NANYANG POLYTECHNIC

Course : Diploma in Electronic Systems
Diploma in Telematics & Media Technology
Diploma in Aerospace Systems & Management
Diploma in Electrical Engineering with Eco-Design
Diploma in Mechatronics Engineering
Diploma in Digital & Precision Engineering
Diploma in Aeronautical & Aerospace Technology
Diploma in Biomedical Engineering
Diploma in Nanotechnology & Materials Science
Diploma in Engineering with Business
Diploma in Information Technology
Diploma in Financial Informatics
Diploma in Cybersecurity & Forensics
Diploma in Infocomm & Security
Diploma in Chemical & Pharmaceutical Technology
Diploma in Biologics & Process Technology
Diploma in Chemical & Green Technology

Module	:	Engineering Mathematics 2B / Mathematics 2B/ Computing Mathematics 2	– EG1761/2008/2681/2916/2961 EGB/D/F/H/J/M207 IT1201/1531/1631/1761 CLB/C/G201

Topic 1 : Descriptive Statistics

Objectives :

At the end of this lesson, the student should be able to:

1 Identify the type of variables
2 Interpret frequency distributions
3 Organise the data and represent through various graphical methods
4 Describe measures of central tendency – mean, median and mode
5 Describe measures of variability – range, variance and standard deviation

Topic 1: Descriptive Statistics

1.1.1 Variables

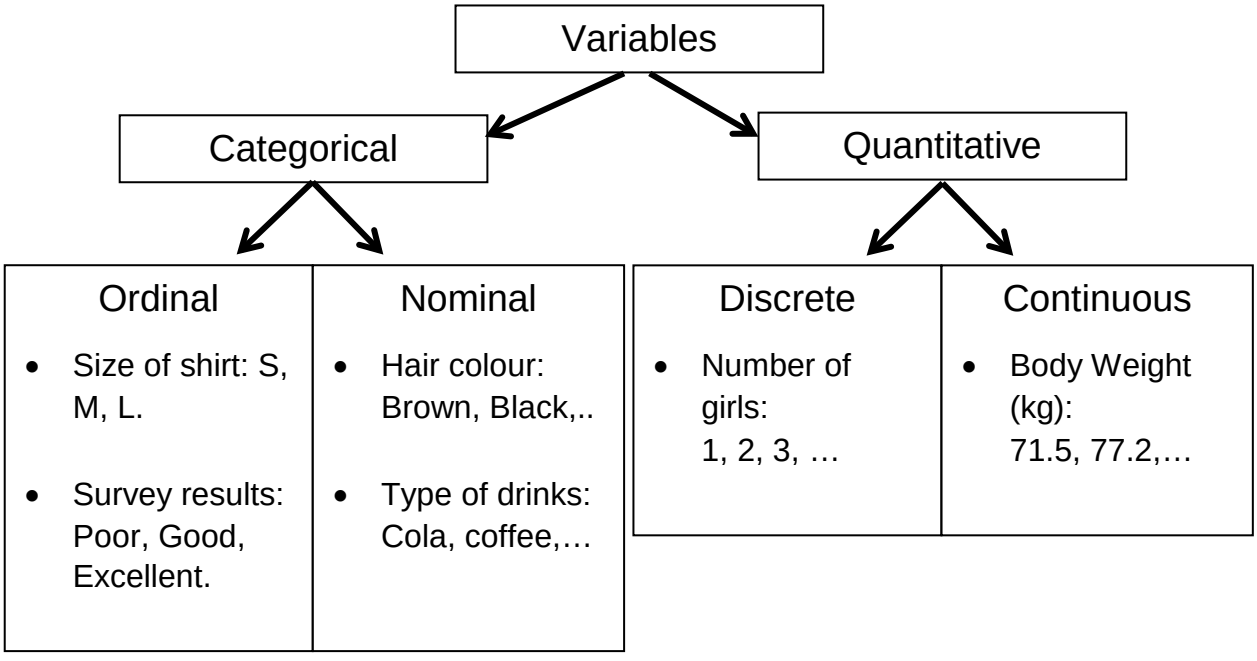
- A variable is the characteristic of an object that can be assigned a value or a category. Variables can be classified as two broad types: **Categorical** and **Quantitative**.

- Categorical variables refer to those that cannot be measured numerically. Quantitative variables are those that can be measured numerically.

- Categorical variables can be further divided in to two sub-categories: **Ordinal** and **Nominal**. Ordinal variables have categories with a natural ranking while nominal variables don't.

- Quantitative variables can be further divided into two sub-categories: **discrete** and **continuous**. Discrete variables have a countable number of values (usually integers) while continuous variables have an uncountable number of values.

The following chart illustrates the relationship between the various types of variables.



1.1.2 Population and sample

- A **population** is a set of **all** the possible observations that can be made. A **sample** is a **subset** of the population.

- Taking samples is usually necessary as it is very tedious to obtain every single data from the population.

- Usually we represent a variable with capital letters such as : X, Y, M,…. We represent observed values of the variable using small letters. For example x = 750.

Example 1.1.2-1

The Manpower Ministry wishes to conduct a survey on a Singapore citizen's monthly salary. A representative from the ministry conducted the survey with 100 people out of 3.34 million Singapore citizens. Let X represents the salary of a Singapore citizen.

_____	is the	**variable of interest**
_____	is the	**observational unit**
_____	is the	sample size
$x = 5000$	is an	_____

1.2 Organising data

- The raw data given will need to be organised neatly before we draw some basic conclusions on the findings. For example, the data below does not show a clear indication on the distribution of the various blood groups.

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

1.2.1 Bar chart, histogram

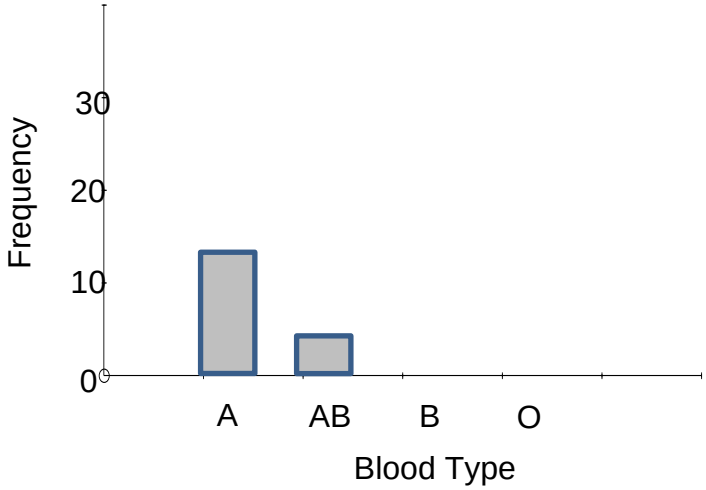
- We can sort the data into two types: **ungrouped** or **grouped**. Ungrouped data is one given as individual data points, while grouped data is one given in intervals.

- A frequency distribution is commonly used to organise data into well specified categories. Grouping of data is usually done when a variable has many different values. We can use a **bar chart** or **histogram** to have a visual view of the data.

Example 1.2.1-1 (Ungrouped data)

Using the data set on blood groups of 50 people above, use the frequency table and complete the bar chart below:

Blood Types	Frequency
A	14
AB	4
B	8
O	24



- To create a histogram for grouped data, we need to calculate the common **interval length** of each class (also known as the "bin width"). To ensure there are **no gaps** between each class, the class intervals will be extended by half a unit of measurement on the left and right.

Example 1.2.1-2 (Grouped data)

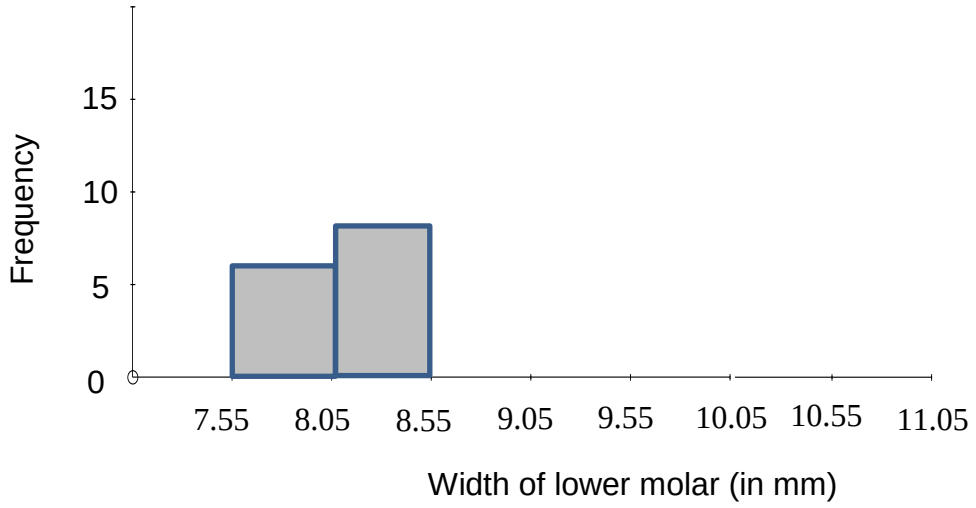
A dentist measured the width (in mm) of the last lower molar of 60 female adult. The results were as follows:

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
8.0	9.3	8.4	9.9	8.7	11.0	8.9	10.0	8.6	8.4

The frequency distribution table is shown below:

Class Interval	Class Boundaries	Class Midpoint	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

Complete the histogram below:



1.2.2 Stem and Leaf Plot

- Stem-and-leaf plot is a method for showing the frequency with which certain classes of values occur. One common approach is to let the last digit be the “leaves” and the remaining digits form the “stems”.

- Unlike the histogram or bar chart, we can still observe the individual data value in the stem and leaf plot.

Example 1.2.2-1

The following are the scores of 20 students on a Statistics test:

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

Construct a stem-and-leaf plot.
Solution:

First we organise the scores in ascending order:

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

The first digit will form the “stems”, while the second digit will form the “leaves”.

1.3 Measures of Central Tendency

1.3.1 Mean

- Given a set of numerical data, $x_1, x_2, ..., x_n$,

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Notation:
  Population mean: μ (data set contains the entire population information).
  Sample mean: $\bar{x}$ (data set contains only a sample's information).

Example 1.3.1-1 (Refer to Appendix 1.2)

Given a sample data set: $2, 5, 7, 10, 11, 13$, calculate the sample mean.
Solution:

1.3.2 Median

- **Median** refers to a value such that **50%** of the data is below this value and **50%** of the data is above this value. It is the middle set of data.

- Suppose our data set has n values.
  Step 1: Arrange the values in **ascending** order.

  Step 2: If n is **odd**, then the median is the $\left(\frac{n+1}{2}\right)^{\text{th}}$ number.

  If n is **even**, then the median is the average of the $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ numbers.

Example 1.3.2-1

For each of the following data set, find its median.

(a) 2, 5, 6, 4, 7, 4, 7, 2, 8, 9, 4, 11, 9, 1, 3

(b) 3, 1, 4, 7, 9, 5, 6, 8, 3, 1, 2, 9, 12, 4, 4, 15

Solution:

a) Arrange the numbers in ascending order:

1, 2, 2, 3, 4, 4, 4, 5, 6, 7, 7, 8, 9, 9, 11

Since there are 15 (odd) data values, median = ____ th value =

b) Arrange the numbers in ascending order:

1, 1, 2, 3, 3, 4, 4, 4, 5, 6, 7, 8, 9, 9, 12, 15

Since there are 16 (even) data values, median is the average of ____th and _____ th values =

1.3.3 Mode

- **Mode** refers to the data value that appears the **most frequent** (highest frequency). A set of data can have more than 1 mode.

Example 1.3.3-1

For each of the following data sets, find the mode.

(a) 2, 5, 6, 4, 7, 4, 7, 2, 8, 9, 4, 11, 9, 1, 3

(b) 2, 3, 5, 9, 6, 4, 7, 4, 7, 2, 8, 9, 2, 9, 1, 3

Solution:

(a) Mode(s) is / are _____

(b) Mode(s) is / are _____

1.4 Measures of Dispersion

1.4.1 Range, Interquartile Range

- Range of a data set = Largest value – Smallest value.

- Quartiles: $Q1$, $Q2$, $Q3$.
  Q1 refers to the value such that **25 %** of the data values are below it. $Q2$ is the **median** of the data and $Q3$ is the value such that **75 %** of the data values are below it.

- Interquartile range = $Q3 - Q1$

- Procedure: Step 1: Find $Q2$ (the median).
  Step 2: Find $Q1$ (the median of the data values below Q2).
  Step 3: Find $Q3$ (the median of the data values above Q2).
  Step 4: Interquartile range = $Q3 - Q1$.

Example 1.4.1-1

The annual profit (rounded to millions of dollars) of 12 randomly selected companies in 2013 are as follows:

8 12 7 17 14 45 10 13 17 13 9 11

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

Solution:
Arrange the numbers in ascending order

7, 8, 9, 10, 11, 12, 13, 13, 14, 17, 17, 45

Range =

Q2 (median) = $\frac{12+13}{2} = 12.5$

Q1 = median of {7, 8, 9, 10, 11, 12} =

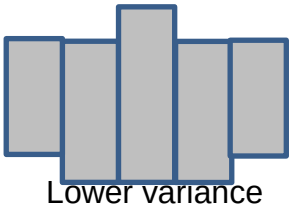
Q3 = median of {13, 13, 14, 17, 17, 45} =

Hence IQR = Q3 – Q1 =

1.4.2 Variance and Standard deviation

- Variance is a measurement of the spread of data, in particular how each data value deviates away from the mean.

Higher variance



Lower variance

- Notation:

  Population variance: $\sigma^2 = \frac{\sum x^2}{N} - \mu^2$, where N is the total population size.

  Sample variance: $s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$, where n is the sample size.

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

**** In this course we will focus on using a scientific calculator to obtain the values of sample mean and variance from a sample data set. ****

Example 1.4.2-1 (Refer to Appendix 1.2)

Consider the sample data marks for a mathematics class test (upon 10)

$$1, 5, 4, 2, 6, 2, 1, 1, 5, 3$$

Calculate the sample standard deviation.

Solution:

Appendix 1.1 Using Excel for Descriptive Statistics

A1.1.1 Create Bar Chart and Histogram

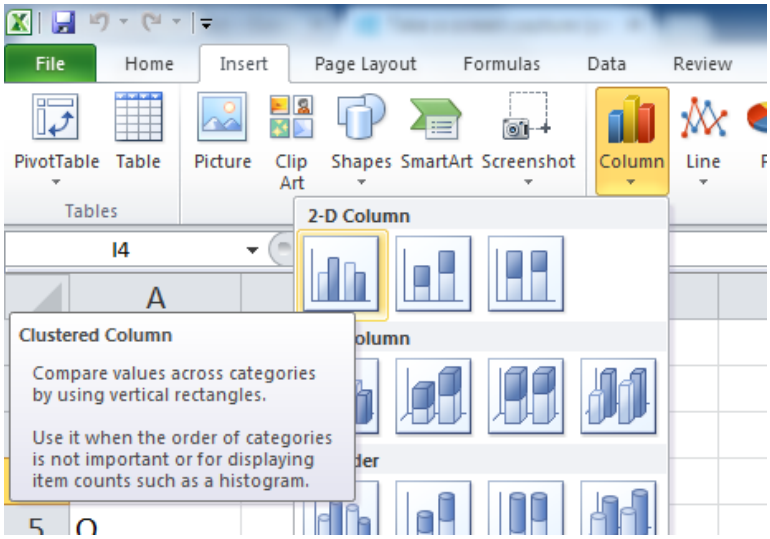
From Excel, go to Options → add-ins →analysis→Analysis TookPak.

- Refer to Example 1.2.1-1 (Bar Chart)

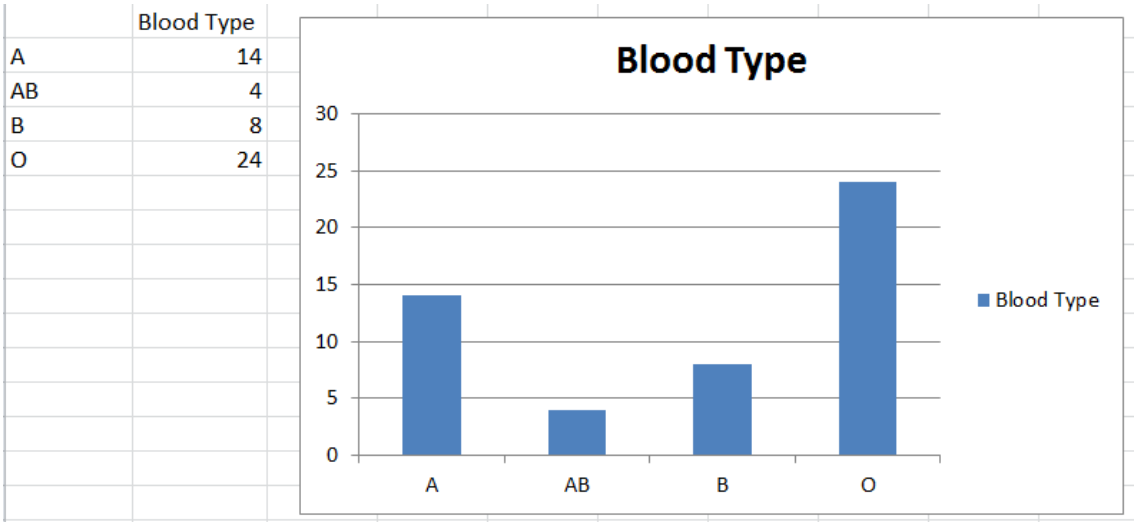
  Step 1: Enter two columns of information, one column for categories and the other for the frequency.

  Step 2: To create a heading for the bar chart, you may key in the heading at the cell above the frequency values' column.

A	B
	Blood Type
A	14
AB	4
B	8
O	24



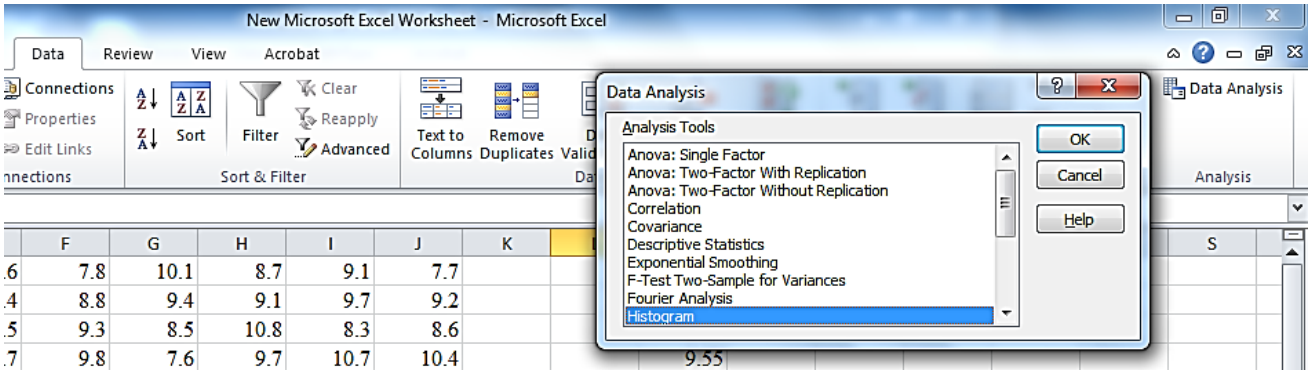
  Step 3: Highlight both columns, proceed to "Insert" tab, select "Column" and "Clustered Column.

- 

Step 1: Copy and paste the data values into an excel spread sheet. Also copy the the upper limit of the class boundaries on a separate column.

Data											bin
7.6	10.6	8.2	10.3	9.6	7.8	10.1	8.7	9.1	7.7		8.05
8.2	9.9	10.9	9.5	10.4	8.8	9.4	9.1	9.7	9.2		8.55
8.7	9.4	9.1	7.9	9.5	9.3	8.5	10.8	8.3	8.6		9.05
10.1	9.8	8.3	10.5	8.7	9.8	7.6	9.7	10.7	10.4		9.55
9.2	9.7	8.6	8.7	8.1	9.2	9.6	10.2	8.9	9.3		10.05
8	9.3	8.4	9.9	8.7	11	8.9	10	8.6	8.4		10.55
											11.05

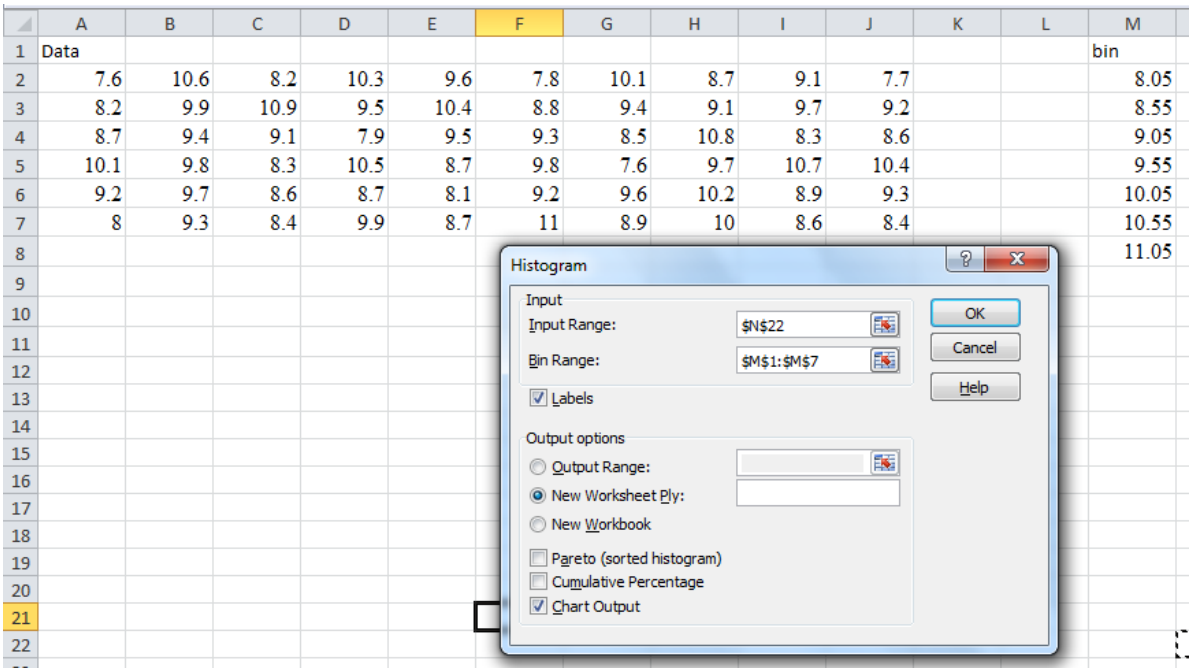
Step 2: Proceed to the “Data” tab, click on the “Data Analysis” option and choose “Histogram”.



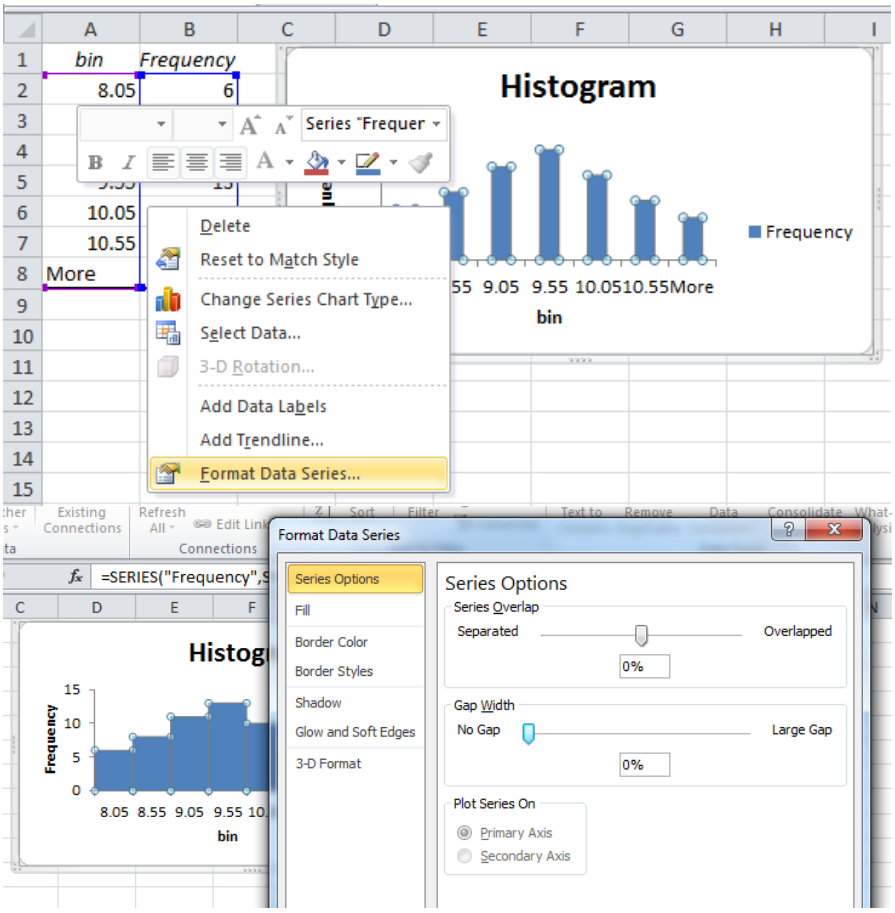
Step 3: Under “Input Range”, highlight all the cells containing the data values.

Under the “Bin Range”, highlight the cells containing the upper limits of the class boundaries (including the header).

Also select the option “Chart Output”.

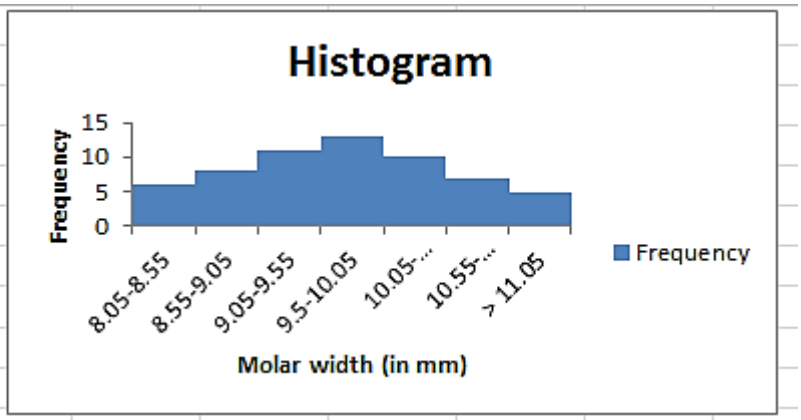


Step 4: Once the diagram is generated, highlight on of the bars and right click to select “Format Data Series”. Select the gap width to 0.



Step 5: You can relabel the class intervals and also the header, etc.

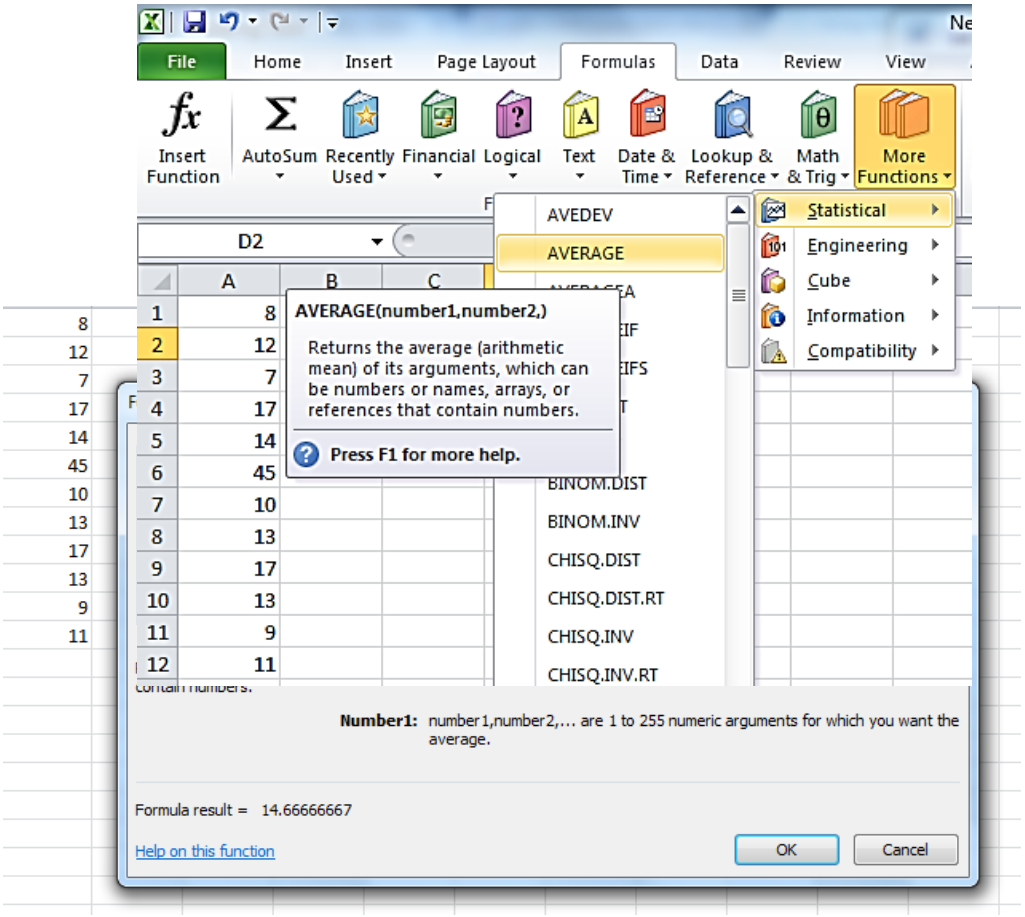
bin	Frequency
8.05-8.55	6
8.55-9.05	8
9.05-9.55	11
9.5-10.05	13
10.05-10.55	10
10.55-11.05	7
> 11.05	5



A1.1.2 Calculate mean, median, standard deviation and quartiles

- Refer to Example 1.4.1-1. To calculate the mean of the data:
  Step 1: Key in the data values in a column and choose the tab “Formulas”, choose “Statistical” and the option “Average”.

  Step 2: Highlight the values to be computed.



- To obtain the median, repeat the same steps as above, choose the function “Median” instead of “Average”.

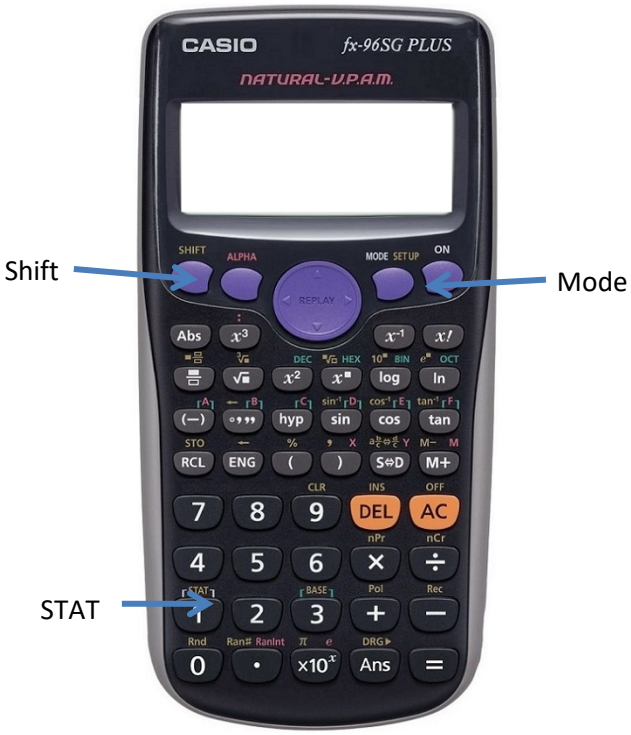
- To obtain the standard deviation, repeat the same steps as above, choose the function “STDEV.P” if the data is the population data or “STDEV.S” if the data comes from a sample.

- To find the quartiles:

  Step 1: Arrange the values in ascending order by highlighting the column of data value, right click and select “Sort” followed by “”smallest to largest”.

  Step 2: Use the “Median” function to find Q_2. Now apply the “Median” function again on the set of data values lower than Q_2 to obtain Q_1. Lastly apply the “Median” function again on the set of data values higher than Q_2 to obtain Q_3.

Appendix 1.2 Using Calculator to obtain mean and standard deviation



Step 1: To enter data into a list:

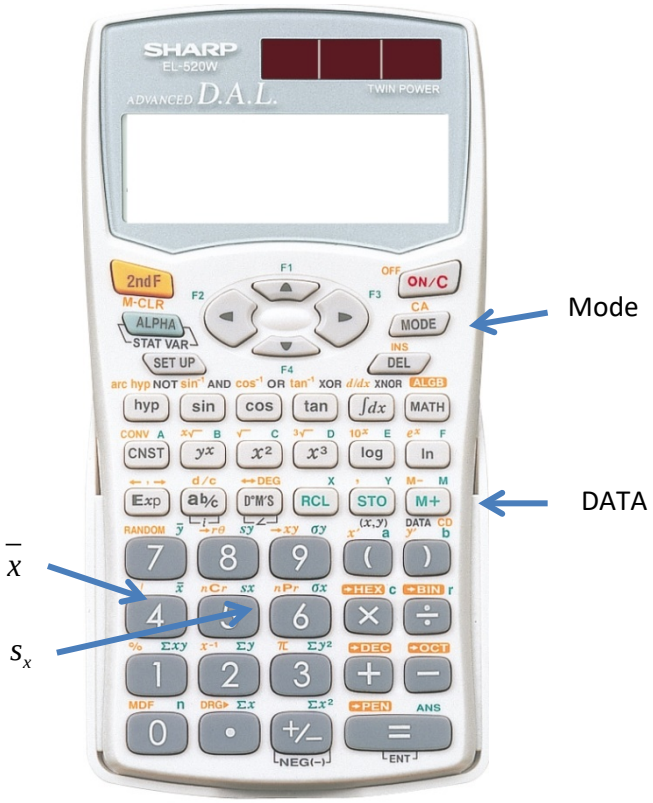
- Mode → 2: STAT → 1: 1 – Var
- Key in a number and press " = " to input the value.
- If frequency list is required:
  Shift → Setup → Down (↓) → 4: Frequency

Step 2: To calculate sample standard deviation

- Exit the data input screen by pressing "AC".
- Press "Shift" → 1: STAT → 4: VAR → 4: s_x

Step 3: To calculate sample mean

- Exit the data input screen by pressing "AC".
- Press "Shift" → 1: STAT → 4: VAR → 2: $\bar{x}$.



Step 1: To key in data in the calculator

- Mode → 1: STAT → 0: SD.
- Key in a number and press " DATA " to store the value in the list.

Step 2: To calculate sample standard deviation and sample mean

- Press "ALPHA" + "5" for sample standard deviation s_x.
- Press "ALPHA" + "4" for sample mean $\overline{x}$.

Tutorial 1: Descriptive Statistics Tutorial

1 Classify each of the variables below by placing a "✓" in the correct categories.

E.g.	Flavour of milk	Categorical ✓	Nominal ✓	Ordinal
		Quantitative	Discrete	Continuous
(i)	Age of a driver (in whole numbers)	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous
(ii)	Gender of a driver	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous
(iii)	Colour of bag	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous
(iv)	Volume of drink	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous
(v)	Size of a shirt (S, M, L, XL)	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous
(vi)	Number of students	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous
(vii)	Examination grades	Categorical	Nominal	Ordinal
		Quantitative	Discrete	Continuous

2 Consider the following 2 sets of sample data

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

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

For each sample, find the following by using the calculator:

(i) mean, (ii) median, (iii) mode,
(iv) interquartile range, (v) range, (vi) standard deviation

3 The daily number of Internet system crashes is observed over 30 days at a university computer centre. The daily Internet system crashes are shown in the table below:

1	3	1	1	0	1	0	1	1	0
2	2	0	0	0	1	2	1	2	0
0	1	6	4	3	3	1	2	4	0

Complete the frequency table and compute its mean, median and mode.

Value, x	0	1	2	3	4	5	6
Frequency							

Answers

1

E.g.	Flavour of milk	Categorical ✓	Nominal ✓	Ordinal
		Quantitative	Discrete	Continuous
(i)	Age of a driver (in whole numbers)	Categorical	Nominal	Ordinal
		Quantitative ✓	Discrete ✓	Continuous
(ii)	Gender of a driver	Categorical ✓	Nominal ✓	Ordinal
		Quantitative	Discrete	Continuous
(iii)	Colour of bag	Categorical ✓	Nominal ✓	Ordinal
		Quantitative	Discrete	Continuous
(iv)	Volume of drink	Categorical	Nominal	Ordinal
		Quantitative ✓	Discrete	Continuous ✓
(v)	Size of a shirt (S, M, L, XL)	Categorical ✓	Nominal	Ordinal ✓
		Quantitative	Discrete	Continuous
(vi)	Number of students	Categorical	Nominal	Ordinal
		Quantitative ✓	Discrete ✓	Continuous
(vii)	Examination grades	Categorical ✓	Nominal	Ordinal ✓
		Quantitative	Discrete	Continuous

2 i) $\bar{x}_A = 9, \bar{x}_B = 8.75$ ii) median $A = 10$, median $B = 9$

iii) mode A = $5, 11$, mode B = 5,

iv) $IQR_A = 7$, $IQR_B = 6.5$,

v) range A = range B = 15,

vi) $s_A = 4.28, \quad s_B = 4.25$

3

Value, x	0	1	2	3	4	5	6
Frequency	9	10	5	3	2	0	1

i mean = 1.43 ii median = 1 iii mode = 1