DESCRIPTIVE STATISTICS

Learning Outcomes

By the end of this lesson, you will be able to:

- organise a set of data by using tables and diagrams such as frequency table and histogram.
- describe the central tendency of a set of numerical data by using the mean, median and mode.
- describe the spread of a set of numerical data by using interquartile range and standard deviation

Introduction to Descriptive Statistics

- Data is all around us. How do use them to make decisions?
- For example, a bank may want to know the spending habits of clientele
- Based on the collected data, it will summarize the data, interpret the quantitative information and devise marketing strategy according to the spending power of its clientele.

- In this topic, you will organise a set of data by using tables and diagrams, as well as describe them using numerical measures.

Why do we need to Organise Data?

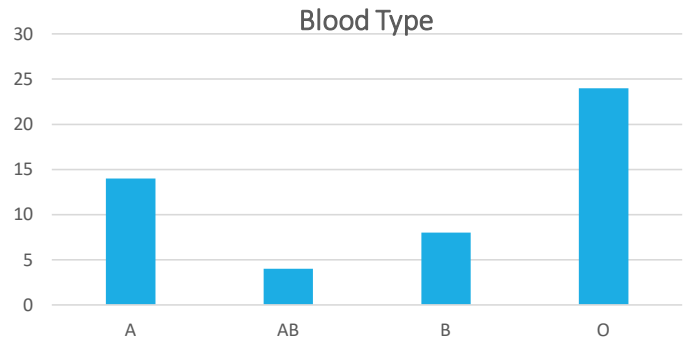
- We need to organise raw data so that we can draw some basic conclusions on the findings
- For example, the raw data on the right does not reveal much about the different blood types.
- But, once we organise data in the forms of table & charts, we can draw some basic information from them

A	A	AB	O	B	B	O	A	O	A
O	O	O	A	O	A	O	O	B	A
B	O	A	O	AB	O	A	O	O	O
O	A	O	O	A	O	O	O	B	B
AB	O	B	O	B	O	A	A	A	AB

Blood Type	Frequency
A	14
AB	4
B	8
O	24

Frequency table

Bar chart



Data Visualisation

- Raw data can be organised and summarised using
  - ✓ Tabulated form (i.e. tables)
  - ✓ Graphical descriptions (i.e. graphs and charts)
  - ✓ Descriptive statistics.

Example 1: FREQUENCY DISTRIBUTION FOR CATEGORICAL DATA

You are given a data set of blood type of 50 randomly selected people. How can we arrange and present the information for easy reading?

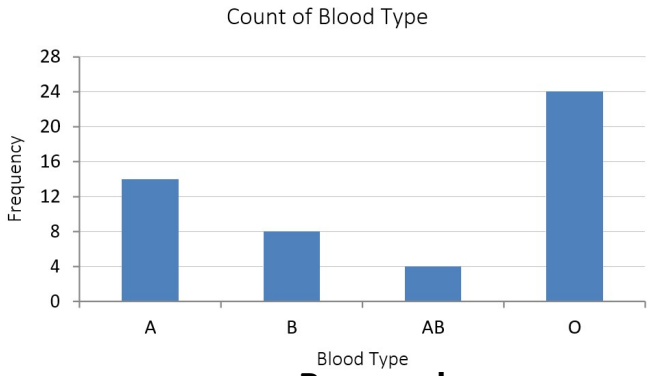
A	*A*	*AB*	*O*	*B*	*B*	*O*	*A*	*O*	*A*
O	*O*	*O*	*A*	*O*	*A*	*O*	*O*	*B*	*A*
B	*O*	*A*	*O*	*AB*	*O*	*A*	*O*	*O*	*O*
O	*A*	*O*	*O*	*A*	*O*	*O*	*O*	*B*	*B*
AB	*O*	*B*	*O*	*B*	*O*	*A*	*A*	*A*	*AB*

Data Visualisation

Blood Type	Frequency
A	*14*
B	*8*
AB	*4*
O	*24*
Total	*50*

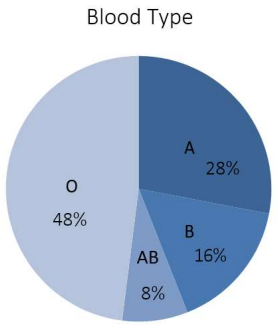
Frequency table

- Used to show the pattern of the data
- Identify where values tend to concentrate
- Expose extreme or unusual values.



Bar graph

- Graph that shows the qualitative classes on the horizontal axis and the class frequencies on the vertical axis.
- Used to compare the number of observations for each class of a qualitative variable.



Pie chart

- Chart that shows the proportion or percentage that each class represents of the total number of frequencies.
- Used to compare relative differences in the percentage of observations for each class of a qualitative variable.

Data Visualisation

- For large data sets, it is necessary to group them into intervals.
- Using Frequency Distribution, quantitative data are grouped into mutually exclusive classes showing the number of observations in each class

Data Visualisation

Example 2: GROUPED FREQUENCY DISTRIBUTION FOR CONTINUOUS DATA

A dentist measured the width (in mm) of the last lower molar of 60 female adult. He wants to present the data in the form of a Frequency Table and Histogram

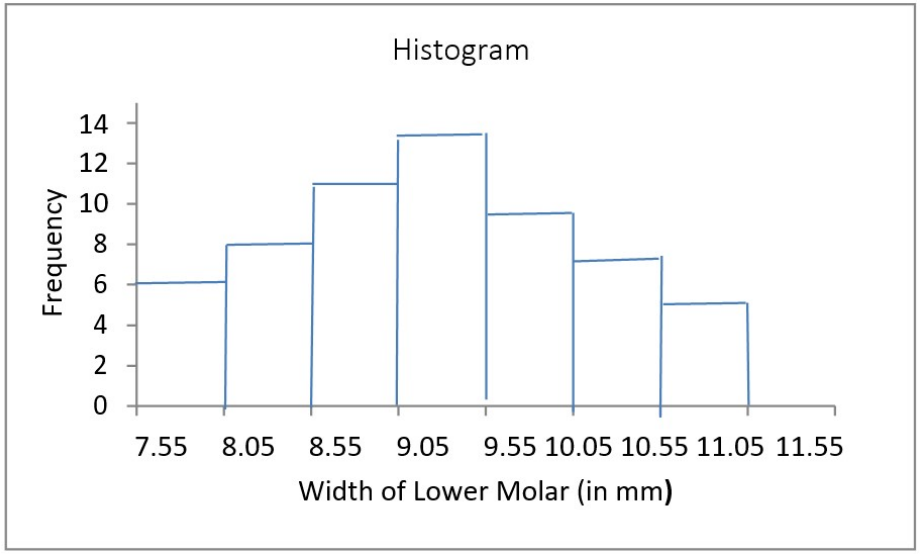
7.6	10.6	8.2	10.3	9.6	7.8	10.1	8.7	9.1	7.7
8.2	9.9	10.9	9.5	10.4	8.8	9.4	9.1	9.7	9.2
8.7	9.4	9.1	7.9	9.5	9.3	8.5	10.8	8.3	8.6
10.1	9.8	8.3	10.5	8.7	9.8	7.6	9.7	10.7	10.4
9.2	9.7	8.6	8.7	8.1	9.2	9.6	10.2	8.9	9.3
8	9.3	8.4	9.9	8.7	11	8.9	10	8.6	8.4

Class Interval	Class	Class	Frequency
7.6 - 8.0	*7.55 – 8.05*	*7.8*	*6*
8.1 – 8.5	*8.05 – 8.55*	*8.3*	*8*
8.6 – 9.0	*8.55 – 9.05*	*8.8*	*11*
9.1 – 9.5	*9.05 – 9.55*	*9.3*	*13*
9.6 – 10.0	*9.55 – 10.05*	*9.8*	*10*
10.1 – 10.5	*10.05 – 10.55*	*10.3*	*7*
10.6 –11.0	*10.55 – 11.05*	*10.8*	*5*

Frequency distribution table

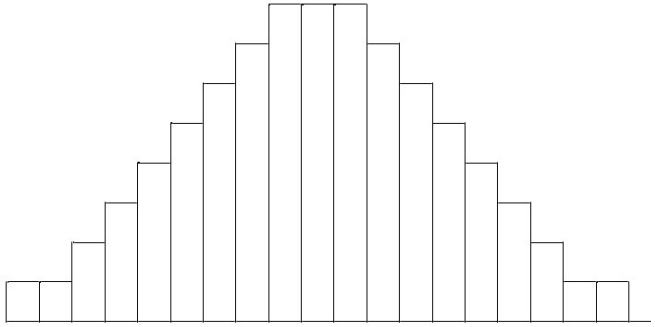
Data Visualisation

- A Frequency Distribution can be presented graphically using a **Histogram**
- A Histogram is a graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis.

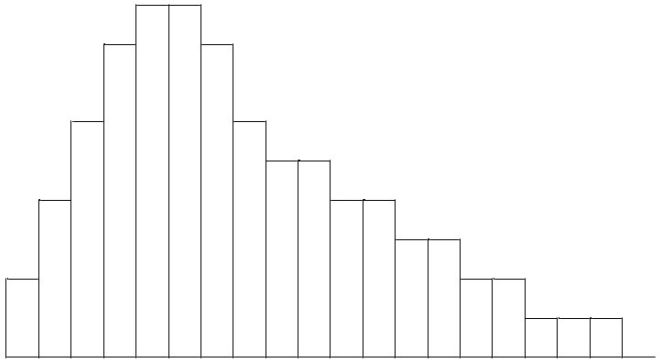


Data Visualisation

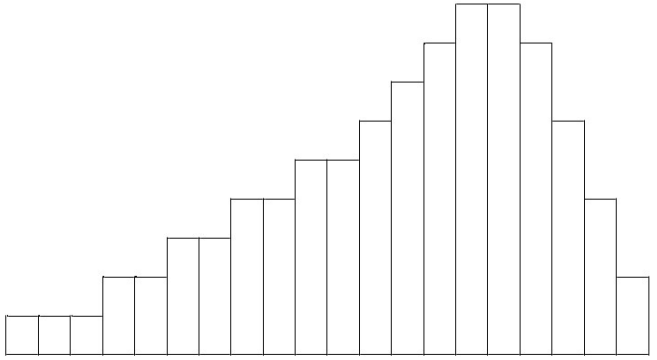
- Histograms come in different shapes.



Symmetric, bell-shaped



Skewed to the right



Skewed to the left

Data Visualisation

- Stem and Leaf Plots
  - Like Histograms, Stem and Leaf Plots helps us to visualise data
  - It provides more information than what a Histogram displays.

Example 3: STEM & LEAF PLOT

You want to visualise the scores of 20 students on a Statistics test using a Stem and Leaf Plot.

83	84	77	64	71	87	72	92	57	92	75	52
80	65	79	71	87	93	96	95				

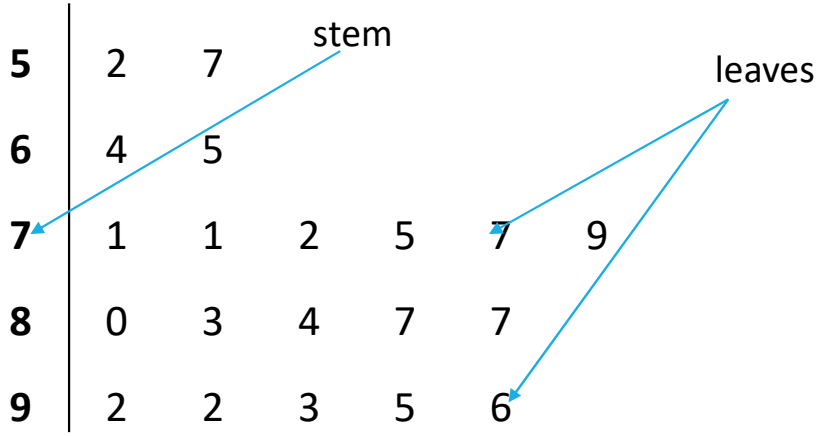
Data Visualisation

- Stem and Leaf Plots

1. Marks are first sorted in ascending order

   52 57 64 65 71 71 72 75 77 79 80 83 84 87 87 92 92 93 95 96

2. Marks are then arranged in form of stem & leaves



Descriptive Statistics

There are 3 general types of statistics that are used to describe data:

Type of descriptive statistics	Purpose/ Measures
Measures of Central Tendency	▪ Describe the central position of a frequency distribution for a group of data. ▪ Measures: Mean, Median, Mode
Measures of Dispersion/Spread	▪ Describe how spread out the data is. ▪ Measures: Standard Deviation, Variance, Range
Measures of Position	▪ Describe the position of a single data relative to other values on the data set. ▪ Measures: First Quartile, Second Quartile, Third Quartile, Inter-Quartile Range

Measures of Central Tendency

- A single value that describes a set of data by identifying the central position within that set of data.
- Measures: Mean, Median, Mode

Mean

- Most common measure of central tendency.
- Can be used with both discrete and continuous data.
- Only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

Population $\mu = \frac{\Sigma x}{N}$

Sample $\bar{x} = \frac{\Sigma x}{n}$

where
Σx is the sum of all values
N or n is the number of observations or values

Mean

Example 4: The following data shows the time spent (in minutes) on Facebook per day by a sample of 10 students. Compute the mean.

35, 35, 35, 46, 48, 53, 54, 55, 55, 145

$$\bar{x} = \frac{\Sigma x}{n}$$

= 561/10

= 56.1 minutes

Median

- Set of observations is arranged either in increasing or decreasing magnitude.
- Median obtained by getting the
  - ✓ middle value if the number of observations is odd or
  - ✓ mean of the middle two values if the number of observations is even.
- 2 steps process :
  1. Arrange data in ascending order (data array)
  2. Determine the position of median using the formula

Median

Example 5: The following data shows the time spent (in minutes) on Facebook per day by a sample of 11 students. Compute the median.

Data Array: 35, 35, 35, 35, 46, 48, 53, 54, 55, 55, 145

Median $= (\frac{n+1}{2})th = (\frac{11+1}{2})th$ =6th item

Median = 48 minutes

Median

Example 6: The following data shows the time spent (in minutes) on Facebook per day by a sample of 10 students. Compute the median.

35, 35, 35, 46, 48, 53, 54, 55, 55, 145

Median $= (\frac{n+1}{2})th = (\frac{10+1}{2})th$ =5.5th item

Median $= \frac{48+53}{2} = 50.5$ minutes

Mode

- The value that occurs most frequently (or the value with the highest frequency).
- On a histogram, it represents the highest bar.

Example 7: The following data shows the time spent (in minutes) on Facebook per day by a sample of 10 students. Compute the mode.

35, 35, 35, 46, 48, 53, 54, 55, 55, 145

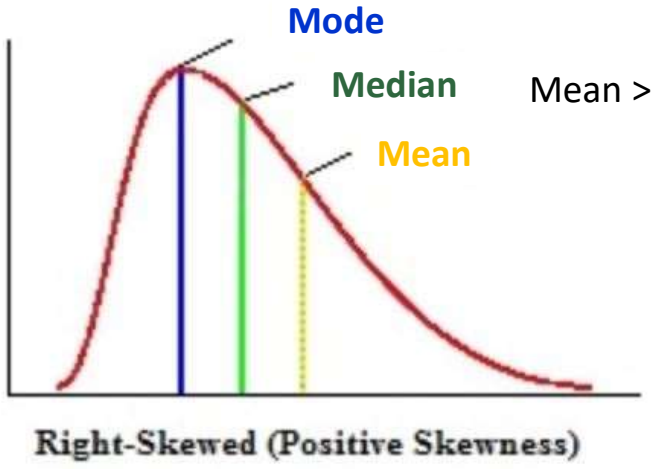
Mode = 35 minutes

Skewness of data

- Skewness refers to distortion in a symmetrical bell curve, or normal distribution, in a set of data.
- If the curve is shifted to the left or to the right, it is said to be skewed.
- Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.
- The three probability distributions are
  - Positively-skewed (or right-skewed)
  - Negatively-skewed (or left-skewed)
  - Zero-skewness

Skewness of data

- Positively-skewed (or right-skewed) distribution
  - ✓ A few extremely large or high values/observations: right skewed.
  - ✓ Mean will be the highest as the mean is influenced by a few extremely big outliers.
  - ✓ Use median instead of mean, as mean will not be representative of the central tendency of data.

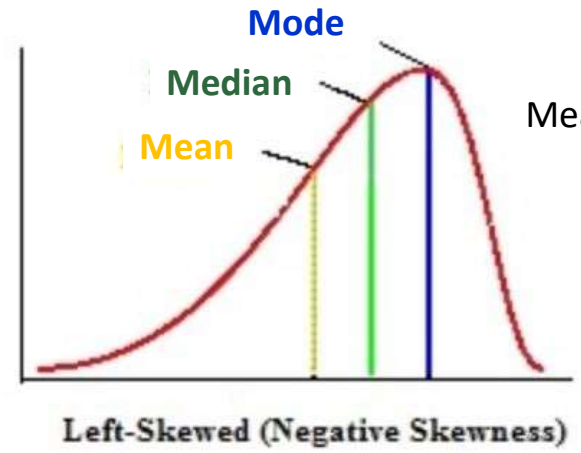


Right-Skewed (Positive Skewness)

Mean > Median > Mode

Skewness of data

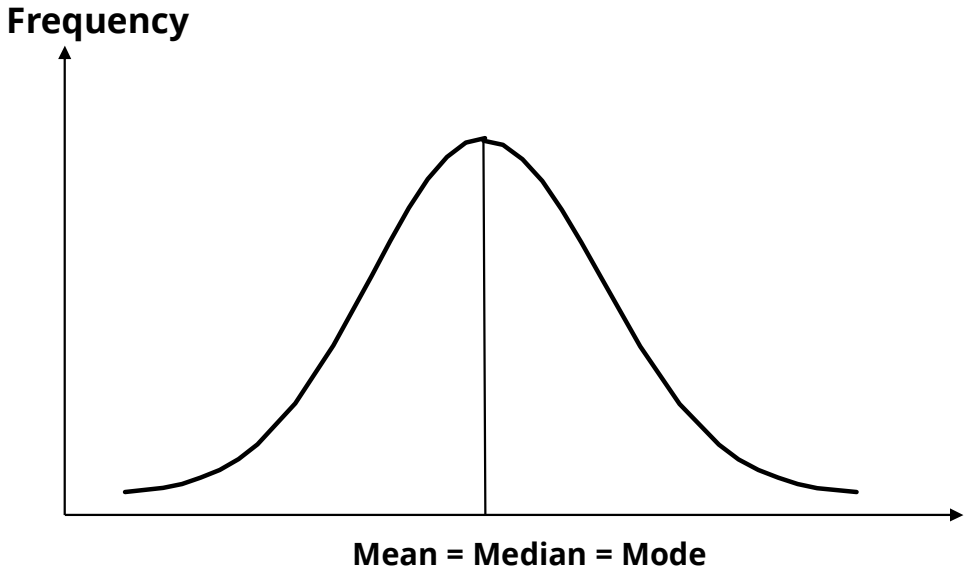
- Negatively-skewed (or left-skewed) Distribution
  - ✓ A few extremely low or small values/observations: Left skewed.
  - ✓ Mean will be the smallest as the mean is influenced by a few extremely small outliers.
  - ✓ Better to use median, as the mean will not be representative of the central tendency of data.



Mean < Median < Mode

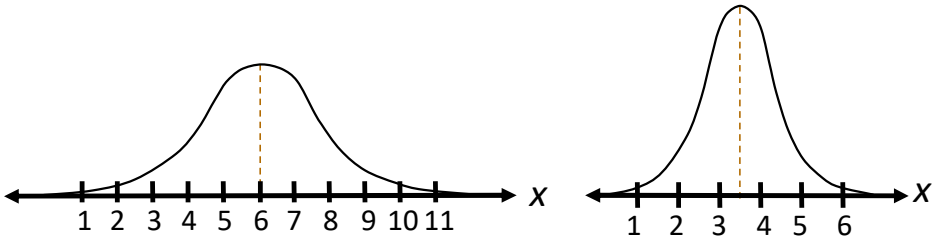
Skewness of data

- Zero skewness
  - ✓ For a symmetrical distribution (normal distribution), mean, median and mode are always equal.
  - ✓ Use mean, mode or median as a representative of the central tendency of data.



Measures of Dispersion

- Describe how spread out the data is.
- **Variance** and **standard deviation** measure the average distance each value in the data set is away from the mean.
- Thus, when the average distance between the data (x) and the mean **increases**, the variance and standard deviation **increases**



Higher Variance | Lower Variance

Population Variance $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$

Sample Variance $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

Standard deviation = $\sqrt{\text{variance}}$

Variance

- Measures the average distance each value in the data set is away from the mean.
- Example 8: On six consecutive Sundays, a call centre operator received a sample of 9, 7, 11, 10, 13, and 7 service calls. Calculate the standard deviation.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{57}{6} = 9.5 \qquad s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1}$$

$$s^2 = \frac{(9-9.5)^2 + (7-9.5)^2 + (11-9.5)^2 + (10-9.5)^2 + (13-9.5)^2 + (7-9.5)^2}{6-1}$$

$$= \frac{27.5}{5} = 5.5$$

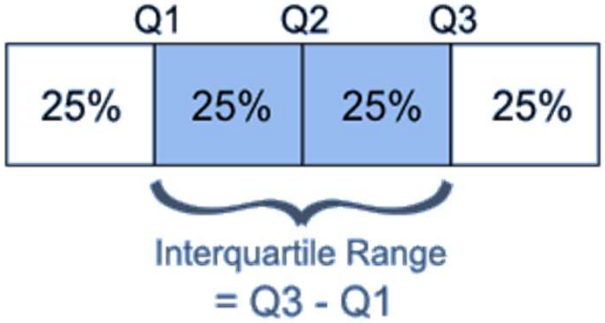
$$s = \sqrt{5.5} = 2.3452$$

Measures of Position

- Describe the position of a single data relative to other values on the data set.
- **Quartiles** are used as the summary measures that divide a ranked data set into four equal parts, denoted by Q1 (First quartile), Q2 (Second quartile, same as median) and Q3 (Third quartile).

Quartile and Interquartile Range

- Quartiles are the values that divide a list of numbers into quarters.
  - ✓ First put the list of numbers in order
  - ✓ Then cut the list into four equal parts
  - ✓ The Quartiles are at the "cuts“



- The difference between the third quartile and the first quartile gives the interquartile range (IQR).

Quartile and Interquartile Range

- Step to finding Interquartile Range.
  1. Sort the observations.
  2. Find the median Q_2.
  3. Find Q_1, the median of the data values that lie below Q_2
  4. Find Q_3, the median of the data values that lie above Q_2.
  5. Find Interquartile range Q_3 - Q_1

Measures of Position

Example 9: Given the marks obtained for a test are as follows:

5, 3, 4, 5, 11, 5, 12, 6, 8, 10, 11, 11 12, 14, 18

Find the values of the three quartiles and the interquartile range.

1. Sort the observations.
2. Find the median Q_2.
3. Find Q_1.
4. Find Q_3.
5. Find Interquartile range Q_3 - Q_1

3, 4, 5, 5, 5, 6, 8, 10, 11, 11, 11, 12, 12, 14, 18

$Q_2 = \frac{16}{2}$ th number = 10

Q_1 = median of {3, 4, 5, 5, 5, 6, 8} = $\frac{8}{2}$ th number = 5

Q_3 = median of {11, 11, 11, 12, 12, 14, 18}

= $\frac{8}{2}$ th number = 12

IQR = Q_3 - Q_1 = 12 - 5 = 7