Descriptive Statistics

Descriptive Statistics

First step in the statistical analysis process.

- Explaining or describing available data.
- Two main methods/tools for describing data.
  - Graphical Methods (e.g. Frequency table, Pie chart, Bar graph, Histogram)
  - Numerical Description(e.g. Mean, Median, Quartiles, Variance, Skewness)
- Each tool depends on the type of data variable that needs to described.

Graphical Methods

For one categorical variable

- Bar charts
- Pie charts
- One-way frequency tables

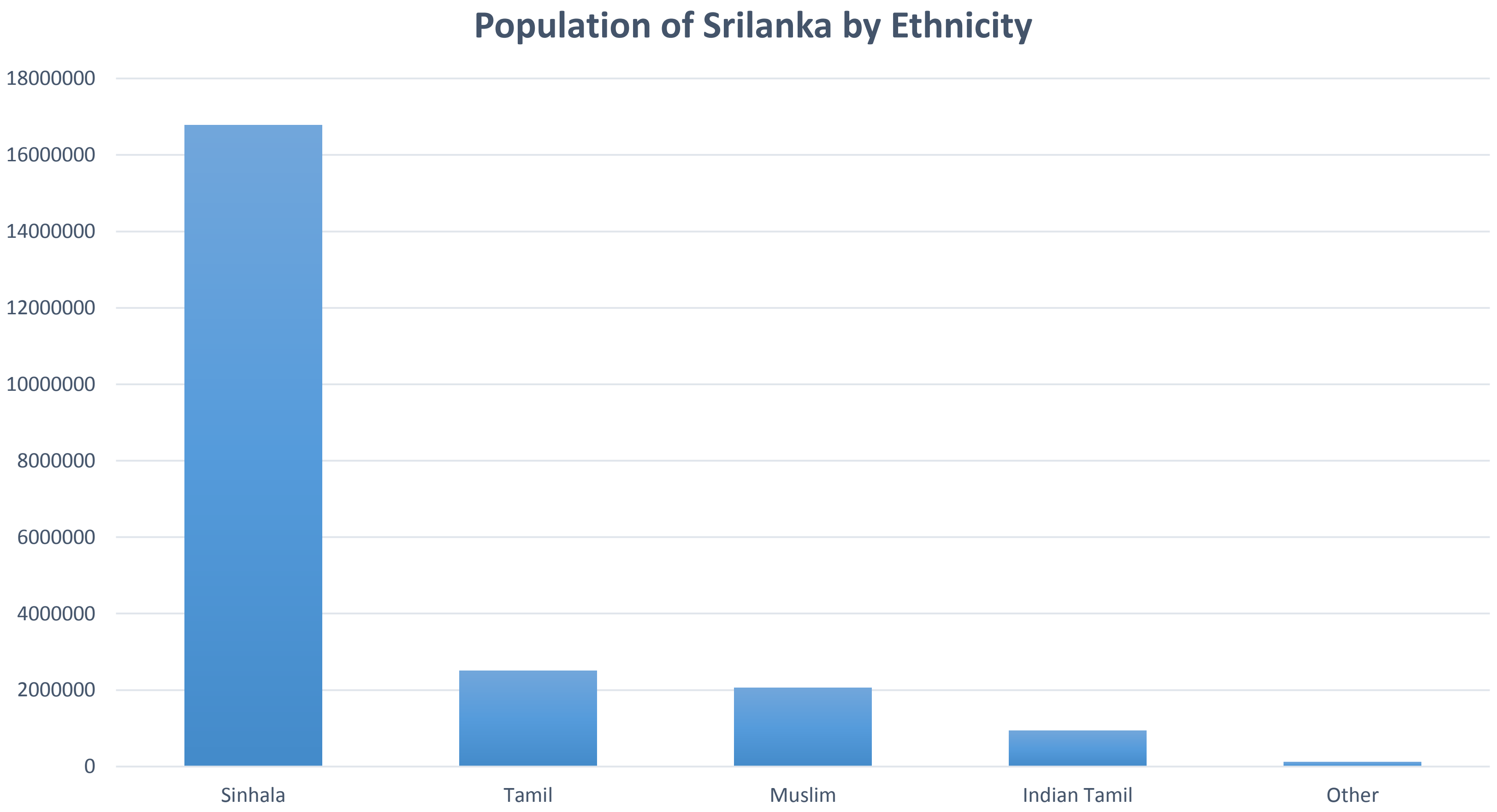
Sri Lanka demographic Profile 2018 – Ethnic groups

1. One Way frequency table

Ethnicity	**Population**
Sinhala	16784626.4
Tamil	2509850.67
Moor	2061663.05
Indian Tamil	941194.002
Other	112046.905

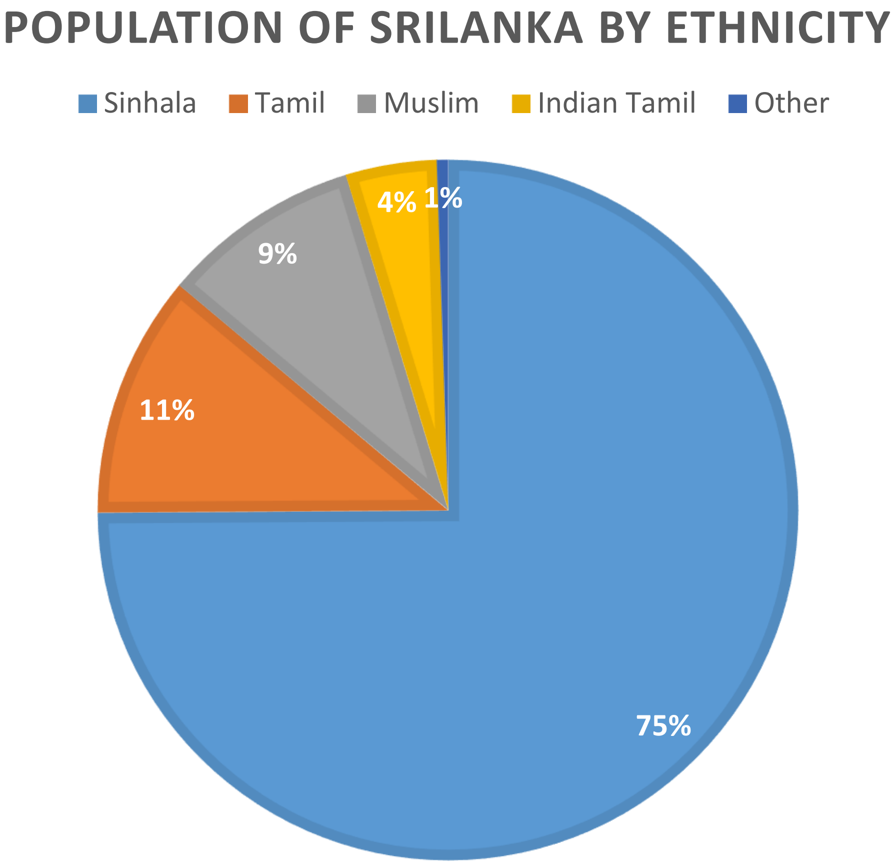
Sri Lanka demographic Profile 2018 – Ethnic groups

2. bar Charts



Sri Lanka demographic Profile 2018 – Ethnic groups

3. Pie Charts



Graphical Methods

For one numerical variable,

- Stem-and-leaf plot
- Histogram
- Boxplot

Stem-and-leaf plot

- When data sets are relatively small, stem-and-leaf plots are particularly useful.
- Each data value is split into a “stem” and a “leaf”.
- The “leaf” is usually the last digit of the number.
- The other digits to the left of the “leaf” form the “stem”.
- Sort the leaves in ascending order.
- Would be easier if the values are sorted at the beginning before drawing the plot.

Example 1.2:

Describe the variable containing data on the “Marks for FCS” of each student at SLIIT Metro.

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

Note : After collecting raw data prepare them in ascending order.

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

Histogram

- Data set is divided into a suitable number of categories/intervals called **classes**.
- Classes with their frequencies (counts) is called a *frequency distribution*.
- A histogram is a graph in which classes are marked on the horizontal axis.
- Classes - The frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis.
- **In a histogram, the bars are drawn adjacent to each other without any gaps**.
- Visual difference between bar charts and histograms???

Example 1.2 (revisited):

Describe the variable containing data on the “marks for FCS” of each student at SLIIT Metro.

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

- Find the range (Maximum – Minimum) of values.

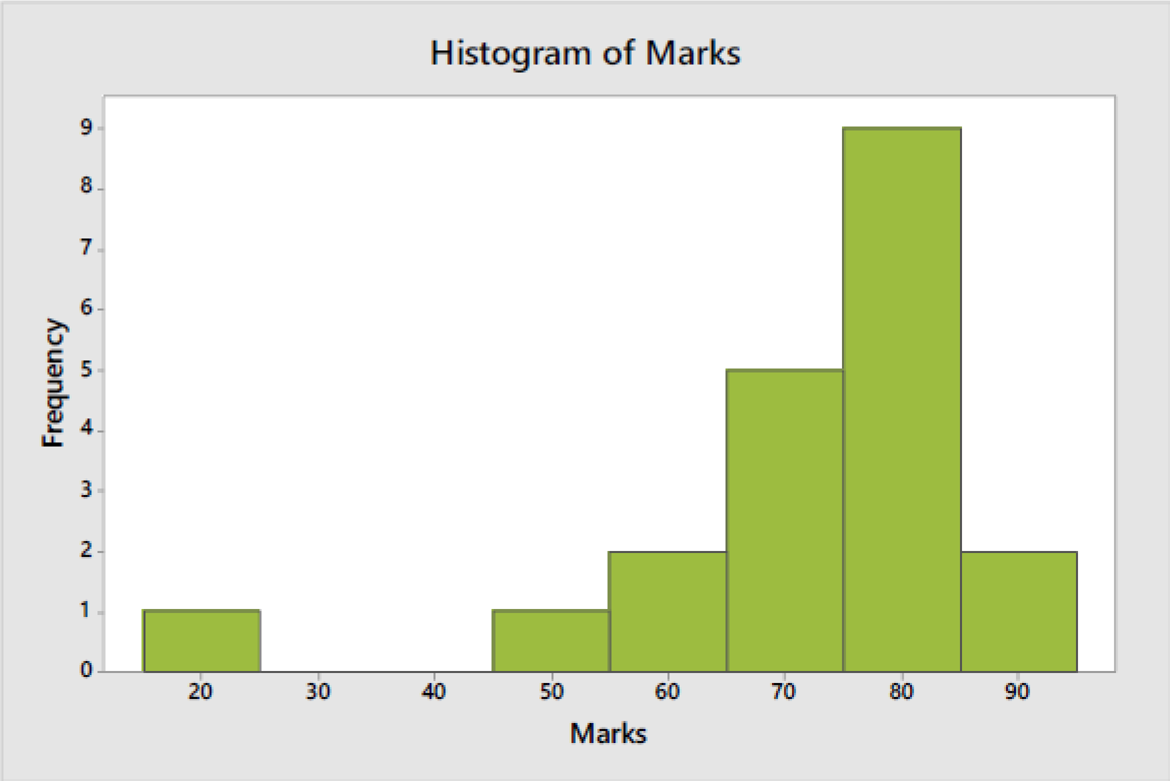
$$91 - 16 = 75$$

- Divide the range into the required number of classes to find the class width. *(E.g.: 8) =*

$$\frac{75}{8} = 9.375 \approx 10$$

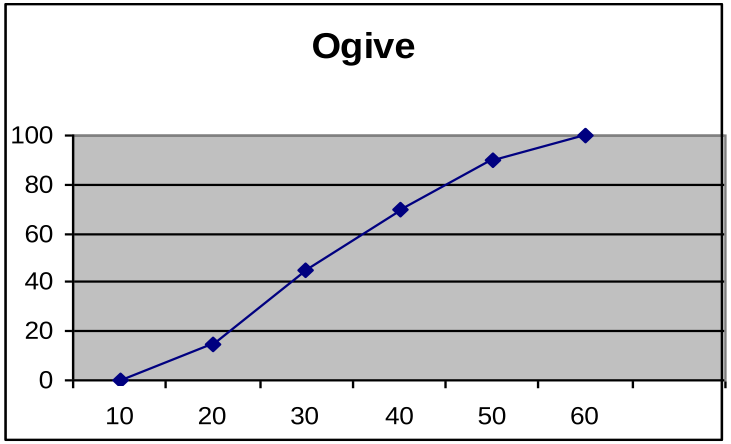
- The classes can be selected by fixing the class width also.

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

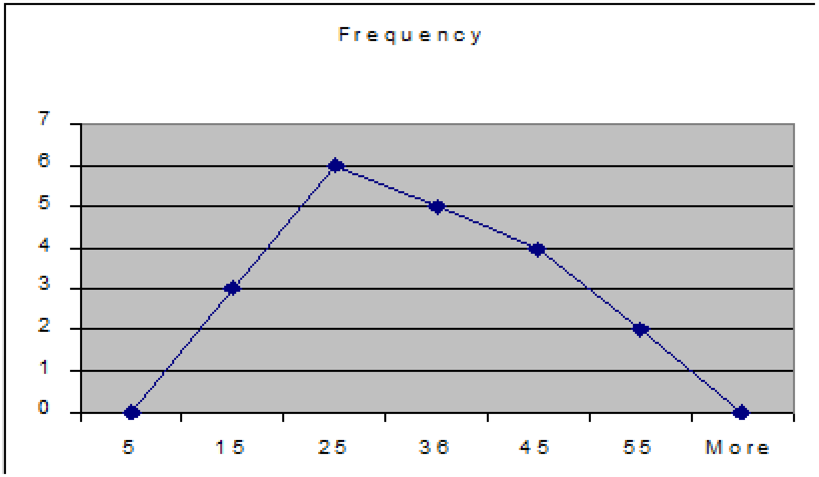


Self Study

The Ogive



The Frequency Polygon



Cumulative Frequency..?

Relative Frequency..?

Box-plot

Will be discussed later in the chapter.

Numerical Methods

- Only for numerical variables.
- Summarizes values/distribution to a single variable.
- Has measurements under 4 main sections.
  - Measures of Central Tendency
  - Measures of Dispersion
  - Measures of Skewness
  - Measures of Kurtosis
- Students should be able to calculate different measures of central tendency and dispersion discussed.

Measures of Central Tendency

- Gives an idea about the **location** of the values as a whole.
- 3 measurements of central tendency/location.
  - 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 ($x_1, x_2, \ldots, x_N$).

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

- Mean of a sample ($\bar{x}$), with n elements ($x_1, x_2, \ldots, x_n$).

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

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

Example 1.2 (revisited):

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

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

Example 1.3:

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

Median

- The median is the value right in the middle of the **ordered** data set.
- The values can be ordered in ascending or descending order.
- Find the position of the median. (n – number of observations)

$$Position\ of\ the\ median = \left(\frac{n+1}{2}\right)^{th}$$

- Find the value that corresponds to the found position from the ordered set of values.
- If n is even, then the median is the average of the two middle numbers.

Example 1.2 (revisited):

Find the median “marks for FCS” of each student at SLIIT Metro.

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

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 median 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

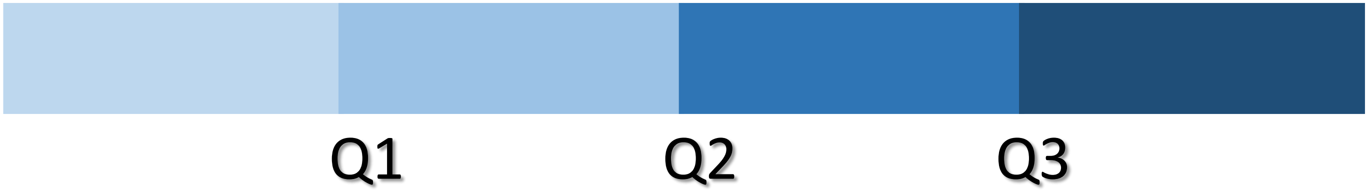
- Mode is the most frequently occurring member of the data set.
- If all the data values are different, the data set has no mode.
- There can be multiple modes for a single set of values.
- This property contradicts with the mean and median.
- Example 1.2 (revisited):

  Find the mode(s) “marks for FCS” of each student at SLIIT Metro.

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

QUARTILES

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



First quartile position: $\mathbf{Q_1 = (n+1)/4}$

Second quartile (median)position: $\mathbf{Q_2 = (n+1)/2}$

Third quartile position: $\mathbf{Q_3 = 3(n+1)/4}$

Range & IQR

- Range is a basic measurement of dispersion.

$$Range = Maximum\ value\ - Minimum\ value$$

- Is suitable mostly for small datasets.
- Extremely sensitive to unusually large and/or small values (*known as outliers*).
- Interquartile range is not sensitive to *outliers* as the range (*more robust for outliers*).

$$IQR = Q_3 - Q_1$$

Box-plot (revisited)

- The five-number summary is displayed.
  - *Minimum*, Q_1, Median (Q_2), Q_3, *Maximum*.

- First, potential outliers should be identified.

- A limit should be defined for the accepted range of values.

$$Upper\ Bound\ (UB) = Q_3 + 1.5\ IQR$$
$$Lower\ Bound\ (LB) = Q_1 - 1.5\ IQR$$

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

- Q_1, Median, Q_3 are marked as a box. The whiskers are marked at the end of the box.

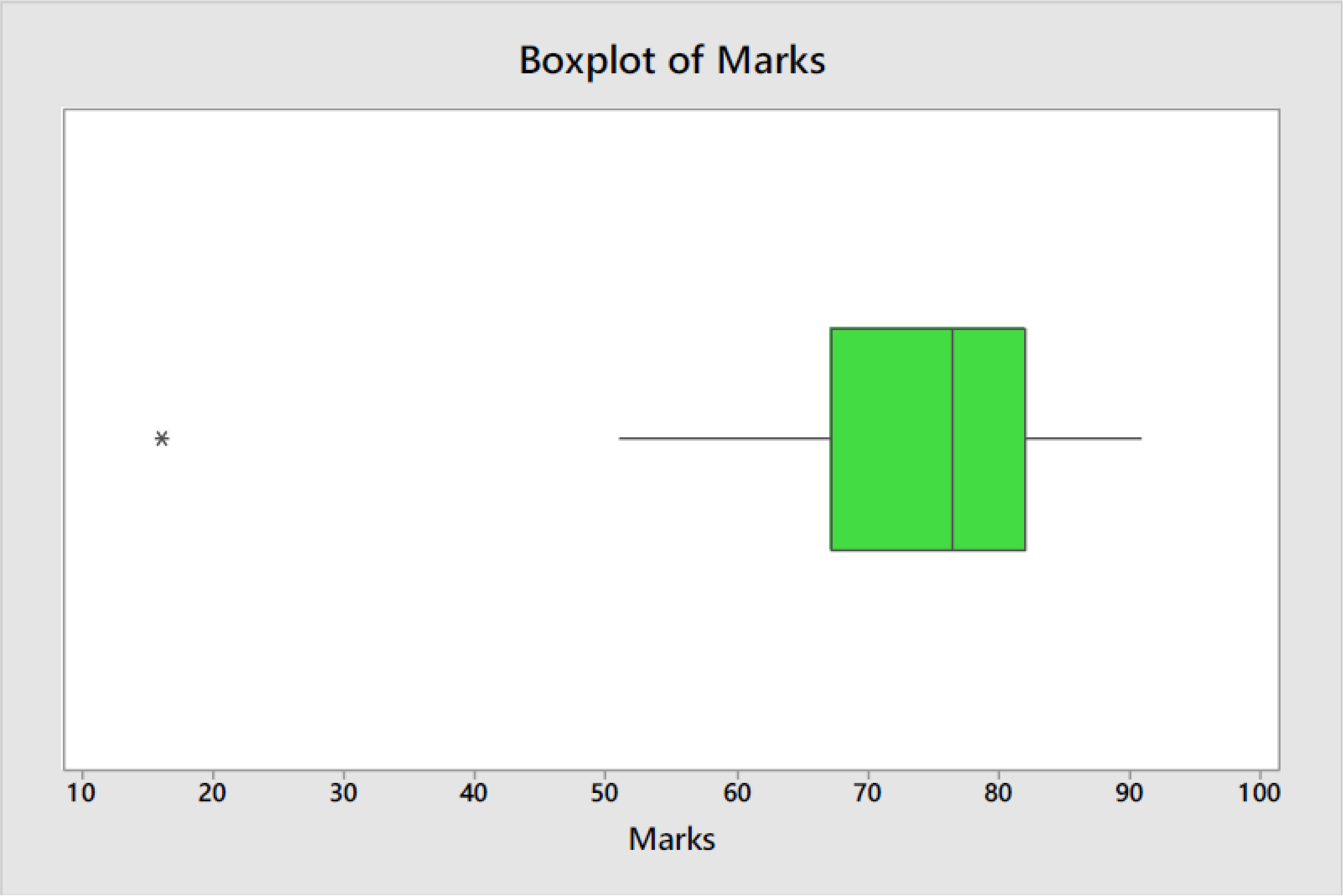
- The **minimum and maximum values that are not outliers** are the endpoints for the whiskers.

Example 1.2 (revisited):

Answer the following questions regarding the “marks for FCS” of each student at SLIIT Metro.

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

- The five-number summary?
- The inter-quartile range (IQR)?
- What is the upper bound and the lower bound to determine outliers?
- Are there any outliers?
- Draw the boxplot for the given set of values.


Boxplot of Marks
*
10
20
30
40
50
60
70
80
90
100
Marks

Self Study

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)

$$Position\ of\ P_k = \left(\frac{n+1}{100}\right) \times k$$

- 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

Self Study

Deciles & Quartiles

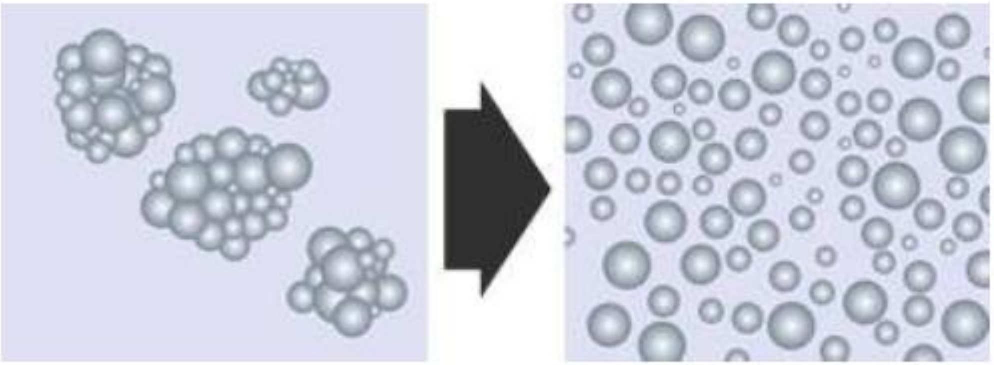
- Deciles divides the entire set of values into 10 equal sections.
- Quartiles divides the entire set of values into 4 equal sections.
- Method is the same as what was used for percentiles.
- Example 1.2 (revisited):

  Find $P_{15}, P_{50}, P_{75}, \mathrm{D}_3, \mathrm{D}_8, \mathrm{Q}_1, \mathrm{Q}_2, \mathrm{Q}_3$ “marks for FCS” of each student at SLIIT Metro.

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

Measures of Dispersion

- Gives an idea about the **spread** of the values as a whole.
- 3 measurements of dispersion/spread.
  - Range
  - Interquartile Range (IQR)
  - Variance/Standard Deviation



Variance & Standard Deviation

- Not sensitive to *outliers* as the range (*more robust for outliers*).
- Variance of a population (σ^2), with N elements ($x_1, x_2, \dots, x_N$).

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

- Variance of a sample (s^2), with n elements ($x_1, x_2, \dots, x_n$).

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

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

Variance & Std. Dev. (cont'd.)

- Using properties of the summation ($\sum$) the equations can be simplified. (*Refer Chapter 3 – Counting in the FCS textbook.*)

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

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

- Standard deviation is the square-root of the variance.
- Population standard deviation (σ).
- Sample standard deviation (s).

Effects of Transformations

- *E.g.: There are observations $x_1, x_2, \dots, x_n$ with a mean $\bar{x}$ and a variance s_x^2.*
- When the same amount (a) is added to all the values to create $y_1, y_2, \dots, y_n$,
  - *What is the relationship between $\bar{x}$ and $\bar{y}$?*
  - *What is the relationship between $s_x^2$and s_y^2?*
- When all the values are multiplied by the same amount (b) to create $z_1, z_2, \dots, z_n$,
  - *What is the relationship between $\bar{x}$ and $\bar{z}$?*
  - *What is the relationship between $s_x^2$and s_z^2?*

Measures of Skewness & Kurtosis

- Skewness gives an idea about the **asymmetry** of the values as a whole.
- If a distribution is symmetric, the Skewness is zero (0).
- 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.
- Kurtosis is a measure of whether the distribution is **peaked or flat**.

Describing two variables

- Describing **two categorical** variables.
  - Two-way frequency tables.
  - Clustered bar charts.
- Describing **one categorical** variable with **one numerical**.
  - Side by side box-plots.
  - Comparison of location measurements for each category.
- Describing **two numerical** variables.
  - Scatter-plots
- *Some of the above results will be obtained using SPSS.*

Summary

- Describing one categorical variable graphically (one-way frequency tables, bar charts and pie charts).
- Describing one numerical variable graphically (Stem-and-leaf plots, histograms and box-plots).
- Describing one numerical variable using summary measures (Find any location measurement or dispersion measurement).
- Describe two variables at the same time.
- Identify the effect of transforming a set of values (adding a constant or multiplying by a constant).