HISTOGRAMS, MEANS, MODES

RAW DATA

Raw data is information that has been collected and has not been organised in any way.

Example

The scores in a maths test of 50 workshop students are given below. The marks are out of ten.

7	6	5	4	3	5	7	7	4
6	5	4	5	4	5	3	6	2
5	5	7	4	3	4	3	5	1
4	3	6	5	3	6	3	4	4
6	5	5	2	7	5	5	3	4
6	5	8	6	4				

In this form it is very difficult to draw any conclusions about the data, or to compare it with last year's scores for example. One important use of statistics is to organise raw data to make it easier to understand.

FREQUENCY DISTRIBUTION

This is a useful way of organising and presenting raw data and in this example the frequency distribution would show how many students scored 0, how many scored 1, how many 2 etc. The simplest way of doing this is first of all to draw up a tally chart.

Tally Chart

Looking at the raw data we can see that the scores go from 1 (the lowest) to 8 (the highest). The scores are written in the first column of the chart. We then go through the raw data line by line recording each score on the chart with a tally mark. It's a good idea to cross out the figures in the raw data as you go on to prevent counting anything twice.

Tally Chart - Scores in Maths Test

Score	Tally	Frequency (Number of students with that score)
1	1	1
2	11	2
3	~~1111~~ 111	8
4	~~1111~~ ~~1111~~ 1	11
5	~~1111~~ ~~1111~~ 1111	14
6	~~1111~~ 111	8
7	~~1111~~	5
8	1	1
		50 = Total Frequency

To make the tally marks easy to read the fifth in a row goes across the previous four to create a "five-bar gate". When the tally chart is complete we can add up the tally marks to find the frequency (the number of students with each of the scores).

We now have a frequency distribution which gives values of a variable (in this case scores in a maths test) alongside the number of times that value of the variable occurs (the frequency).

GROUPED FREQUENCY DISTRIBUTION

Example

In another maths test the time taken by 30 students to complete the test was recorded (times to nearest minute). These results are shown below.

25	28	10	15	18	23
13	21	25	26	16	17
30	20	16	22	23	26
24	24	29	21	27	28
23	27	25	22	24	23

In this case to give a picture that will be useful it is better to group the data according to time rather than try to give a frequency distribution showing each individual time. This grouped frequency distribution is shown below.

Times Taken by Workshop Students to Complete a Maths Test

Time in minutes (to nearest minute)	**Tally**	**Frequency** (number of students)
10 – 14	11	2
15 – 19	~~1111~~	5
20 – 24	~~1111~~ ~~1111~~ 11	12
25 – 29	~~1111~~ ~~1111~~	10
30 - 34	1	1
		Total **30**

The groups into which the data has been arranged are called **classes** or **class intervals**.

Note

1. How the classes have been arranged. We have 10 - 14, 15 - 19 etc **not** 10 - 15, 15 - 20 etc as that would give an overlap ("15" could come into either class).

2. The data in this example were given to the nearest minute. Any time which was actually recorded as being between 9.5 minutes and 14.5 minutes would fall into the first class of 10 - 14 minutes. This shows us that the dividing line between the first and second class is 14.5 minutes. This divide is called the **class boundary.** The boundaries between the other classes are 19.5, 24.5, 29.5 and 34.5 minutes.

3. The difference between the upper and lower-class boundaries tells us the size of the class or class width. In this case the **class width** is 5 minutes. (19.5 - 14.5 minutes)

Exercise 1

1. The shoe sizes of 54 female students are recorded below. Draw up a tally chart from this data and use it to form a frequency distribution.

5	6	4	6	5.5	4	5	3.5	5
6	5	4	7	3	6	5	6	7
5.5	7	7	5	4	7.5	4	6	5
4	6	4	5	5.5	4	5.5	4.5	5
6	5.5	5	5	6	3	3.5	6	6
4.5	6	4.5	4.5	5	4	4.5	4.5	4

2. How many cigarettes do you smoke per day? The replies given to this question by 105 students are shown below. Draw up a grouped frequency distribution from this data using equal class intervals starting with 0 - 9.

20	0	0	0	20	15	15	15	25	25	23	0	0	25	0
0	30	0	0	0	0	0	15	0	0	50	0	0	0	15
0	0	20	30	3	0	20	0	40	0	0	0	0	0	0
10	0	50	10	45	0	0	0	0	0	20	15	0	0	22.5
15	0	15	0	20	40	0	20	38	20	0	20	0	36	0
0	0	10	0	20	0	0	20	0	40	0	15	0	0	0
0	0	0	0	40	0	40	0	0	45	0	0	0	20	35

DISCRETE AND CONTINUOUS VARIABLES

In this pack we have worked with a number of different **variables**:- scores in a math's test, times taken to do a test, shoe sizes.

A variable that can in theory take any value in a certain range is called a **continuous variable**. For example the "time taken to do a maths test" was measured to the nearest minute but with an accurate clock we could have measured it to seconds, tenths of seconds or even hundredths of seconds. Time is a continuous variable. Other examples of continuous variables would be the length and the weight of pieces of wood and the ages of students.

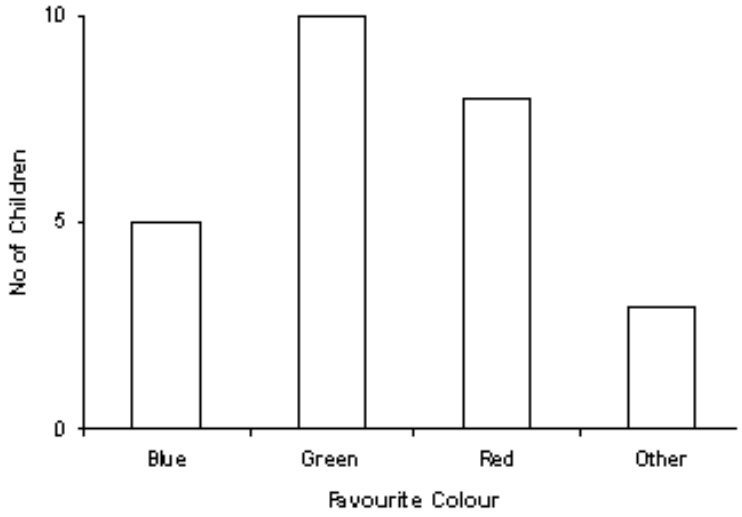
A variable that can only take certain distinct or separate values is a **discrete variable**. For example the number of children in a family or the number of cars in a car park. A discrete variable does not have to be a whole number. For example, shoe size is a discrete variable even though half sizes are included – 3, 3 ½, 4, 4½……

DIAGRAMS TO ILLUSTRATE FREQUENCY DISTRIBUTIONS

1. You have already met one type of statistical diagram, i.e., the **BAR CHART**.

Example

Bar chart to show favourite colour of children in a class.



Notice

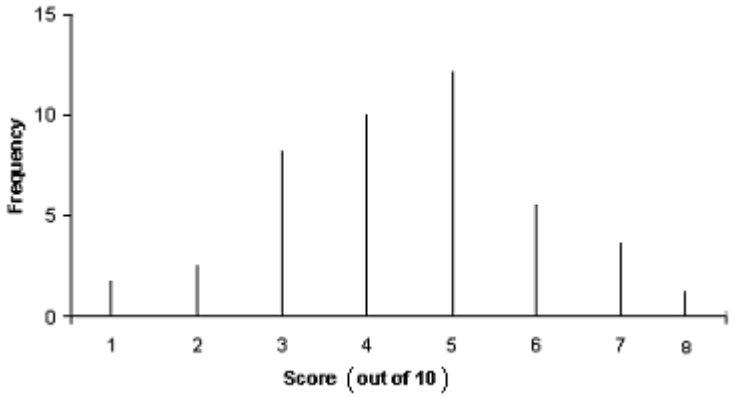
a) The "towers' are **separate** from each other.

b) Descriptive **words** are used along the bottom axis. There is **no horizontal scale.**

c) Frequency (or number of children) goes on the vertical axis and height of bar measures frequency.

This type of diagram is best for **"qualitative"** data which cannot be counted or measured.

2. When illustrating discrete data, it is best to use a **VERTICAL LINE GRAPH**. The information from example 1 could be shown in the chart below.



This is similar to a bar chart in that we have separate "towers" but we now have numbers along the horizontal axis.
If we make the towers slightly fatter this will look just like a bar chart.

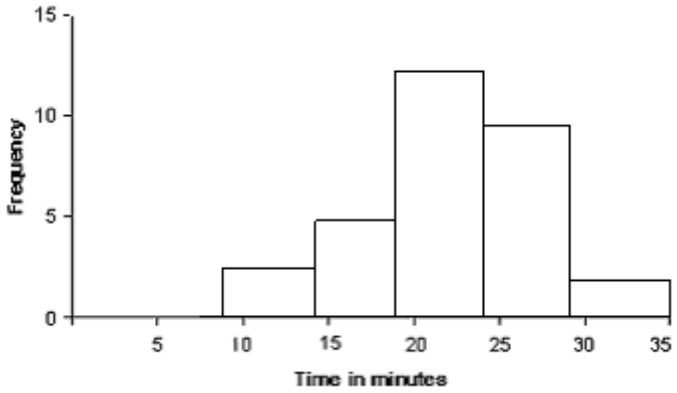
3. For **continuous data** (or discrete data which has been grouped) a **HISTORGRAM** is used.

This may also look at first glance to be similar to a bar chart but notice 2 important points:

1. **Continuous scale along the base line** and

2. "Towers" **touch** each other.

Histogram to show the times taken by students to complete a Maths Test.
(Times to nearest minute)



*Notice how the bars start at 9.5, 14.5, 19.5 etc.

Because we have a frequency distribution of a continuous variable the dividing lines between the rectangles of the histogram come on the class boundaries.

Here are some notes to help you to draw histograms.

1. Form a frequency distribution (if one is not already given).

2. If there are any open-ended intervals decide what the limits should be to make all classes of equal size.

3. Use graph paper. Draw and label the axes and mark in the scales. Mark frequency density on the vertical axis. On the horizontal axis mark the scale and state the units.

4. Decide where the class boundaries will fall. Although the scale may be marked 10, 20, 30 the boundaries may be on 9.5, 19.5, 29.5.

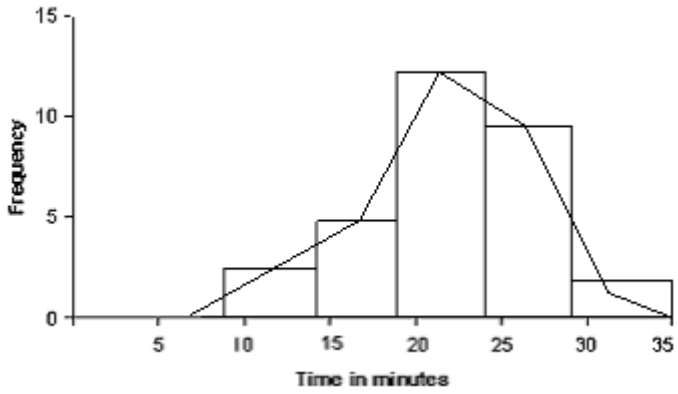
Exercise 2

1. Draw a vertical line graph from the data in question 1, Exercise 1.
2. Draw a histogram using the information given in question 2, Exercise 1.

FREQUENCY POLYGONS

A frequency polygon is formed from a histogram by joining the mid-points of the tops of the rectangles by straight lines.
The tops of the first and last rectangles are joined to the axis to either side at the points where the mid-points of the next class intervals would be.
Look again at example 2 of Histograms.

Frequency Polygon of the Times Taken by Students to Complete a Maths Test



MEAN AND MODE

Average is a word used in common English. An average is used to give quick information about a group. It is a single value which is used to represent all the values of a distribution. We may talk of: the average height of 12-year-old boys:

- the batting average of a cricketer
- the average wage
- the average day-time temperature.

There are **three** types of **averages:**

1. The arithmetic **mean** (often called the mean)
2. The **mode**
3. The **median**

The mean is exactly the same as the everyday "average" mentioned above. That and the mode are to be dealt with in this pack. The median is covered in pack MS6.

The Arithmetic Mean

The arithmetic mean of a set of data is found by adding up all the values in the set and dividing this total by the number of values.

$$\text{Arithmetic Mean} = \frac{\textbf{Sum of all the values}}{\textbf{The number of all the values}}$$

Example 1

Four children are aged 3,4,7 and 6 years. What is the mean value of these children's ages.

SUM OF AGES = 3 + 4 + 7 + 6 = 20 years
NUMBER OF AGES = 4

MEAN VALUE IS $\frac{20}{4}$ = 5 years

Sometimes we need to know the total of the values when we have been given the mean and the number of values.

SUM = MEAN x NUMBER OF VALUES

It is very important that you understand questions of the following type.

Example 2

A cricketer has a mean score of 34 runs for his eight innings of the season. How many runs must he score in his ninth innings in order to increase his mean score to 37?

Total runs scored in first eight innings = 8 x 34 = 272 runs

Total runs which must be scored in first nine innings = 9 x 37 = 333 runs

Therefore, the cricketer must score 333 - 272 = 61 runs in the ninth innings to increase his mean score to 37.

The Mean of a Frequency Distribution

Example 3

The scores in a test of 50 workshop students (marks out of ten), were:

7	6	5	4	3	5	7	7	4
6	5	4	5	4	5	3	6	2.
5	5	7	4	3	4	3	5	1
4	3	6	5	3	6	3	4	4
6	5	5	2	7	5	5	3	4
6	5	8	6	4				

The mean of this data could be found by adding up all the scores and then dividing this sum by 50. This working is made much simpler if the raw data is first organised into a frequency distribution.

Score	**Frequency**	**Score x Frequency**	
1	1	1 x 1	1
2	2	2 x 2	4
3	8	3 x 8	24
4	11	4 x 11	44
5	14	.	70
6	8	.	48
7	5	.	35
8	1	.	8
	Total = 50		Total = 234

The total of all the scores is then found by adding up the last column since 1 person scored 1, 2 people scored 2, 8 people scored 3 etc.

Therefore, **Mean Score** $= \dfrac{\textbf{Sum of all the scores}}{\textbf{The number of students}} = \dfrac{234}{50} = 4.68$

Statistics has its own special language and a shorthand is usually used for these totals. The variable (in this case test score) is x
The frequency (in this case the number of students with that score) is f.
Variable multiplied by frequency is fx

The total frequency (the total number of students) is Σf.

(Σ (say "sigma") is the Greek letter S used to stand for **SUM** or **TOTAL** so Σf ("sigma f") means total frequency).

Σfx means multiply f by c for each value and then add all these (fc) together.

Here is another example to show you how these symbols are used.

Example 4

The shoe sizes of a group of men are shown below. Find the mean shoe size of this group.

Male Shoe Size

Shoe Size (x)	**Frequency (f)**	**f. x.**
6	1	6
6.5	1	6.5
7	10	70
7.5	1	7.5
8	15	120
8.5	7	59.5
9	7	63
9.5	0	0
10	3	30
10.5	1	10.5
11	3	33
11.5	1	11.5
	Σf=50	Σfx =417.5

Mean shoe size $= \dfrac{\sum \mathbf{f}x}{\sum \mathbf{f}} = \mathbf{8.35}$

NOTE

1. The mean shoe size is not an exact shoe size. This is usually the case when we work with discrete data.

2. A calculator makes the working out of statistics problems much easier and quicker. Use one now so that you are confident in its use for the exam.

Finding the Mean of a Grouped Frequency Distribution

When presented with grouped data we have to use the mid-point of the interval to stand for the whole group. Then the method is the same.

Example 5

The heights of 80 children in a first school were found and grouped together as shown below. Calculate the mean height.

Height (cm)	**Frequency**	**Mid-point**	
	f	x	fx
80 - 89	2	84.5	169
90 - 99	10	94.5	945
100 - 109	16	104.5	1672
110-119	24	114.5	2748
120 - 129	16	124.5	1992
130 - 139	11	134.5	1479.5
140 - 149	1	144.5	144.5
	Σf = 80		Σfx = 915

Mean height = $\frac{\sum \mathbf{f}\chi}{\sum \mathbf{f}} = \frac{\mathbf{9150}}{\mathbf{80}} = \mathbf{114.375\ cm}$

= **114.4 (to 1 dp)**

Exercise 3

1. Find the mean of £43, £51, £54, £47, £49.

2. 5 parcels weigh 84 kg, 3 weigh 76 kg and 2 weigh 88 kg. What is the average weight of these10 parcels.

3. The marks obtained by a group of 60 students in an examination were as follows:

Marks	20 – 29	30 - 39	40 - 49	50 – 59	60 - 79
No of students	4	9	25	19	3

Find the mean mark.

The Mode of a Frequency Distribution

The mode of a frequency distribution is the most commonly occurring value.

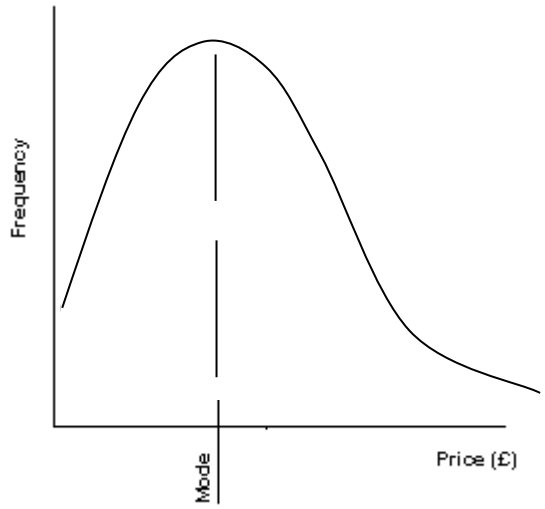
Example 1

Take a set of prices - £12, £13, £15, £13, £12, £13, £11.

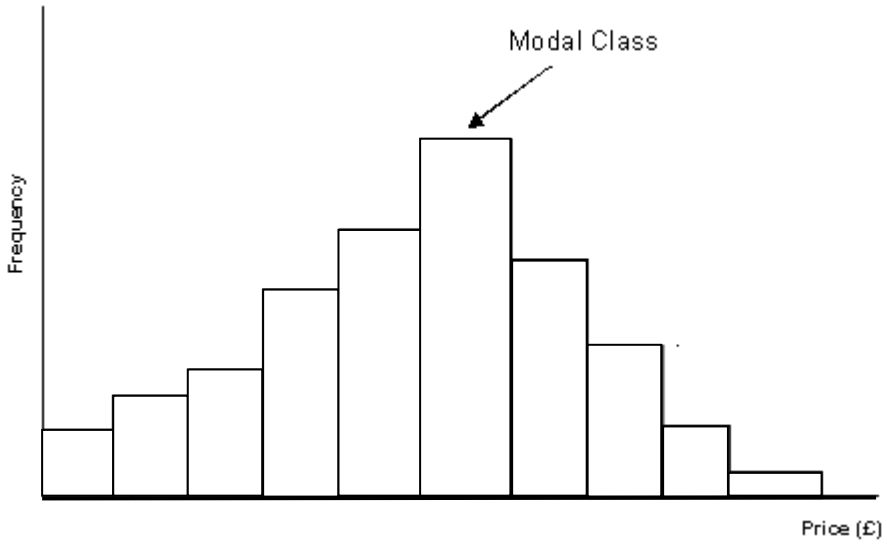
The **mode** of this set is **£13** as it occurs three times in the set.

The mode can be obtained directly from a frequency curve or from the histogram of a distribution.

Frequency Curve



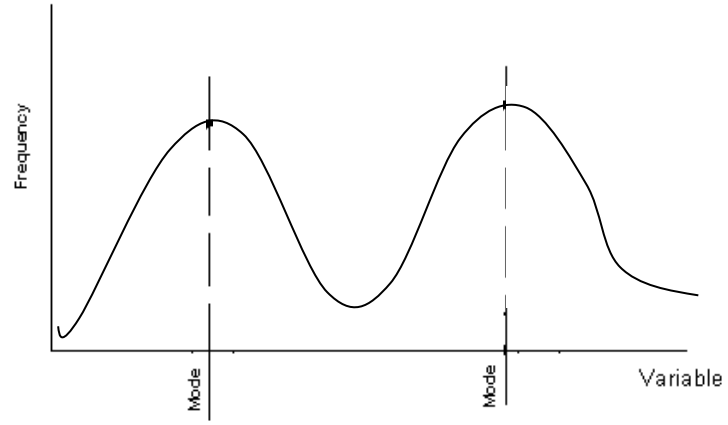
Histogram



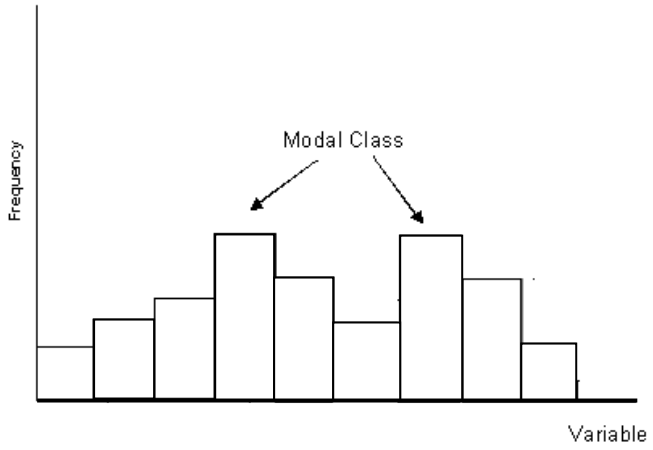
For grouped data we can pick out a modal class.
It is possible for a distribution to have more than 1 mode.

For example, if a frequency curve has more than one peak or histogram has classes with equally large frequency, the diagrams will appear as follows:

Example 2



Example 3



Example 4

Find the mode of the following numbers

1, 2, 7, 5, 2, 1, 5

The modes are 1, 2 and 5 (they all occur twice).

The Different Statistical Averages

In this pack you have met two statistical averages:- the arithmetic mean and the mode.

Which average to use in any particular case depends on a number of factors. Each has advantages and disadvantages.

The arithmetic mean is the most well known and most used average but it has a serious disadvantage when used for distributions with a small number of extreme values.

For example. The marks of 5 students were 15, 14, 11, 14, 46. The arithmetic mean of the marks is 20 (100 + 5). This is obviously affected by the extreme value of 46 and may not give the true impression of the level of marks. In this case it may be that 14 (the mode) is a better average to use.

A shoe shop will not find it useful to know the mean shoe size of ladies' feet are, for example, 6.21. It will be far more valuable to know that the mode is, for example, size 7, as this tells the shop keeper that he or she can expect to sell more shoes of this size than any other.

Exercise 4

1. Look at the distribution of marks in example 3 Exercise 3. What is the modal class of this distribution. Draw the histogram.

ANSWERS

Exercise 1

1.

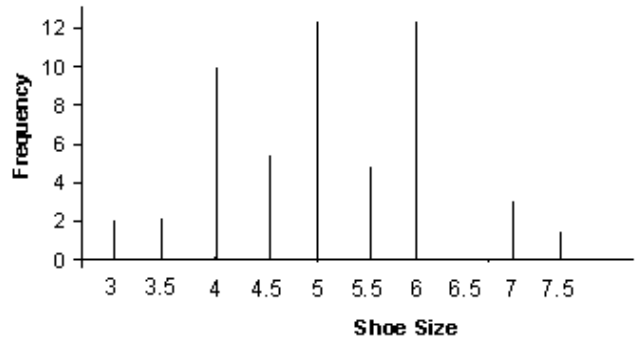
Shoe size	Tally	Frequency
3	11	2
3.5	11	2
4	𝍸 𝍸	10
4.5	𝍸 1	6
5	𝍸 𝍸 11	12
5.5	𝍸	5
6	𝍸 𝍸 11	12
6.5		0
7	1111	4
7.5	1	1

2.

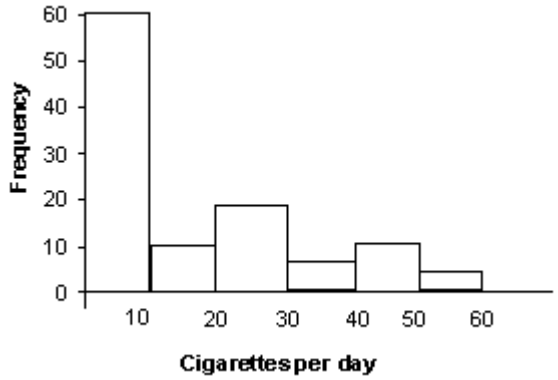
Cigarettes per day	Tally	Frequency
0-9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 11	62
10-19	𝍸 𝍸 11	12
20- 29	𝍸 𝍸 𝍸 11	17
20 – 39	𝍸	5
40 - 49	𝍸 11	7
50 - 59	11	2

Exercise 2

1.



2.



Exercise 3

1. $\text{Mean} = \frac{£224}{5} = £48.80$

2. Total Weight = 5 + 84 = 420
   3 x 26 = 228
   2 x 88 = 176
   824 Kg

   $\text{Mean} = \frac{824}{10} = 82.4 \text{ kg}$

3.

	MARK (x)	FREQUENCY (f)	MID POINT f(x)
20-29	24.5	4	98
30-39	34.5	9	310.5
40-49	44.5	25	1112.5
50-59	54.5	19	1035.5
60-79	69.5	8	208.5
		Σf =60	

$\text{Mean} = \frac{2765}{60} = 46.08$ **NOTE** The class interval (60-79) is wider than the others.

Exercise 4

Modal Class = (40-49)