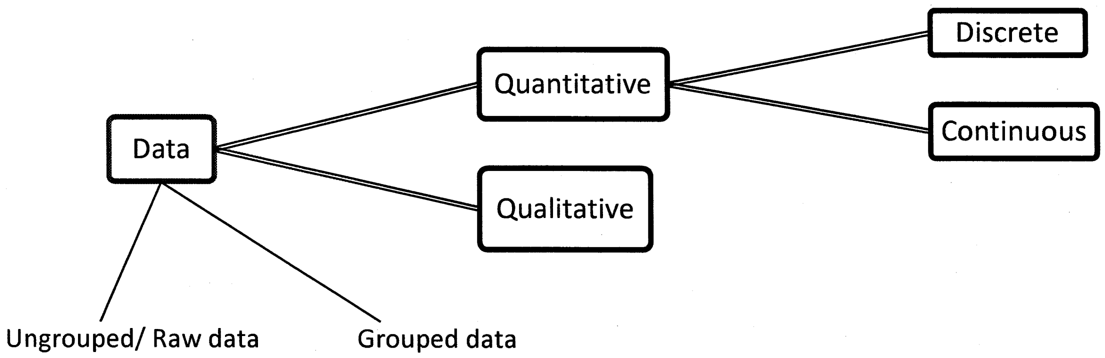
Paper 6 – Topic 1 - REPRESENTATION OF DATA

1. Representation of data

- select a suitable way of presenting raw statistical data, and discuss advantages and/or disadvantages that particular representations may have;
- construct and interpret stem-and-leaf diagrams, box-and-whisker plots, histograms and cumulative frequency graphs;
- understand and use different measures of central tendency (mean, median, mode) and variation (range, interquartile range, standard deviation), e.g. in comparing and contrasting sets of data;
- use a cumulative frequency graph to estimate the median value, the quartiles and the interquartile range of a set of data;
- calculate the mean and standard deviation of a set of data (including grouped data) either from the data itself or from given totals such as Σx and Σx and Σx^2, or $\Sigma(x - a)$ and $\Sigma(x - a)^2$.



Stem and Leaf Diagram (Stemplot)

- One way of arranging the values that gives some information about the patterns within the data.
- A useful way of grouping data into classes while still retaining the original data(advantage).

Example 1 – Below are the marks of 20 students in an assignment.

84	17	38	45	47	53	76	54	75	32
66	65	55	54	51	44	39	19	54	72

Important steps:

1. Lowest value
2. Highest value
3. Equal width for intervals
4. Stem represents ____________ and the leaf represents _________

stem(tens)	leaf(units)

5. The entries in each leaf must be arranged in numerical order.

6. The key is essential in explaining how the stemplot has been formed.

Key 1|7 means 17 mark

Paper 6 – Topic 1 - REPRESENTATION OF DATA

Stem **(tens)**	**Leaf** **(units)**
1	7
2	
3	8
4	5 7
5	
6	
7	
8	4

Stem	**Leaf**	
1	7 9	(2)
2		(0)
3	8 2 9	(3)
4	5 7 4	(3)
5	3 4 5 4 1 4	(6)
6	6 5	(2)
7	6 5 2	(3)
8	4	(1)
		Total 20

Example 2

The lengths, in metres, of 20 measurements in a physics experiment are recorded as follows.

1.78, 1.87, 1.89, 1.72, 1.68, 2.04, 1.96, 1.76, 1.90, 1.73,
1.78, 1.61, 1.78, 1.77, 1.85, 1.65, 1.89, 1.95, 2.01, 1.83

(i) Represent this information on a stem-and-leaf diagram.

(ii) State the mode.

Lengths

16	1 5 8
17	2 3 6 7 8 8 8
18	3 5 7 9 9
19	0 5 6
20	1 4

Key: 17 | 3 means 1.73 m

(ii) The mode is 1.78 m.

Example 3

The maximum temperature in °C, measured to the nearest degree, was recorded each day during June in a particular city. The temperatures were as follows:

19 23 19 19 20 12 19 22 22 16 18 16 19 20 17
13 14 12 15 17 16 17 19 22 22 20 19 19 20 20

Draw a stem-and-leaf diagram to illustrate the temperatures and write down the mode.

Back to Back Stemplot

- can be used to compare two samples by showing the results together on a back to back stemplot.
- A comment is necessary to compare both the samples.
- Two keys are essential.

Example 4

These are the examination marks for French and for English achieved by pupils in a particular class.

French	43	55	29	49	36	55	61	34	42	42	54	60	48	23	44	31	55	45	37	57
English	80	65	74	59	79	92	52	71	43	86	60	74	57	41	79	74	58	52	64	84

(i) Draw a back-to-back stem-and-leaf diagram.

(ii) Compare the two sets of marks.

(i) Examination marks

French		English
	1	
9 3	2	
7 6 4 1	3	
9 8 5 4 3 2 2	4	1 3
7 5 5 5 4	5	2 2 7 8 9
1 0	6	0 4 5
	7	1 4 4 4 9 9
	8	0 4 6
	9	2

Key: 2 | 4 | 1 means 42 for French
41 for English

(ii) The lowest marks are 23 for French and 41 for English. The highest marks are 61 for French and 92 for English.

The marks for English are more spread out (more variable) than the marks for French.

The mode for French is 55 and the mode for English is 74.

It would appear that the pupils performed better in English. However, this would depend on the standards of marking used in the two examinations.

Ways of Grouping Data

- For large sets of data stemplot is not the best method of displaying data.
- The best way is to divide the data into classes.
- The information is more concise than the raw data, but the disadvantage is that the original information has been lost.

Frequency Distribution table

- Shows how many values lie in each class.
- Losses the original values.
- Tally column is optional.
- Class boundaries (upper and lower).
- Class width (equal and unequal).

Paper 6 – Topic 1 - REPRESENTATION OF DATA

Frequency distribution for Discrete Data

This is the frequency distribution for the letters in the solutions of the crossword puzzle.

Number of letters in word	3	5	6	7	9	12	
Frequency	5	6	5	10	2	2	Total 30

Frequency distribution for Continuous Data

The following data were obtained in a survey of the heights of 20 children in a sports club. Each height was measured to the nearest centimetre.

133 136 120 138 133 131 127 141 127 143

130 131 125 144 128 134 135 137 133 129

To form a frequency distribution of the heights of the 20 children, group the information into **classes**

Height (to the nearest cm)	Height (cm)
120–124	$119.5 \leqslant h < 124.5$
125–129	$124.5 \leqslant h < 129.5$
130–134	$129.5 \leqslant h < 134.5$
135–139	$134.5 \leqslant h < 139.5$
140–144	$139.5 \leqslant h < 144.5$

The values 119.5, 124.5, 129.5, ... are called the **class** (or **interval**) **boundaries**. The upper class boundary of one interval is the lower class boundary of the next interval.

Note:

1. Class width = upper class boundary – lower class boundary

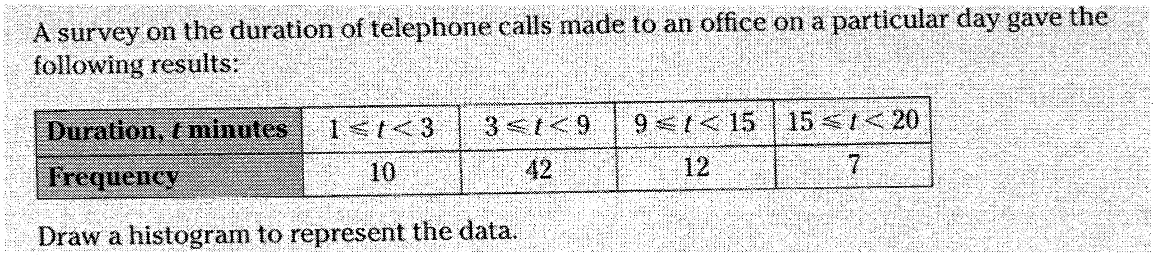
2. mid-point of an interval = (upper class boundary + lower class boundary)/2

3. Modal class – when data have been grouped it is not possible to state the mode. Instead the modal class can be given. The modal class is the interval with the greatest frequency.

Histogram

Grouped data can be displayed in a histogram.

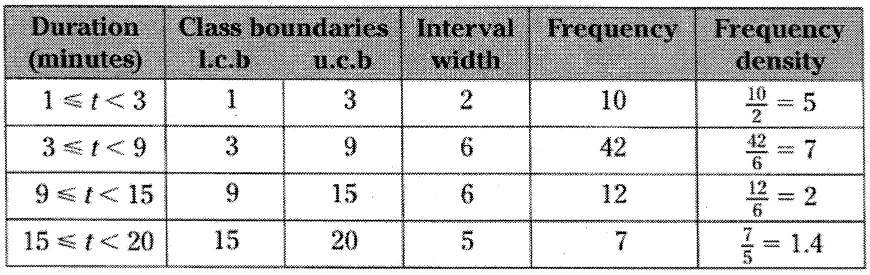
Example 5



A survey on the duration of telephone calls made to an office on a particular day gave the following results:

Duration, t minutes	$1 \leq t < 3$	$3 \leq t < 9$	$9 \leq t < 15$	$15 \leq t < 20$
Frequency	10	42	12	7

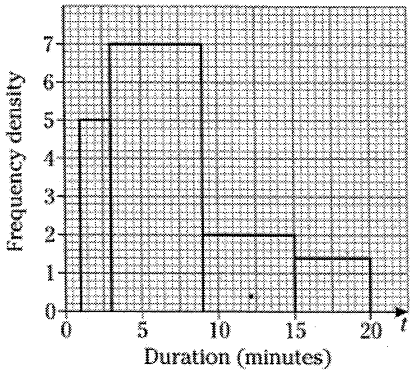
Draw a histogram to represent the data.



Duration (minutes)	**Class boundaries l.c.b**	**u.c.b**	**Interval width**	**Frequency**	**Frequency density**
$1 \leq t < 3$	1	3	2	10	$\frac{10}{2} = 5$
$3 \leq t < 9$	3	9	6	42	$\frac{42}{6} = 7$
$9 \leq t < 15$	9	15	6	12	$\frac{12}{6} = 2$
$15 \leq t < 20$	15	20	5	7	$\frac{7}{5} = 1.4$

$$\boxed{\text{frequency density} = \frac{\text{frequency}}{\text{interval width}}}$$ →

Histogram to show duration of telephone calls



Example 6

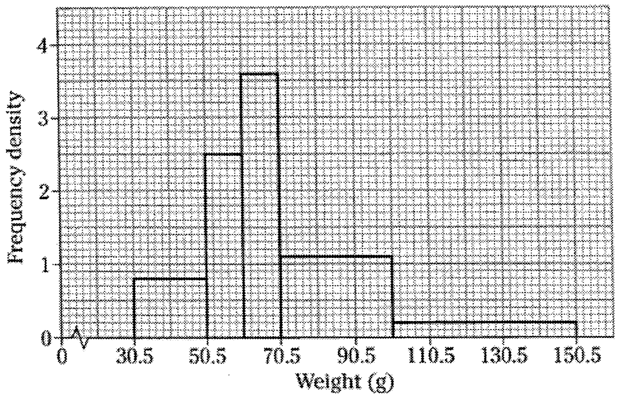
The grouped frequency table records the weights, to the nearest gram, of the letters delivered to an apartment block on a particular day.

Weight (grams)	31–50	51–60	61–70	71–100	101–150
Frequency	16	25	36	33	10

Draw a histogram to illustrate the data and state the modal class.

Weight (nearest gram)	Class boundaries l.c.b	u.c.b	Interval width	Frequency	Frequency density
31–50	30.5	50.5	20	16	$\frac{16}{20} = 0.8$
51–60	50.5	60.5	10	25	$\frac{25}{10} = 2.5$
61–70	60.5	70.5	10	36	$\frac{36}{10} = 3.6$
71–100	70.5	100.5	30	33	$\frac{33}{30} = 1.1$
101–150	100.5	150.5	50	10	$\frac{10}{50} = 0.2$

Histogram to show weights of letters



Remember to include a title.

Remember to label the axes.
Note that
- the vertical axis (frequency density) **must** start from zero.
- the horizontal axis need not start from zero.

HINT: You will find it easier to draw the histogram if you mark the class boundaries (30.5, 50.5, etc.) at the thicker lines on your graph paper.

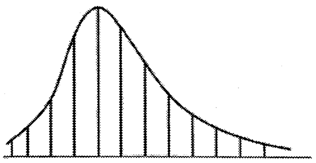
In this example, the interval with the greatest frequency density (represented by the highest bar) is also the interval with the greatest frequency.

The modal class is 61–70.

The shape of a distribution

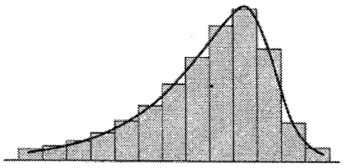
If you superimpose a curve on a histogram or a vertical line graph, it is easier to see the general 'shape' of the distribution.

Positive skew



In a *positively* skewed distribution there is a long tail to the *right* (in the positive direction).

Negative skew



In a *negatively* skewed distribution there is a long tail to the *left* (in the negative direction).

Symmetrical