3. DESCRIPTIVE STATISTICS [IT2110]

By Department of Mathematics and Statistics
Faculty of Humanities and Sciences, SLIIT

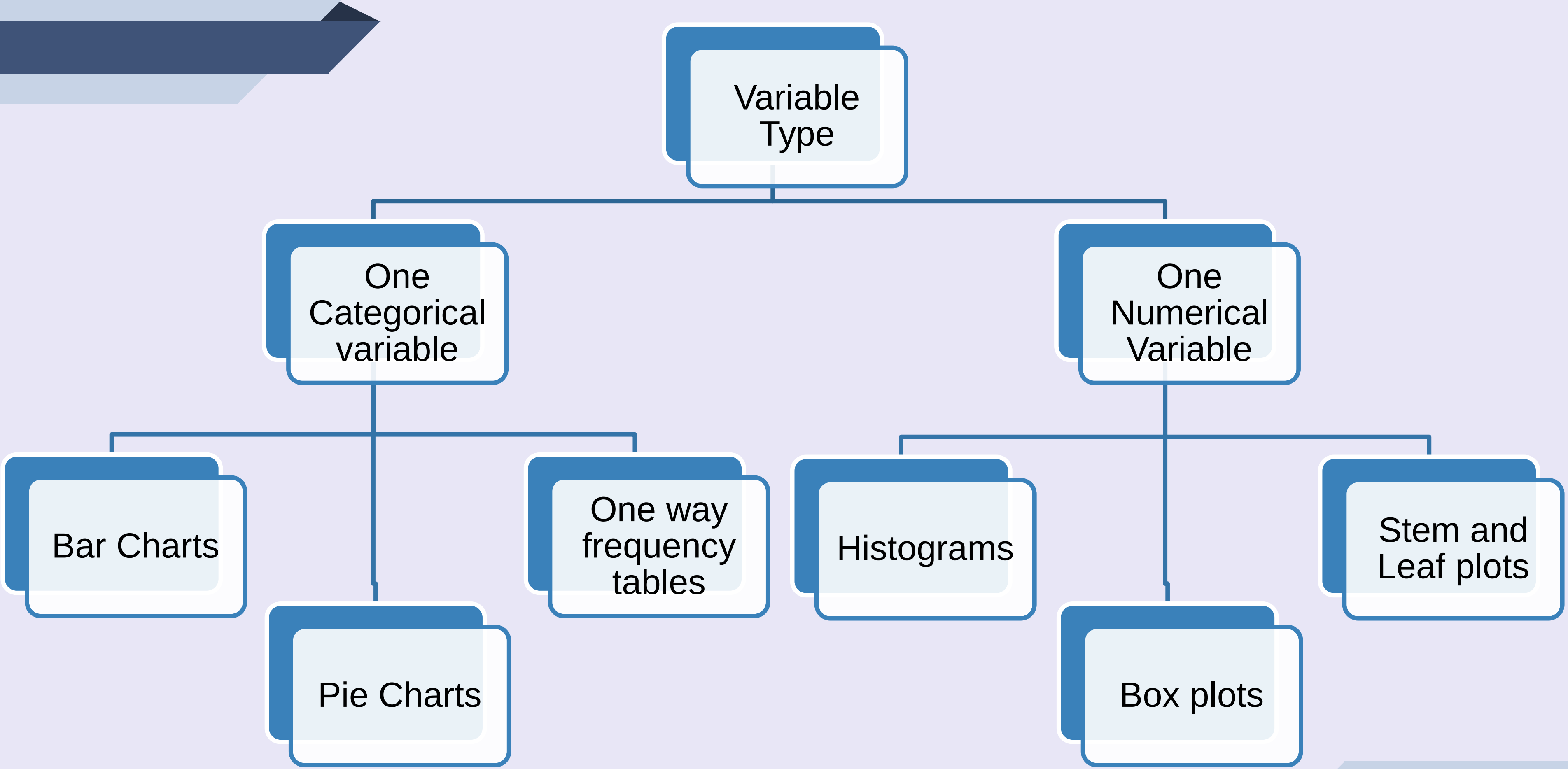
“ *This will give you an idea about the behavior of data.*

- This is also known as ***preliminary analysis***.
- It describes how the each of the variables in your analysis behave.
- There are ***two methods*** that you can use under exploratory analysis. They are,
  - *Graphical Methods &*
  - *Numerical Methods*
- Each method depends on the type of the data available

Graphical Methods

- You can use graphical methods to analyze both categorical and numerical variables.

- Type of graph you use depends on the type of the data available


Variable Type
One Categorical variable
One Numerical Variable
Bar Charts
Pie Charts
One way frequency tables
Histograms
Box plots
Stem and Leaf plots

One-way Frequency Tables

Categorical Variable

Gender	Frequency
Male	48
Female	52

Numerical Variable

Marks	Frequency
0-20	12
21-40	8
41-60	42
61-80	56
81-100	10

Bar Charts

- In bar charts, each bar will represent each category level. These bars can be drawn in vertically or horizontally.
- Frequency, cumulative frequency or percentages can be used for the y axis while x axis will represent the categorical variable.
- Length of the bar will proportional to the value it represent.

Bar Charts Cont…

- There are several type of bar charts. For example,

Simple Bar Charts

Component bar charts / Stacked bar charts

Percentage component bar charts

Multiple bar charts / Clustered bar charts

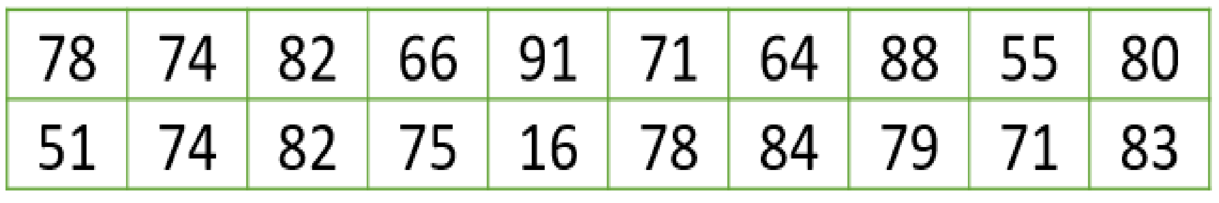
Pie Charts

- Pie charts are used to analyze *one categorical* variable.
- In pie charts, area of each sector will be proportional to the value of category it represent.
- This is *appropriate*, when there are *few number of categories* for the variable or when *value* of each *category* is *varying* widely.

Histograms

- First, divide the given data set into suitable number of classes (intervals/categories) which have the same width.
- Classes with their frequencies (counts) is called a frequency distribution.
- Frequency, relative frequency or percentages can be used for the y axis while x axis will represent the classes of the variable.
- In histograms, each bar will represent each class and length of the bar will proportional to the frequency of respective class.
- In histograms, ***bars are drawn adjacent with each other*** (No gaps between two bars).

Example:-



78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

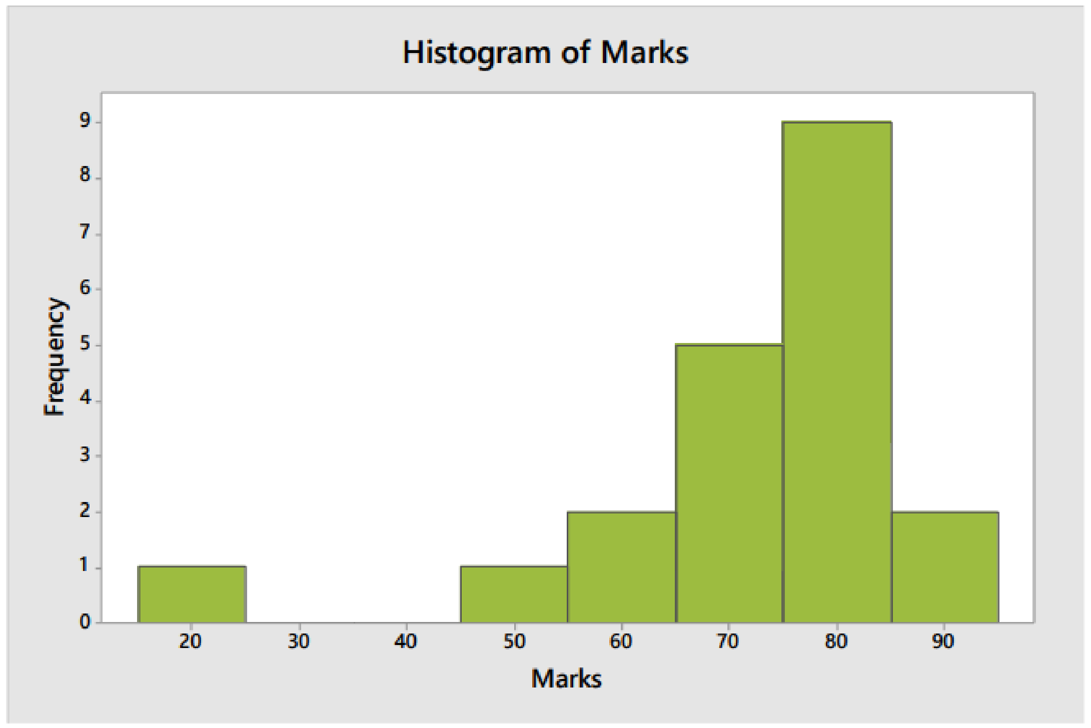
- Range = Max – Min = 91-16 = 75

- Divide the range into required number of classes to find class width (*Eg:- 8*):-

$$75 / 8 = 9.375 \approx 10$$

- Classes can be selected by fixing the class width also

Example:-



Class	Frequency
14.5 – 24.5	1
24.5 – 34.5	0
34.5 – 44.5	0
44.5 – 54.5	1
54.5 – 64.5	2
64.5 – 74.5	5
74.5 - 84.5	9
84.5 – 94.5	2

Box Plots

- To draw a box plot, it is need to identify the *five number summary* & *outliers* for the variable.
- Five Number Summary:
  - *Minimum*
  - *Maximum*
  - *Q1*
  - *Q2 (Median)*
  - *Q3*

Outliers

- Before drawing the box-plot we should identify the potential outliers.
- A limit should be defined for the accepted range of values.

$$Upper\ Bound = Q_3 + 1.5 * IQR$$

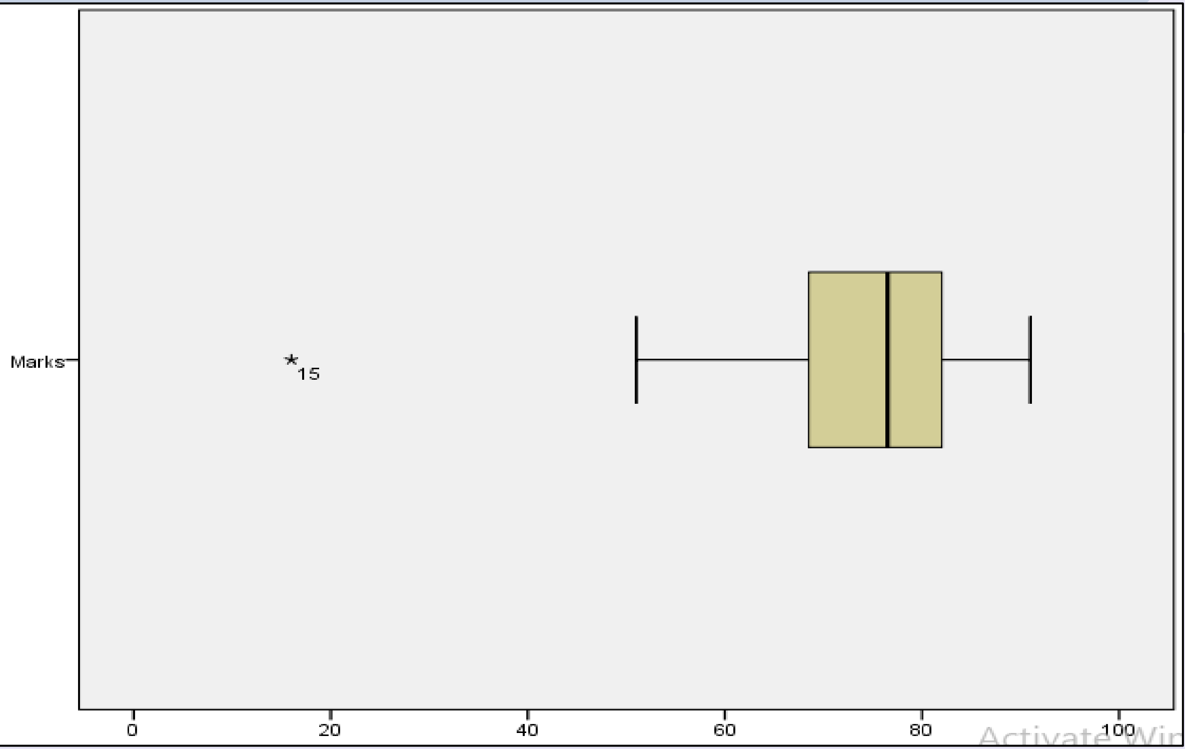
$$\text{Lower}\ Bound = Q_1 - 1.5 * IQR$$

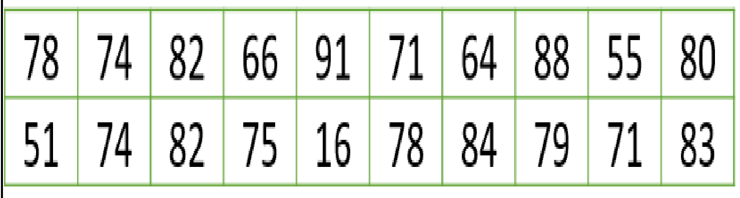
- Values outside the range are considered as outliers and marked with asterisks (*).

Outliers

- $Q1$, Median, $Q3$ are marked as a box.
- Minimum & maximum values ***which are not outliers***, will be end point for whiskers of the box plot.

Example:-



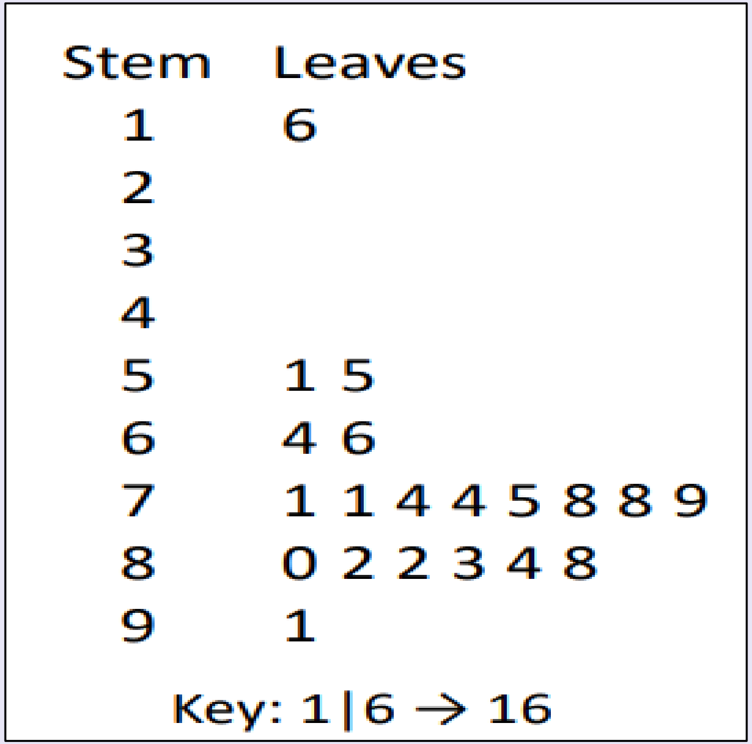


78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

Stem-and-Leaf Plots

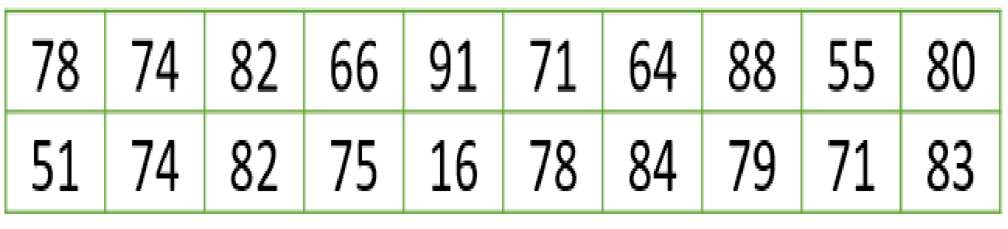
- These plots are useful when the data set is very small.
- Before drawing this plot, it is need to arrange the data in ascending order.
- Then, each data value will split into two parts known as ***stem*** and *leaf*.
- The “leaf” is usually the last digit of the number.
- The other digits to the left of the “leaf” form the “stem”.
- These plots are not much use in preliminary analysis.

Example:-



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

Key: 1|6 → 16



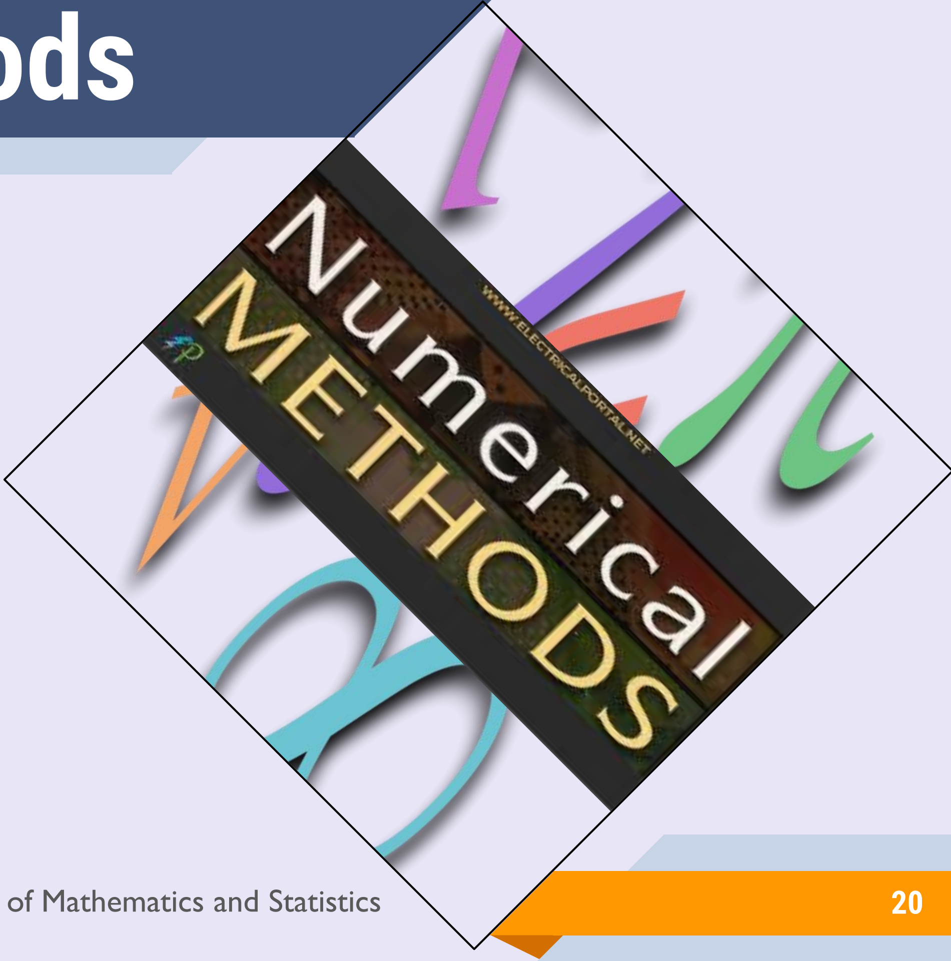
78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

Describing Two Variables

- Two categorical variables:
  - *Two-way frequency table*
  - *Clustered bar chart*
  - *Stacked bar chart*
- One categorical & one numerical variable:
  - *Parallel box plots*
  - *Comparison of location measurements for each category.*
- Two numerical variables:
  - *Scatter plot*

Numerical Methods

- Numerical methods are applied only for numerical variables.
- These methods summarize the variable into a single value.



Numerical Methods Cont…

- This has measurements under four main sections. They are,

Measures of central tendency

Measures of dispersion

Measures of skewness

Measures of kurtosis

Measures of Central Tendency

- This gives an idea about the *location* of the data as a whole.
- Following three measurements can be used for this.
  - *Mean*
  - *Median*
  - *Mode*
- Other location measurements :
  - *Percentiles / Deciles / Quartiles*

Mean

- Different types of means
  - Arithmetic mean
  - Geometric mean
  - Harmonic mean

- Only the arithmetic mean is discussed (referred to as the mean).

- Mean of a population (μ), with N elements ($x1$, $x2$, … , xN),

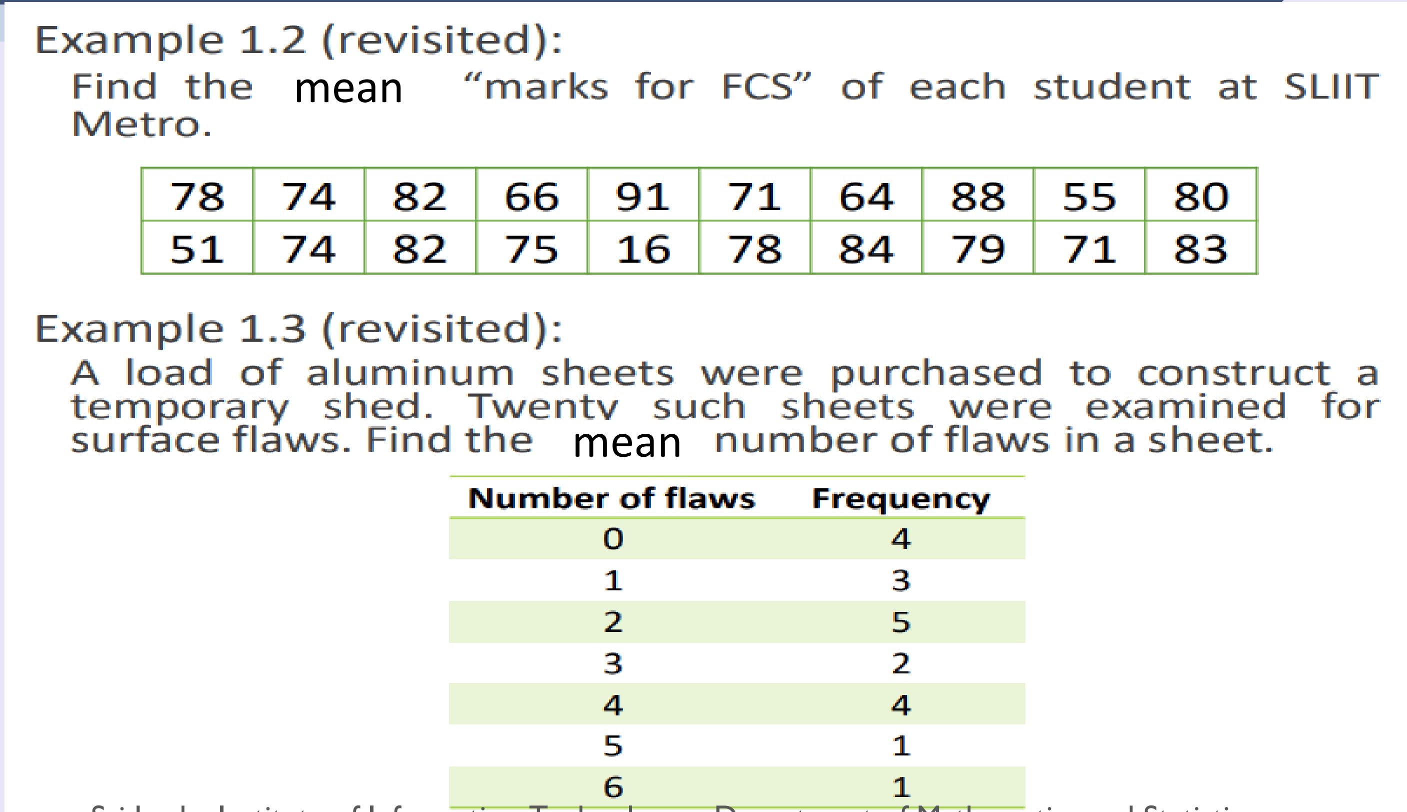
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Mean of a sample ($\bar{x}$), with n elements ($x1$, $x2$,…, xn),

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- If not specified, consider the data are coming from a sample.

Examples:-



Example 1.2 (revisited):

Find the mean "marks for FCS" of each student at SLIIT Metro.

78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

Example 1.3 (revisited):

A load of aluminum sheets were purchased to construct a temporary shed. Twenty such sheets were examined for surface flaws. Find the mean number of flaws in a sheet.

Number of flaws	Frequency
0	4
1	3
2	5
3	2
4	4
5	1
6	1

Mode

- A value with the highest frequency in a data set.
- There can be multiple modes in a data set.
- If all the data values are different, the data set has no mode.

Percentiles

- Divides the entire set of values into 100 equal sections.
- The values should be ordered in ascending order.
- Position of the $k\ th$ percentile. (n – number of observations)

$$\boldsymbol{Position\ of\ P_k = (\frac{n+1}{100}) * k}$$

- Find the value that corresponds to the found position from the ordered set of values.
- If the position is not an integer the following methods can be used.
  - *Nearest Rank method*
  - *Linear Interpolation*

Deciles & Quartiles

- Deciles divides the entire set of values into 10 equal sections.
- Quartiles divides the entire set of values into 4 equal sections.
- Method is the same as what has used for percentiles.

Measures of Dispersion

- This gives an idea about the ***dispersion / spread*** of the data as a whole.
- Following three measurements can be used for this.
  - *Range (Max - Min)*
  - *IQR (Q3 – Q1)*
  - *Variance & Standard Deviation* ($\sqrt{\boldsymbol{Variance}}$)
- Range is more suitable for small data sets.
- Range and variance are highly sensitive for outliers while, IQR is not sensitive for outliers.

Variance & SD

- This is a measurement of ***dispersion/spread*** of the data. This describes how the data has dispersed around its mean.

- Sensitive to outliers.(Not robust for outliers).

- Variance of a population (σ^2), with N elements ($x1$,$x2$, … ,xN),

$$\boldsymbol{\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

- Variance of a sample (s^2), with n elements ($x1$, $x2$, … , xn),

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- If not specified, consider the data are coming from a sample.

- Standard deviation (SD) is the square-root of the variance.
  - ▻ *Population SD – σ*
  - ▻ *Sample SD - s*

Measures of Skewness

- This gives an idea about the ***asymmetry*** of the data as a whole.
- If the data set is symmetric skewness will be zero.
- A negative skew means that the left tail is longer; the mass of the distribution is concentrated on the right.
- A positive skew means that the right tail is longer; the mass of the distribution is concentrated on the left.

Measures of Kurtosis

- This gives an idea about whether the distribution is *peaked* or *flat.*

Example

A sample of 25 plastic hinges was subjected to repealed stress cycles until failure. The number of cycles which each survived is given below.

72, 35, 63, 67, 87, 71, 64, 47, 60, 81, 39, 52, 57, 74, 43, 55, 37, 83, 48, 91, 53, 44, 94, 65, 75

I. Find five number summary

II. Find mode, p_{15}, D_3, mean, variance & sd.

III. Draw box plot & stem & leaf plot.

IV. Comment on the distribution of data.

THANKS!

Any questions?