CHAPTER 1
DESCRIPTIVE STATISTICS

- **L1 – Basic of data measurements – mean, variance and standard deviation**

Learning Objectives:

At the end of the lesson, students should be able to:

- Explain the concepts of
  - sample mean, population mean,
  - sample variance, population variance, sample standard deviation,

Compute and interpret the sample mean, sample variance, sample standard deviation, sample median, an sample range

Population – Sample (Definition)

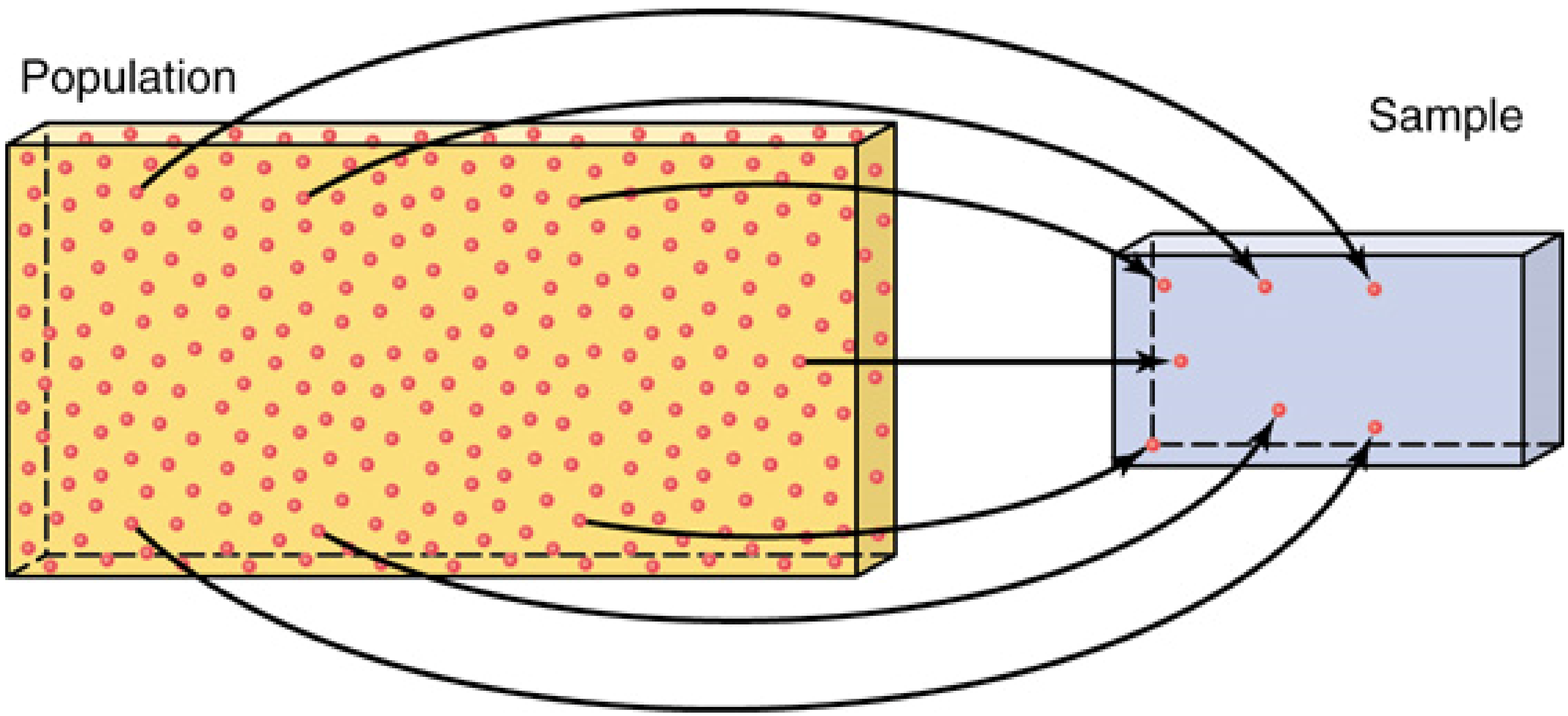
Population:

- A collection, or set, of individuals or objects or events whose properties are to be analyzed.
  (the number UTP students)

Sample:

- A subset of the population. The number of individuals of a sample is called the sample size.
  (the number of engineering students in UTP)

Illustration of selection of a sample from a population



Population - Sample

Variable:

- A characteristic of the objects in a population.
  - CGPA of UTP students (number)
  - Gender of an engineering graduate (category: male or female)
- Its value may change from one object to another in the population

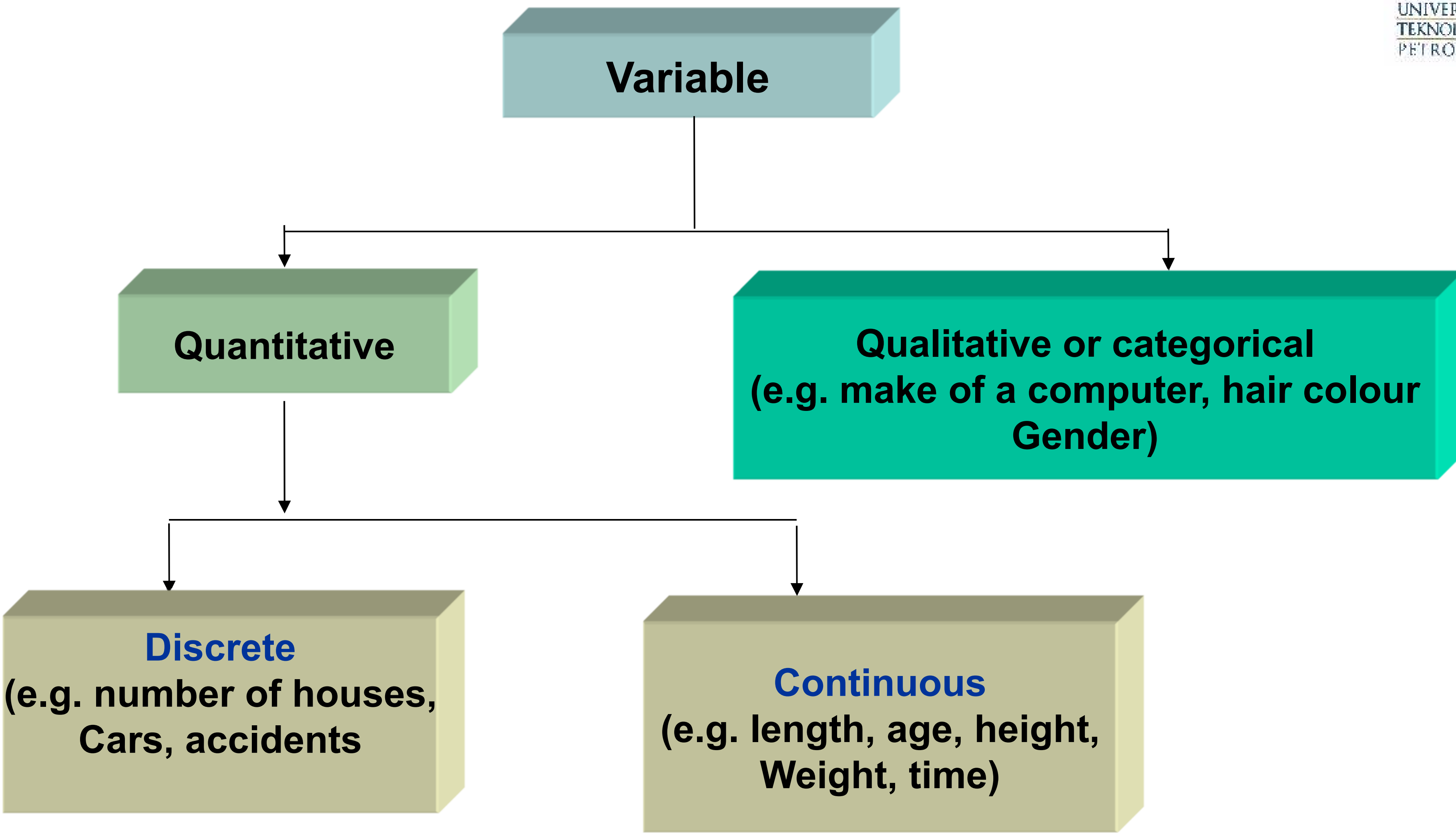
Univariate:

- A data set consists of observations on a single variable. (type of transmission in a car, automatic or manual)

Multivariate:

- A data set arises when observations made on more than one variable (height and weight)

Types of variables



Statistics

Descriptive Statistics

-Methods of organizing, display, and describe important features of data by

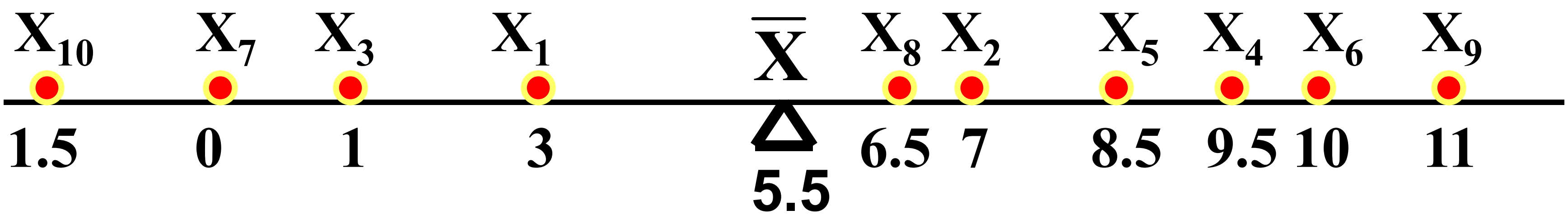
* tables,
* graphs, and
* summary measures

Inferential Statistics

-Methods that use sample results to help make decisions (inferences) or predictions about a population

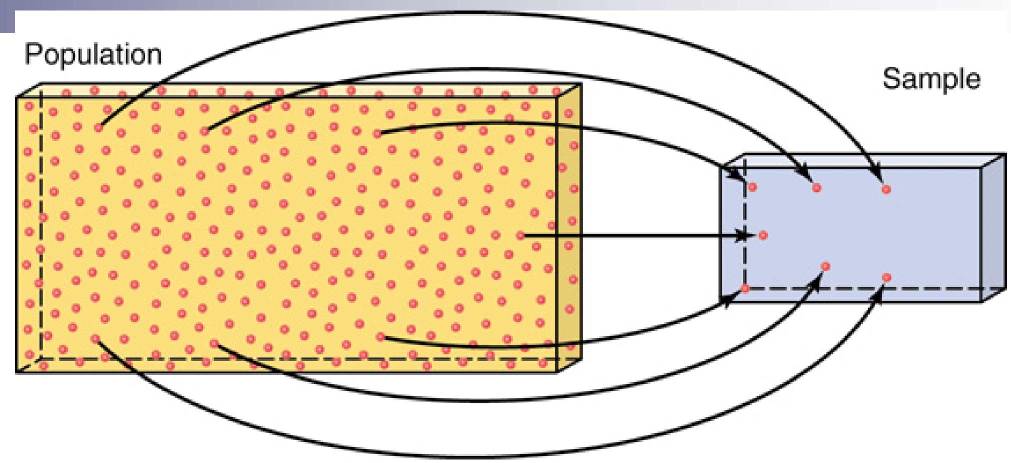
Numerical Summary : Mean

- **The mean is the balance point for a system of unit weights at points $x_1, x_2, \ldots, x_n$**



$$\sum x = x_1 + x_2 + \cdots + x_{10} = 55; \qquad \bar{x} = \frac{55}{10} = 5.5$$

MEAN



Population mean (mu) :

Sum of all values
In the population

$$\boxed{\mu = \frac{\sum x}{N}}$$

The population size

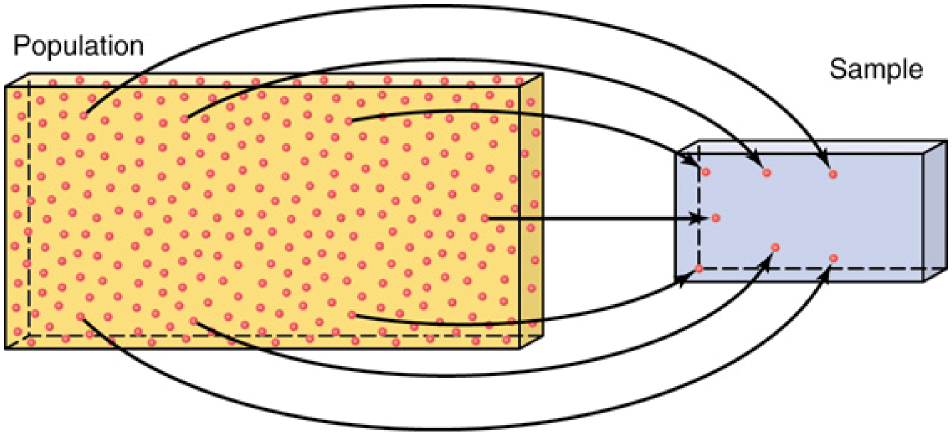
Sample mean

Sum of all values
In the sample

$$\boxed{\bar{x} = \frac{\sum x}{\mathbf{n}}}$$

The sample size

Numerical Summary : Variability



Population variance :

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

Population standard deviation is σ

Numerical Summary : Variability

- **Sample Variance**

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{1}{n-1}S_{XX};\ S_{XX} = \sum(x-\bar{x})^2$$

$$S_{xx} = \sum x^2 - n(\bar{x})^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$\boxed{s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

Sample Standard Deviation: SD = s

Exercise 1: (Example 4.1)

Find the mean, variance and standard deviation for the following observations:

55 68 90 42 89 70

$$\bar{x} = \frac{\sum x}{\mathbf{n}} = \frac{55+68+90+42+89+70}{6} = \frac{414}{6} = 69$$

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{30334 - \frac{(414)^2}{6}}{5} = 353.6$$

$$s = 18.804$$

Exercise 3: (L1)

Seven oxide thickness measurements of wafers are studied to assess quality in a semiconductor manufacturing process. The data (in angstroms) are: 1264, 1280, 1301, 1300, 1292, 1307, and 1275. Calculate the sample average, variance and standard deviation.

CHAPTER 1
DESCRIPTIVE STATISTICS

- **L2 - Graphical display of Data**

Learning Objectives:

At the end of the lesson, students should be able to:

- Construct and interpret pictorial and tabular display of data

Pictorial & Tabular Methods

1. Stem-and-Leaf Displays:

How to construct a Stem-and-Leaf Display:

1. Each numerical data is divided into two parts:
   - The leading digit(s) becomes the stem, and the remaining digit(s) becomes the leaf

2. List the stem values in a vertical column.

3. Record the leaf for each observation beside its stem.

4. Write the units for stems and leaves on the display.

Stem & Leaf Display

- Result of Math. Exam. of a 50-student class:

35	42	56	41	63
26	37	66	92	16
49	28	56	64	72
59	17	45	56	29
30	45	39	37	43
76	73	64	51	60
40	52	57	65	83
68	52	84	91	64
45	76	56	90	73
34	26	57	41	56

- **Stem-and-Leaf Display**

Stem	Leaf
1	6 7
2	6 6 8 9
3	0 4 5 7 7 9
4	0 1 1 2 3 5 5 5 9
5	1 2 2 6 6 6 6 7 7 9
6	0 3 4 4 4 5 6 8
7	2 3 3 6 6
8	3 4 6
9	0 1 2

Stem: tens digit

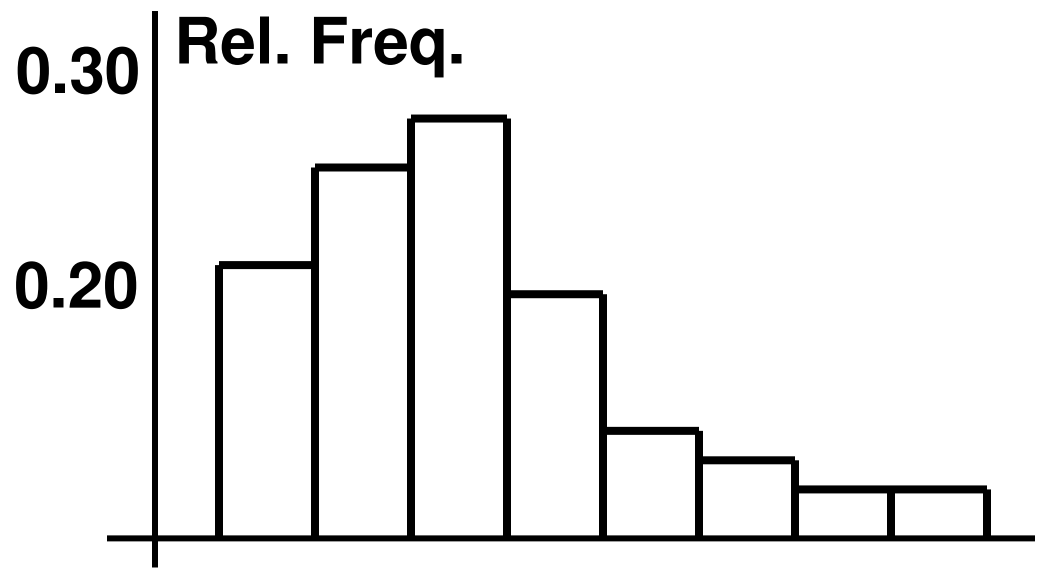
Leaf: ones digit

2. Histogram:

A bar graph representing a frequency distribution of a quantitative variable.A histogram is made up of the following components. Histograms are used to summarize large data sets.

Age	Freq.	Rel. Freq.
18	20	0.20
19	24	0.24
20	26	0.26
21	18	0.18
22	5	0.05
23	3	0.03
24	2	0.02
25	2	0.02
Sum	100	1.00

Histogram: ages of 100 students



3. Box plot:

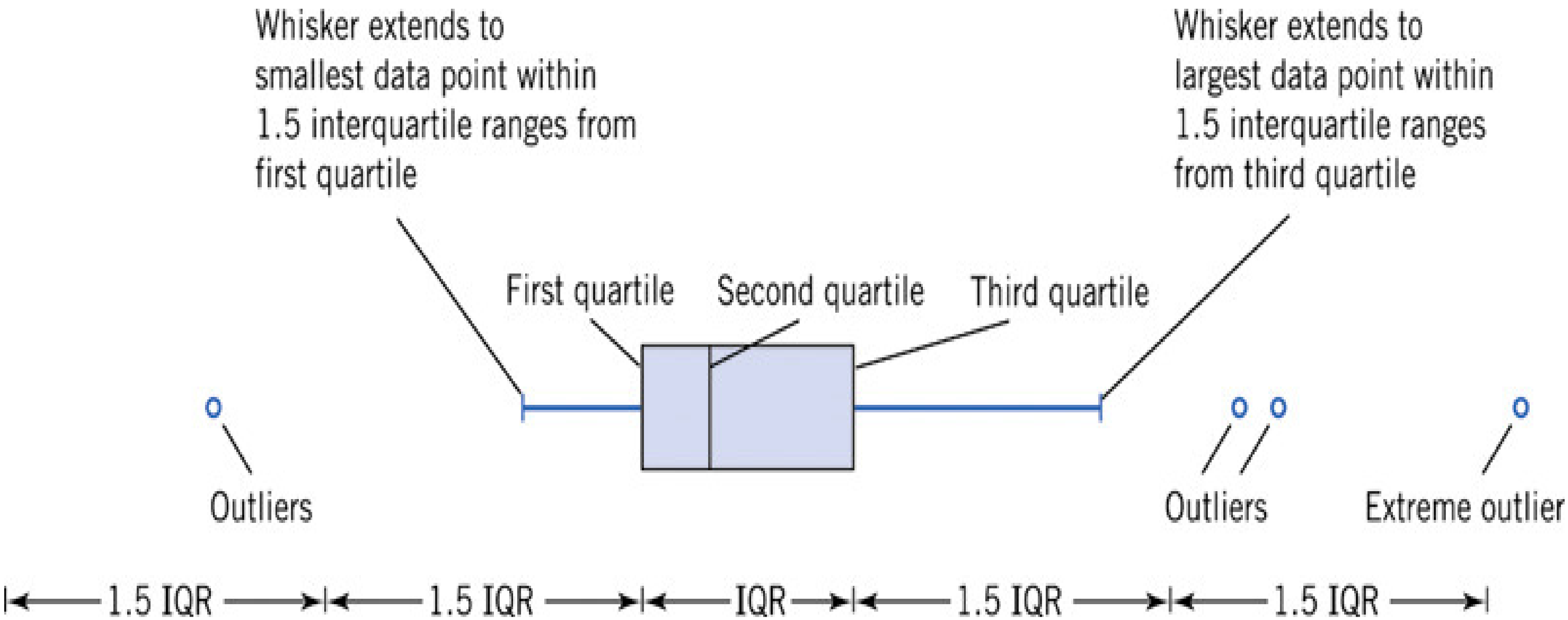
a graphical display that simultaneously describes several important features of a data set:

- center
- Spread
- departure from symmetry
- identification of outliers

a box plot displays the median, the first quartile and the third quartiles on a rectangular box, aligned either horizontally or vertically.

sometimes called box whiskers plot.

HOW TO CONSTRUCT A BOX PLOT



Numerical Summary : Sample Median

The median of a sample depends on whether the number of terms in the sample is even or odd.

- If the number of terms is odd, then the median is the value of the term in the middle.
- If the number of terms is even, then the median is the average of the two terms in the middle

➢ **Arrange the observations $x_1, \ldots, x_n$ in increasing order:**

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

Use the following rule:

$$\boxed{\tilde{x} = \begin{cases} \dfrac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd.} \end{cases}}$$

Numerical Summary : Sample Median

- **Example 1: Find Median for the following observations:**

0.3 7.8 4.6 3.7 9.2 12.1 -5 -2.5 10.8

Numerical Summary : Sample Median

- **Example 1: Find Median for the following observations:**

0.3 7.8 4.6 3.7 9.2 12.1 -5 -2.5 10.8

Arrange the observations in increasing order: n = 9

- 5 -2.5 0.3 3.7 4.6 7.8 9.2 10.8 12.1

$$\tilde{x} = \begin{cases} \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd.} \end{cases}$$

- **Example 2: Find Median for given observations :**

2.8 5.2 -2.3 2.6 3.6 1.4 6.9 4.3 8.4 2.8

➢ Example 2: Find Median for given observations :

2.8 5.2 -2.3 2.6 3.6 1.4 6.9 4.3 8.4 2.8

Rearrange the observations in increasing order:

- 2.3 1.4 2.6 2.8 $\underline{2.8 \quad 3.6}$ 4.3 5.2 6.9 8.4

Median = (2.8 + 3.6)/2 = 3.2

$$\boxed{\tilde{x} = \begin{cases} \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd.} \end{cases}}$$

LOWER QUARTILE, UPPER QUARTILE, INTERQUARTILE RANGE

UNIVERSITI TEKNOLOGI PETRONAS

Percentile:

Measure of central tendency that divide a group of data into 100 parts.

Nth percentile:

At least n% of the data lie between the nth percentile and at most (100-n)% of the data lie above the nth percentile

90 percentile:

At least 90% of the data lie between the 90th percentile and at most (10)% of the data lie above the 90th percentile

LOWER QUARTILE, UPPER QUARTILE, INTERQUARTILE RANGE

- LQ (Q_1) is 25 percentile
- Median (Q_2) is 50 percentile
- UQ (Q_3) is 75 percentile

25 percentile = Q_1

At least 25% of the data lie between the 25th percentile and at most (75)% of the data lie above the 25th percentile

LOWER QUARTILE, UPPER QUARTILE, INTERQUARTILE RANGE

- LQ (Q_1) and UQ (Q_3) are defined as follows

Step 1. Arrange the values in increasing order

Step 2. Q_1 is the value in position 0.25(n+1)

Q_3 is the value in position 0.75(n+1)

Step 3. If the positions are not integers, Q_1 and Q_3 are found by *interpolation*, using adjacent values

- IQR = $Q_3 - Q_1$

LOWER QUARTILE, UPPER QUARTILE, INTERQUARTILE RANGE

- **Example 1: (values are arranged in increasing order)**

- 5   -2.5   0.4   3.7   4.6   7.8   9.2   10.8   12.1   13.5   14

$n = 11$, $0.25(n+1) = 0.25(12) = 3$;

$0.75(n+1) = 0.75(12) = 9$

$Q_1 = x_{(3)} = 0.4$,

$Q_3 = x_{(9)} = 12.1$,

and IQR $= 12.1 – 0.4 = 11.7$

LOWER QUARTILE, UPPER QUARTILE, INTERQUARTILE RANGE

- **Example 2: (values are arranged in increasing order)**

- 5 - 4 2 6 6.5 7.8 9.2 10.8 12.5 14.5 15 16.4

n=12,

$0.25(n+1) = 0.25(13) = 3.25$; $0.75(n+1) = 0.75(13) = 9.75$

$$Q_1 = x_{(3)} + 0.25(x_{(4)} - x_{(3)}) = 2 + 0.25(6 - 2) = 2 + 0.25(4) = 3$$

$$Q_3 = x_{(9)} + 0.75(x_{(10)} - x_{(9)}) = 12.5 + 0.75(14.5 - 12.5) = 14$$

LOWER QUARTILE, UPPER QUARTILE, INTERQUARTILE RANGE

- **Example 3**: (values are arranged in increasing order)

2   5   9   9.8   10.2   10.8   12.5   14   16.4   18.7

n=10,

0.25(n+1) = 0.25(11)=2.75;   0.75(n+1)=0.75(11) = 8.25

$Q_1 = x_{(2)} + 0.75(x_{(3)} - x_{(2)}) = 5 + 0.75(9 - 5) = 5 + 0.75(4) = 8$

$Q_3 = x_{(8)} + 0.25(x_{(9)} - x_{(8)}) = 14 + 0.25(16.4 - 14) = 14.6$

Example 4 :

The following " cold start ignition time" of an automobile engine obtained for a test vehicle are as follows:

1.75	1.92	2.62	2.35	3.09	3.15	2.53	1.91

a) Calculate the sample median, the quartiles and the IQR

b) Construct a box plot of the data.

Example 4:

The following " cold start ignition time" of an automobile engine obtained for a test vehicle are as follows:

1.75 1.92 2.62 2.35 3.09 3.15 2.53 1.91

a) Calculate the sample median, the quartiles and the IQR

b) Construct a box plot of the data.

Solution:

Rank the n = 8 measurements from smallest to largest

1.75 1.91 1.92 2.35 2.53 2.62 3.09 3.15

sample median: since n is even $\tilde{x} = \frac{1}{2}(x_{(n/2)} + x_{(n/2\ +1)})$

$$\Rightarrow \tilde{x} = \frac{1}{2}(x_{(4)} + x_{(5)}) = \frac{1}{2}(2.35 + 2.53) = 2.44$$

Solution:

1.75 1.91 1.92 2.35 2.53 2.62 3.09 3.15

Lower quartile: $Q_1 = x_{(0.25(n+1))} = x_{(0.25(8+1))} = x_{(2.25)}$

$$Q_1 = x_{(2)} + 0.25(x_3 - x_2) = 1.91 + 0.25(1.92 - 1.91) = 1.913$$

Upper quartile: $Q_3 = x_{(0.75(n+1))} = x_{(0.75(8+1))} = x_{(6.75)}$

$$Q_3 = x_{(6)} + 0.75(x_7 - x_6) = 2.62 + 0.75(3.09 - 2.62) = 2.973$$

IQR:

$$Q_3 - Q_1 = 2.973 - 1.913 = 1.06$$

b) Construct a box plot of the data.