Sri Lanka Institute of Information Technology

Faculty of Computing

IT2120 - Probability and Statistics

Ms. K. G. M. Lakmali

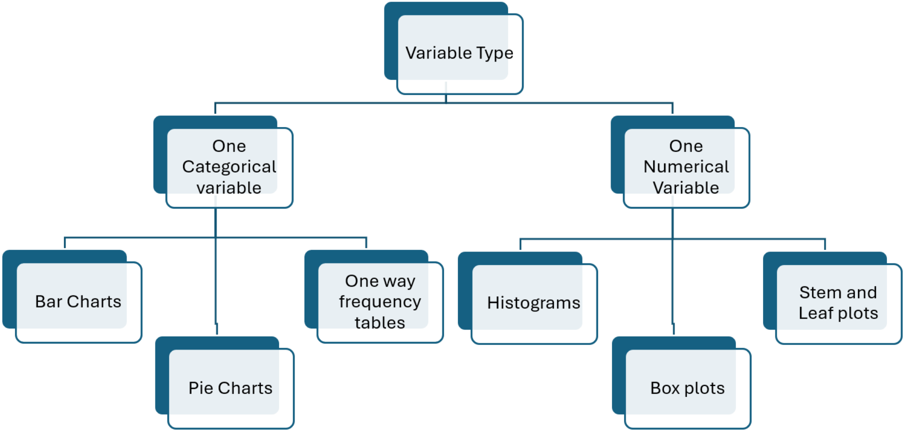
Year 02 and Semester 01

2. DESCRIPTIVE STATISTICS

- This is also known as **preliminary analysis**.
- This will give you an 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

One way frequency tables

Histograms

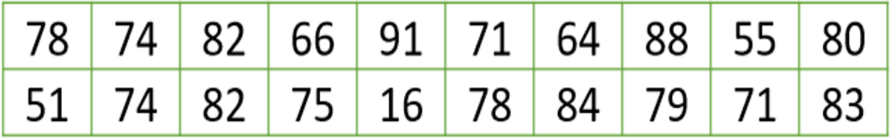
Box plots

Stem and Leaf plots

Histograms

- First, divide the given data set into suitable number of classes (intervals/categories) which have the same width.
- Then, obtain frequency distribution. 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).

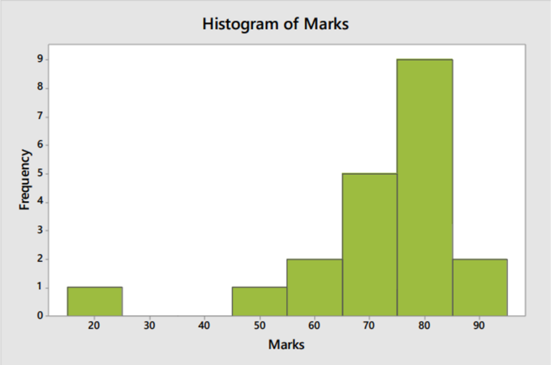
Example



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 (Most of the time, suitable value for number of classes will be between 5 10) to find class width
  (Eg:- 8):- 75 / 8 = 9.375 10
- Classes can be selected by fixing the class width as well.

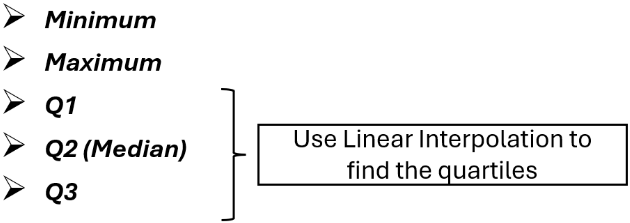
Example



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***

Use Linear Interpolation to find the quartiles

Outliers

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

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR}$$

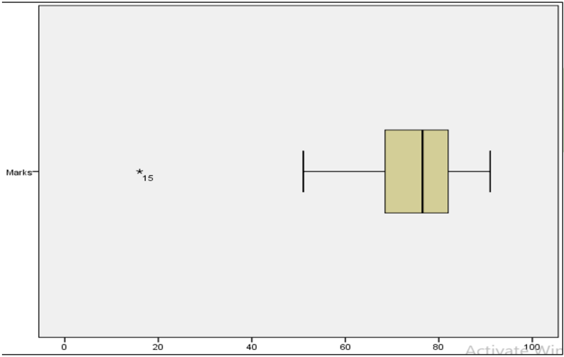
$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR}$$

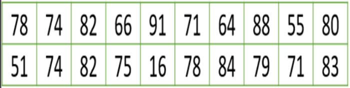
- Values outside the range are considered as outliers and marked with asterisks (*).

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

Numerical Methods

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

Numerical Methods

- 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 ($x1, x2, \ldots, xN$)

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

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

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

- In a question, 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.

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** ($Q_3 - Q_1$)
    - **Variance & Standard Deviation** ($\sqrt{\textbf{Variance}}$)
- 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 ($x1, x2, \dots, xN$)

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

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

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

- In a question, 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 Relative Dispersion

- Coefficient of Variation (CV) is the most popular relative measure of dispersion that indicates the magnitude of variation relative to the magnitude of the mean.

$$Coefficient\ of\ Variation\ (CV) = \frac{sd}{\bar{x}} * 100$$

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

1. Find five number summary.
2. Find mode, mean, variance & sd.
3. Calculate relative dispersion for the data set.
4. Draw the box plot.
5. Comment on the distribution of data.

THANKS!

Any questions?