DESCRIPTIVE STATISTICS

By SLIIT Mathematics Unit
Faculty of Humanities and Sciences

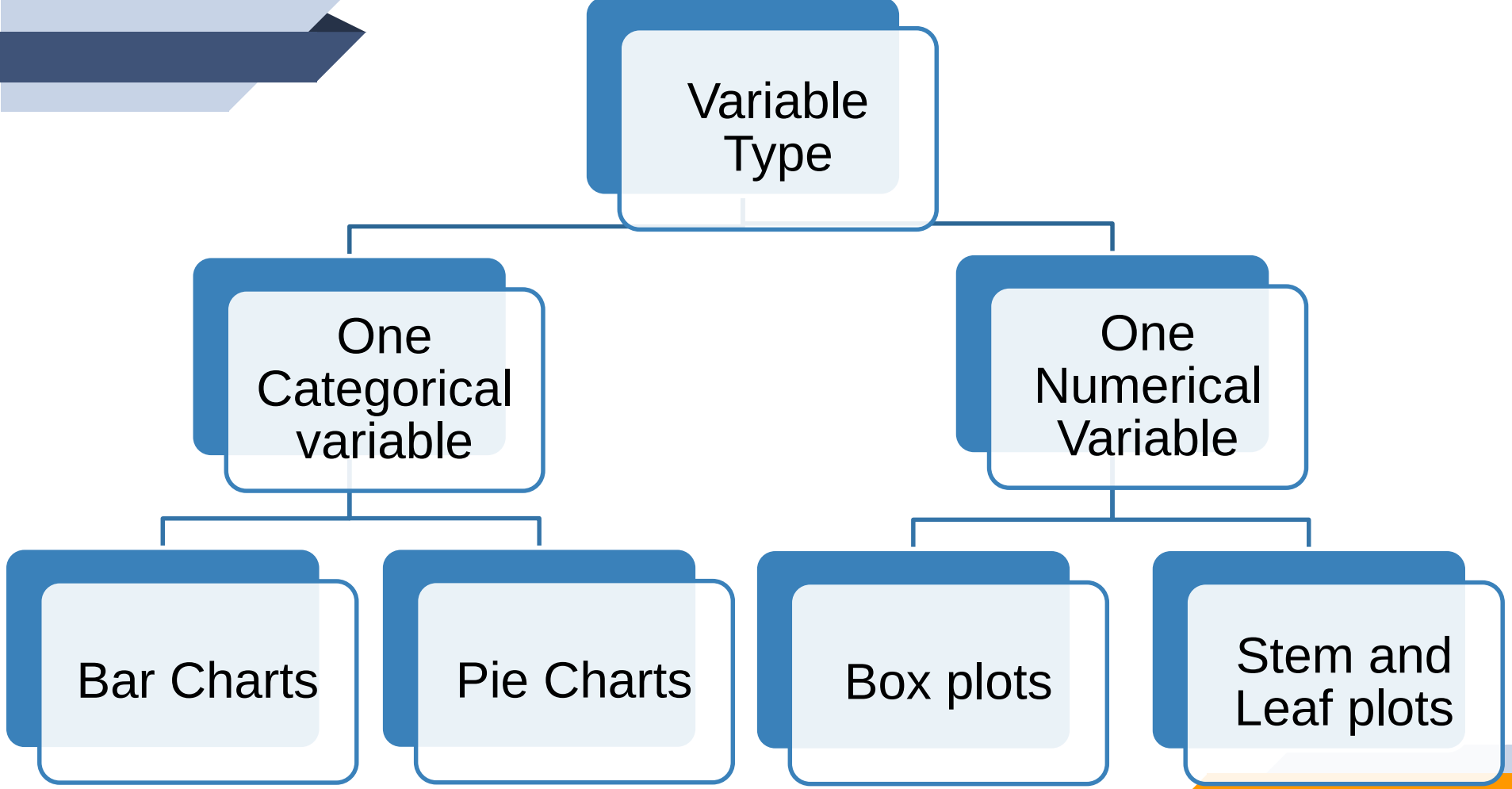
> *This will give you a rough idea about the behavior of data.*

- 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
Box plots
Stem and Leaf plots

Box Plots

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

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”.

Example:-

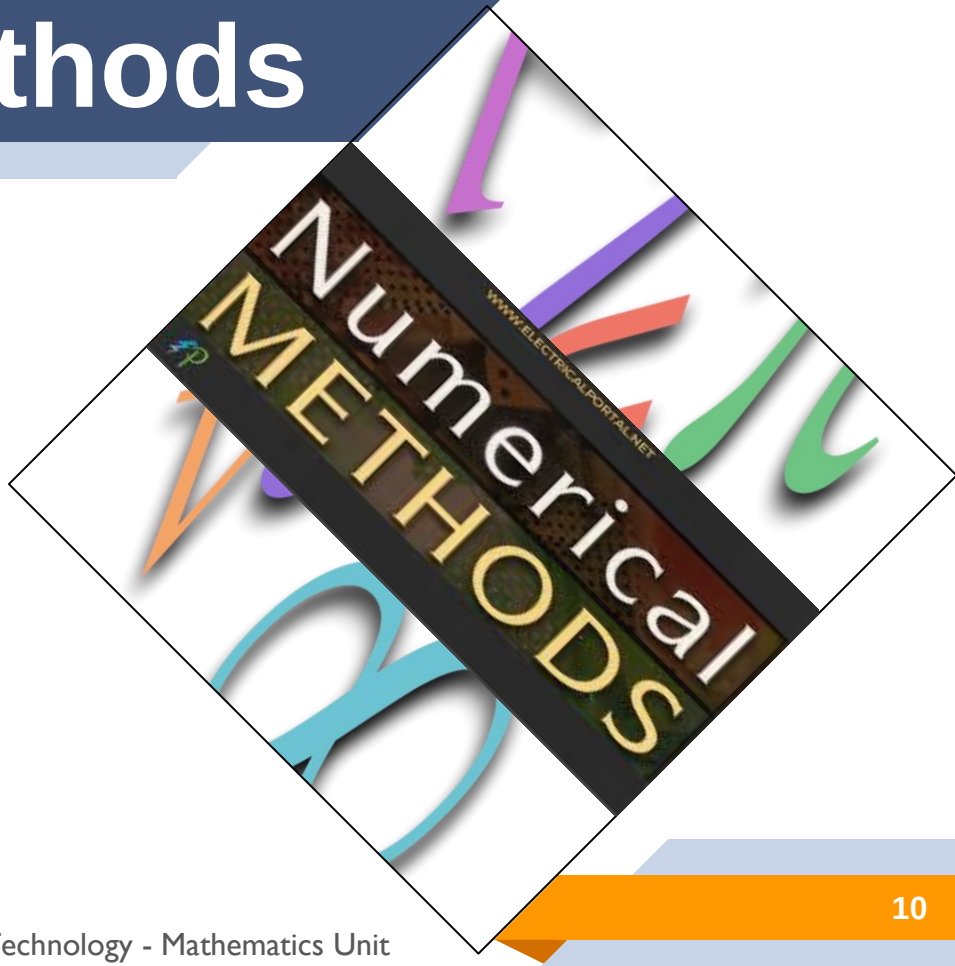
Stem	Leaves
1	6
2	
3	
4	
5	1 5
6	4 6
7	1 1 4 4 5 8 8 9
8	0 2 2 3 4 8
9	1

Key: 1|6 → 16

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

Numerical Methods

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



Numerical Methods Cont…

- We will discuss numerical measurements under **two** main sections.
  - *Measures of **Central Tendency***
  - *Measures of **Dispersion***

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***

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),

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

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

Examples:-

Example 1.2 (revisited):

Find the median "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

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.

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 ($\sqrt{Variance}$)***
- Range is more suitable for small data sets.

Variance & SD

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

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

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

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

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

- 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*

Example

A sample of diameter of 25 plastic balls 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, mean, variance & sd.
III. Draw box plot & stem & leaf plot.

THANKS!

Any questions?