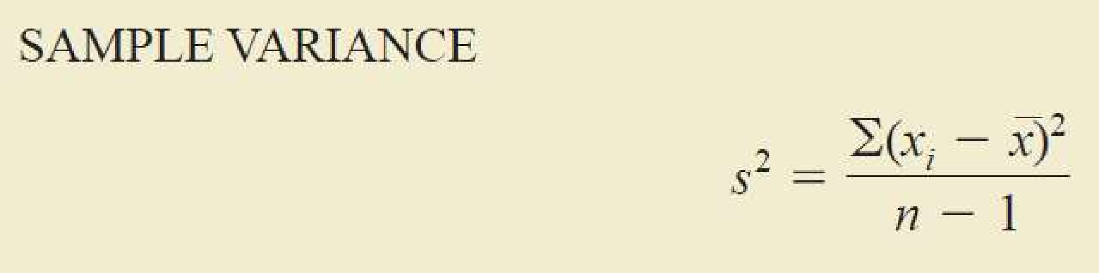


POPULATION VARIANCE

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{SS}{DOF}$$

If sum of the squared deviations about the sample mean is divided by n-1, the resulting sample variance provides an unbiased estimate of the population variance

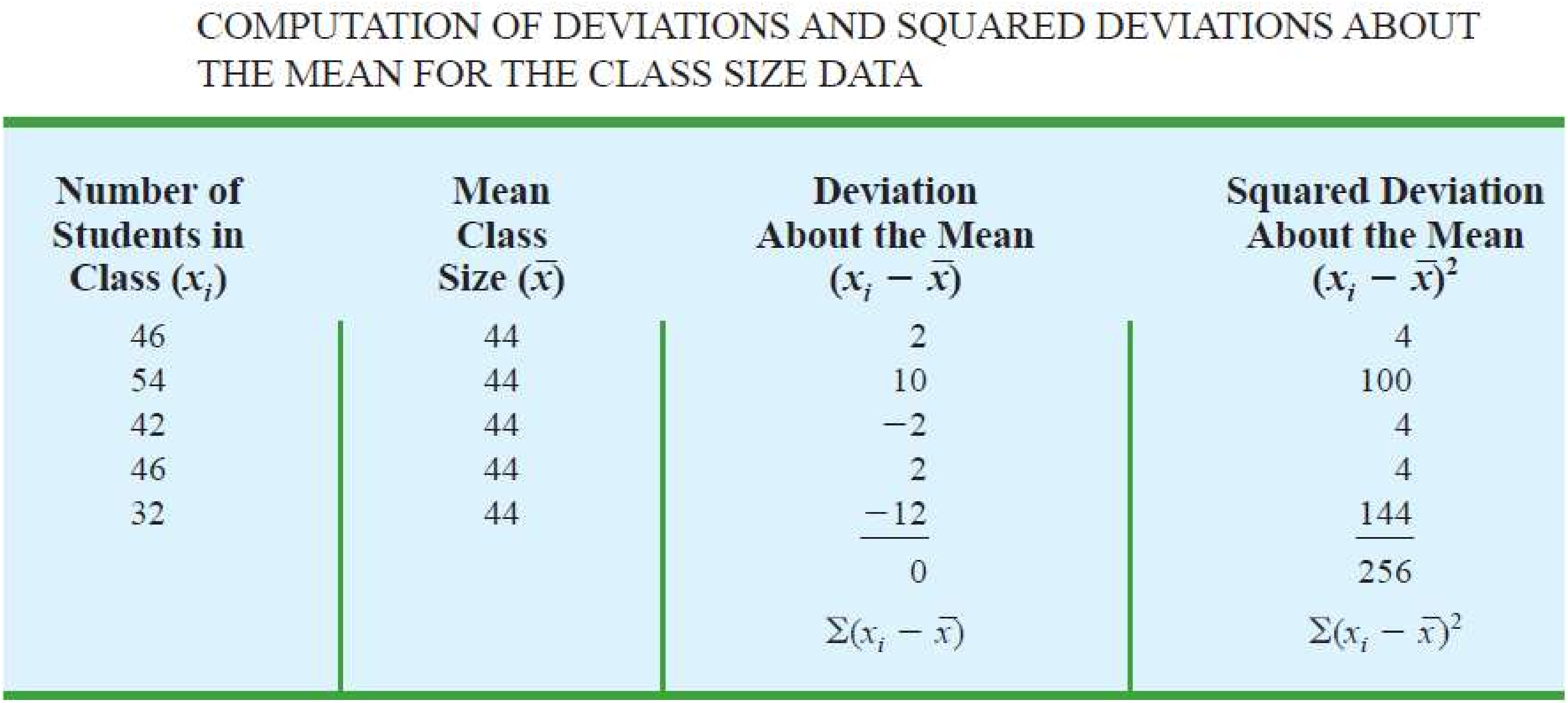


SAMPLE VARIANCE

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

$$s^2 = \frac{SS}{DOF}$$

Calculation of Variance of Data Set



COMPUTATION OF DEVIATIONS AND SQUARED DEVIATIONS ABOUT THE MEAN FOR THE CLASS SIZE DATA

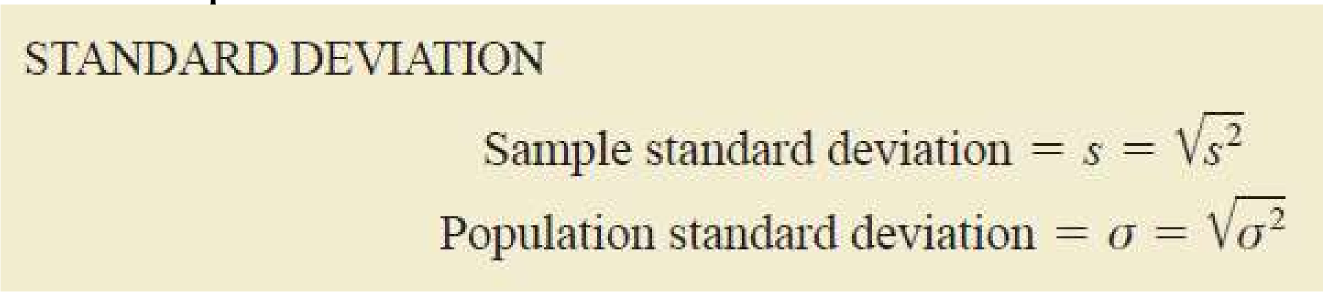
Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

The sum of squared deviations about the mean is (*xi*-*x*)^2 =256. Hence, with *n*-1=4, the sample variance is

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

SD and COV

4. Standard Deviation:- The **standard deviation** is defined to be the positive square root of the variance.



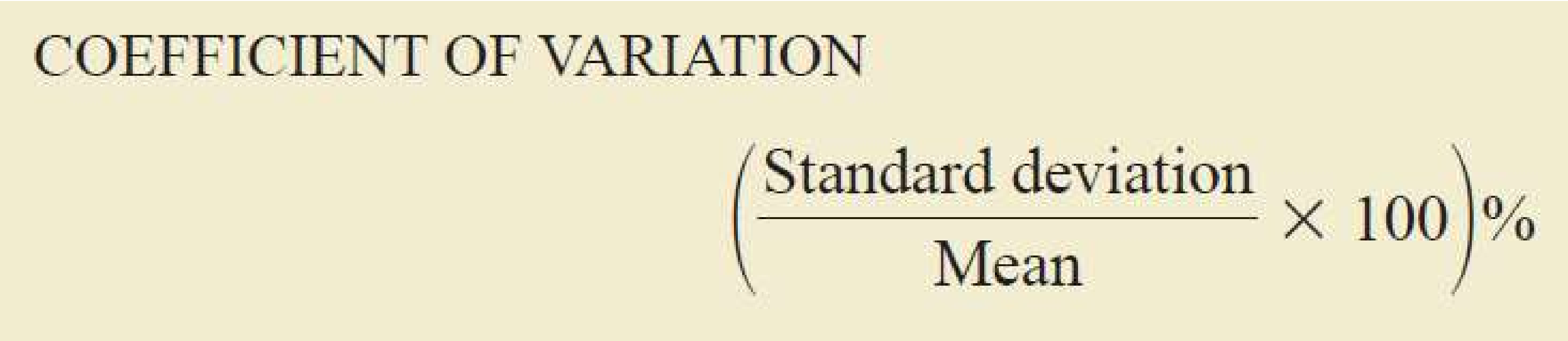
STANDARD DEVIATION

$$\text{Sample standard deviation} = s = \sqrt{s^2}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

Hence the Sample Standard Deviation=8, for the latest example

5. Coefficient of Variation:- In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient of variation** and is usually expressed as a percentage.



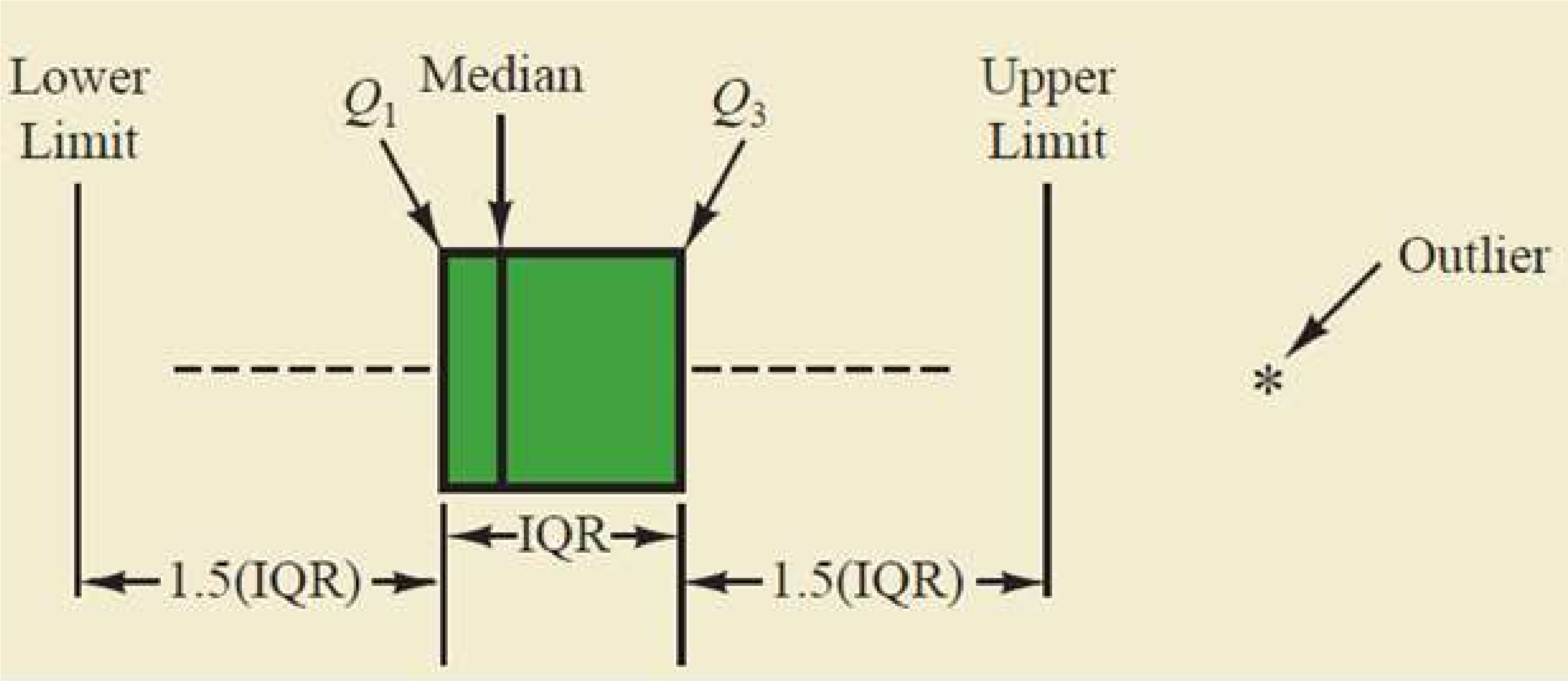
COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

Five Number Summary and Box plot

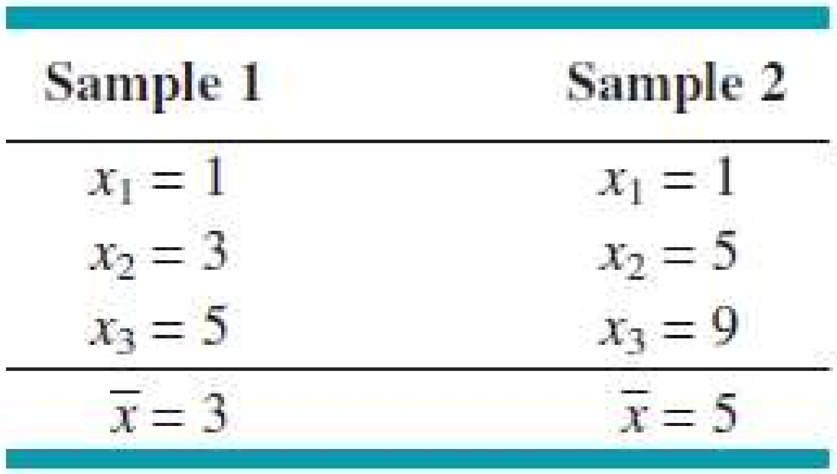
6. Five Number Summary- The following five numbers are used to summarize the data: 1.Smallest value 2.First quartile ($Q1$) 3.Median ($Q2$) 4. Third quartile ($Q3$) 5. Largest value

7. Box Plot:- Graphical display that simultaneously displays several important features of the data, such as central tendency, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data ("outliers").



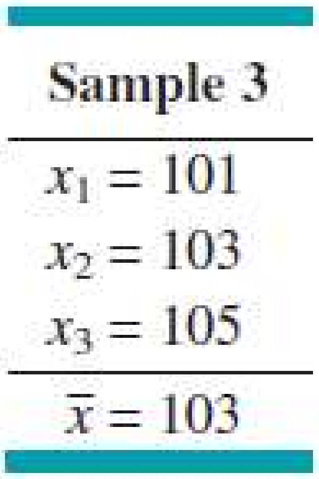
Standard Deviation - Comparision

- For class size data, sample mean =44 and a sample standard deviation =8. The coefficient of variation is [(8/44)x100]% =18.2%. In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean.



Sample 1	Sample 2
$x_1 = 1$	$x_1 = 1$
$x_2 = 3$	$x_2 = 5$
$x_3 = 5$	$x_3 = 9$
$\bar{x} = 3$	$\bar{x} = 5$

s = 2 s = 4



Sample 3
$x_1 = 101$
$x_2 = 103$
$x_3 = 105$
$\bar{x} = 103$

s = 2

Sample 2 has greater variability than Sample 1

Standard Deviation only reflects scatter about the average