Section 2: Data presentation and interpretation

This section deals with

- Types of data
- Frequency tables
- Stem-and-leaf diagrams
- Grouping data
- Bar charts, pie charts & vertical line charts
- Histograms

Much of the work in these notes will be familiar to you from G.C.S.E.

Types of data

Before deciding on how to present data, it is important to consider what sort of data you are dealing with. Here are some examples.

Coin tosses:
{H, H, T, H, T, T, H, T, H, H, T, H, H, H, T, T, H, H, T, T}

Dice throws:
{6, 1, 2, 5, 5, 2, 5, 6, 1, 2, 2, 4, 4, 5, 6, 2, 5, 3, 3, 2}

Heights:
{1.85, 1.78, 1.65, 1.70, 1.66, 1.85, 1.80, 1.77, 1.67, 1.73,
1.82, 1.88, 1.73, 1.71, 1.68, 1.90, 1.79, 1.82, 1.65, 1.70}

Number of coin tosses until a Head is obtained:
{1, 3, 2, 2, 1, 5, 2, 1, 3, 1, 2, 2, 6, 3, 2, 2, 1, 5, 8, 1}

Marks awarded by judges at a world champion skating event:
{5.9, 5.8, 4.9, 5.2, 5.7, 5.6}

Judgement of level of pain from 120 patients in clinical trial of a drug:
{none, mild, acute, mild, moderate, none, mild, mild, moderate, moderate, mild, mild, moderate}

Edexcel AS Maths Data 2 Notes and examples

Categorical and numerical data

You can first classify these sets as either **categorical** or **numerical**:

Categorical	Numerical
Coin tosses Pain levels	Scores when dice are thrown Heights Number of coin tosses until it lands 'heads' Judges' marks in an ice-skating competition

Categorical data need not be expressed in numbers. They are usually given as categories such as heads or tails, pain level, gender, eye colour, car type, etc.

Numerical data are expressed as numbers and the values of these numbers have a numerical meaning.

Beware: numbers can be categorical data if they do not have a numerical meaning. Examples are numbered menu items, such as you might see in a Chinese restaurant, or the numbers on rugby shirts. It would make no sense to find the mean value of the numbers of the menu items ordered in a Chinese restaurant!

Notice that the pain levels could be given a numerical value, e.g. mild 1, moderate 2, etc. This would convert the data to numerical data, which could allow calculations to be made, e.g. mean pain level. However, this is not good practice because the numbers are subjective; one person's judgement of the crossover point between moderate and acute pain is unlikely to be the same as another's. Assigning numerical values to such subjective judgements make the data seem more accurate than they really are and can produce misleading results.

Discrete and continuous data

Numerical data can be further divided into **discrete** and **continuous**.

If all the possible values for the data can be listed, the data is discrete.

Numerical data	
Discrete	**Continuous**
Number of coin tosses until it lands 'heads' Judges' marks in an ice-skating competition Scores from throwing dice	Heights

You might think that the ice-skating scores are continuous, but in fact they are discrete, because we can list the possible scores which can be given: 6.0, 5.9, 5.8, 5.7, 5.6, 5.5, … etc. With continuous data, this is not possible – not only are there an infinite number of possible heights, but they also can't be listed without missing some out.

Notice that it is not simply a question of whether the set of possible values for the data is finite or infinite: the number of throws until a Head is tossed could be any whole number, and so this set is infinite but discrete. Also, it is not a question of whether the data consist of whole numbers. The skating data are decimal but discrete.

In practice, all continuous data have to be rounded – the heights given above are all given to 3 significant figures. Once rounded, the set of possible values is in fact finite and listable:

…, 2.00, 1.99, 1.98, 1.97, 1.96, … etc.

Nevertheless, these data are clearly measuring a continuous quantity, and are therefore regarded as continuous rather than discrete.

Example 1
Decide whether each of the following sets of data is categorical or numerical, and if numerical whether it is discrete or continuous.

A Cards drawn from a set of playing cards:
{2 of diamonds, ace of spades, 3 of hearts etc…}
B Number of aces in a hand of 13 cards:
{1, 2, 3, 4}
C Time in seconds for 100 metre sprint:
{10.05, 12.31, 11.20, 10.67, 11.56, …etc}
D Number of weeds in a 1 m by 1 m square in a biology experiment:
{2, 8, 12, 3, 5, 8, …}
E Number of spectators at a football match:
{23 456, 40 132, 28 320, 18 214, …etc}
F Day of week when people were born:
{Wednesday, Monday, Sunday, Sunday, Saturday, etc…}
G Times in seconds between 'blips' of a Geiger counter in a physics experiment:
{0.23, 1.23, 3.03, 0.21, 4.51, …etc}
H Percentages gained by students for a test out of 60:
{20, 78.33, 80, 75, 53.33, …etc}

Solution
A and F are categorical data, all the others are numerical.
B - discrete
C - continuous
D - discrete, as there must be a whole number of weeds.
E - discrete
G - continuous
H - discrete, as there are only 60 possible percentage scores.

Edexcel AS Maths Data 2 Notes and examples

Frequency tables

When data contains items which are repeated, it makes sense to use a frequency table to record them.

Example 2
The data set below shows the scores when a die was thrown repeatedly.

{6, 1, 2, 5, 5, 2, 5, 6, 1, 2, 2, 4, 4, 5, 6, 1, 4, 3, 3, 2}

Show this data in a frequency table.

Solution
There are three 1s, five 2s, two 3s, three 4s, four 5s, three 6s. In a frequency table:

Score	Frequency
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

Always add up the frequencies to check that this is the same as the number of data items.

Grouping data

Grouped frequency tables are used when the data are widely spread. Consider the following data on spectators at football matches:

23456	40132	28320	18214	12250	13302	17359	18605
14567	16321	16002	19925	20451	18491	22902	19358
18314	21359	32304	22561	12912	25600	28614	10540
17312	27347	29902	41354	38401	16592	18610	15482
34012	22782	38427	15384	18921	16349	26210	8265

Only when the data is grouped does it start to make some sense:

No. of spectators	Tally	Frequency
0 – 10000	𝍷	1
10000 – 20000	𝍸 𝍸 𝍸 𝍸 𝍷𝍷𝍷	23
20000 – 30000	𝍸 𝍸	10
30000 – 40000	𝍷𝍷𝍷𝍷	4
40000 – 50000	𝍷𝍷	2
Total		40

Immediately, you can see that most of the crowds are between 10000 and 30000, with crowds below 10000 and over 30000 unusual.

Are the groups above well defined? Where would you place a crowd of 10000? For discrete data, the groupings should not overlap, so it would be better to group as follows:

No. of spectators	Tally	Frequency
0 – 9999	|	1
10000 – 19999	𝍸 𝍸𝍸𝍸𝍸 |||	23
20000 – 29999	𝍸 𝍸	10
30000 – 39999	||||	4
40000 – 49999	||	2
Total		40

With continuous data, the overlap problem does not really apply, as in theory it should be possible to decide whether continuous data lies above or below any class boundary. In practice, the data have been rounded, so you clarify which group to place data in using ≤ or < symbols, as shown in the following example:

1.85 1.78 1.65 1.70 1.66 1.85 1.80 1.77 1.67 1.73
1.82 1.88 1.73 1.71 1.68 1.90 1.79 1.82 1.65 1.70

So 1.70 goes in this class

Height h	Frequency
$1.65 \le h < 1.70$	5
$1.70 \le h < 1.75$	5
$1.75 \le h < 1.80$	3
$1.80 \le h < 1.85$	3
$1.85 \le h < 1.90$	3
$1.90 \le h < 1.95$	1
Total	20

The classes are sometimes presented like this:

Height h	Frequency
1.65 –	5
1.70 –	5
1.75 –	3
1.80 –	3
1.85 –	3
1.90 –	1
1.95 –	0
Total	20

How do you decide how to group? This depends on the amount of data you have to group, and how many classes you want to end up with. Look at the football crowd data again.

If the number of classes is too large for the amount of data you have, then the frequencies are too small to build up an idea of the ‘shape’ of the distribution:

Class	Frequency	Class	Frequency	Class	Frequency
5000 - 5999	0	20000 - 20999	1	35000 - 35999	0
6000 - 6999	0	21000 - 21999	1	36000 - 36999	0
7000 - 7999	0	22000 - 22999	3	37000 - 37999	0
8000 - 8999	1	23000 - 23999	1	38000 - 38999	2
9000 - 9999	0	24000 - 24999	0	39000 - 39999	0
10000 - 10999	1	25000 - 25999	1	40000 - 40999	1
11000 - 11999	0	26000 - 26999	1	41000 - 41999	1
12000 - 12999	2	27000 - 27999	1	42000 - 42999	0
13000 - 13999	1	28000 - 28999	2	43000 - 43999	0
14000 - 14999	1	29000 - 29999	1	44000 - 44999	0
15000 - 15999	2	30000 - 30999	0	45000 - 45999	0
16000 - 16999	4	31000 - 31999	0	46000 - 46999	0
17000 - 17999	2	32000 - 32999	1	47000 - 47999	0
18000 - 18999	6	33000 - 33999	0	48000 - 48999	0
19000 - 19999	2	34000 - 34999	1	49000 - 49999	0

On the other hand, if there are too few classes, grouping becomes too crude, and we lose detail:

Class	Frequency
0 - 19999	22
20000 - 39999	16
40000 - 59999	2
Total	40

The best choice of class intervals gives enough detail to get a feel for the distribution:

Class	Frequency
0 - 4999	0
5000 - 9999	1
10000 - 14999	5
15000 - 19999	16
20000 - 24999	6
25000 - 29999	6
30000 - 34999	2
35000 - 39999	2
40000 - 44999	2
Total	40

Stem-and-leaf diagrams

The problem with grouping data is that the raw data is lost – all you know is the class each item of data lies in. A stem-and-leaf diagram groups the data but keeps the raw data intact.

This data set is the heights of a group of 20 ‘A’ level students.
{1.85, 1.78, 1.65, 1.70, 1.66, 1.85, 1.80, 1.77, 1.67, 1.73, 1.82, 1.88, 1.73, 1.71, 1.68, 1.90, 1.79, 1.82, 1.65, 1.70}

The stems are 16, 17, 18, 19. The second decimal places are used as the leaves.

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

Once the leaves are placed, they should be ordered:

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

$n = 20$
17|3 represents 1.73

A key needs to be added to show the size of the data set and the place value of the data

You may want to group the data in a different way, as in the diagram below.

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

$n = 20$
17| 3 represents 1.73

Two sets of data can be compared by means of a side-by-side stem plot:

Girls (n = 18)		Boys (n = 20)
8 7 3	15	
7 4 3 3 0	16	5 5 6 7 8
9 8 7 5 0	17	0 0 1 3 3 7 8 9
6 3 3 1 0	18	0 2 2 5 5 8
	19	0

17|3 represents 1.73

This diagram packs lots of statistical punches:

- The raw data is preserved
- The data is ordered, making it easy to find the median and quartiles
- The length of the lines of leaves gives the shape of each distribution
- Comparing these enables us to compare the distributions

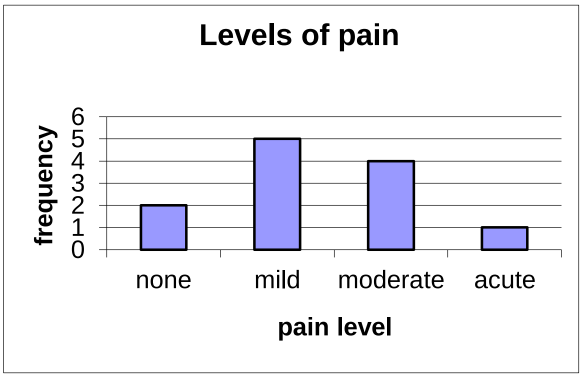
In the above, it is clear that the boys are on average taller than the girls. However, the spread of the boys' heights and the girls' heights appear to be similar.

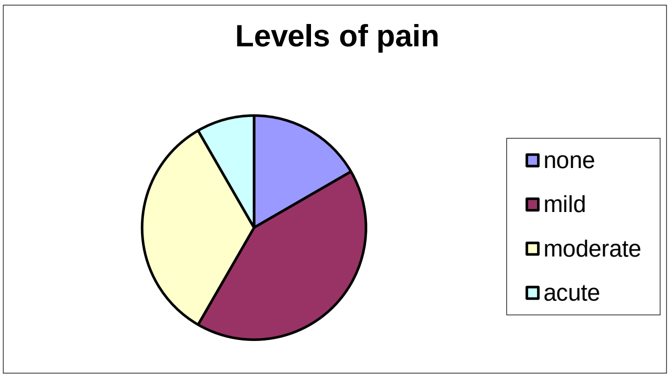
Bar charts, pie charts and vertical line charts

Categorical data is best presented using either a pie chart or a bar chart. Which is better depends on what features you want to highlight.

E.g. Data from a drug trial: Judgement of level of pain from 12 patients in the clinical trial of a pain-killing drug
{none, mild, acute, mild, moderate, none, mild, mild, moderate, moderate, mild, moderate}
As a frequency table:

Level of pain	Frequency
none	2
mild	5
moderate	4
acute	1





Compare the bar chart with the pie chart:

- The bar chart compares the frequencies – you can easily read these off the chart
- The pie chart compares proportions, but obscures the individual frequencies

The pie chart automatically scales the data to fractions of 360°. This is an advantage when you want to compare two data sets of different sizes.

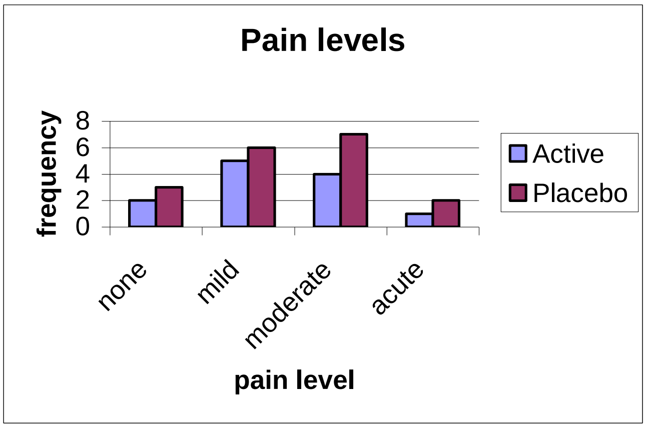
Edexcel AS Maths Data 2 Notes and examples

A placebo drug has no active ingredient, but is used as a control in drug trials – everyone who takes medication tends to feel better psychologically even if the drug has no therapeutic effect.

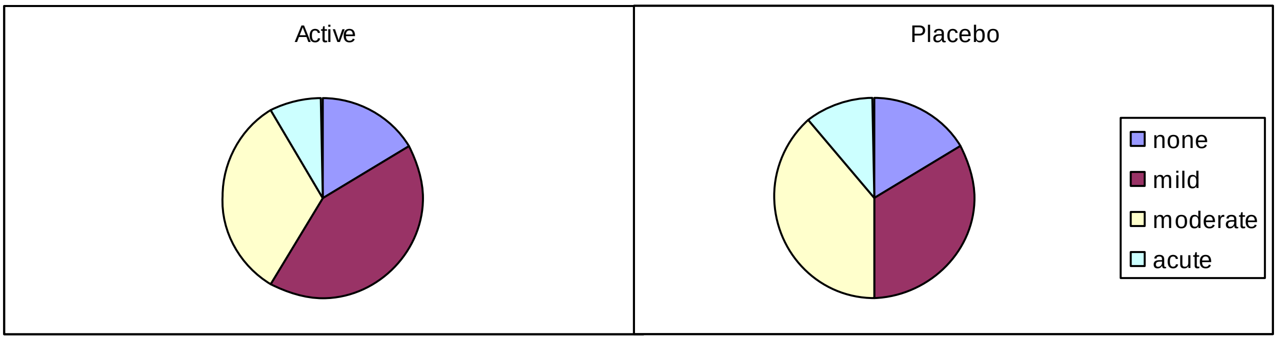
Suppose the placebo drug gave the following results:

Level of pain	Frequency
none	3
mild	6
moderate	7
acute	2

The two sets of data could be compared using a comparative bar chart:



or as two pie charts:



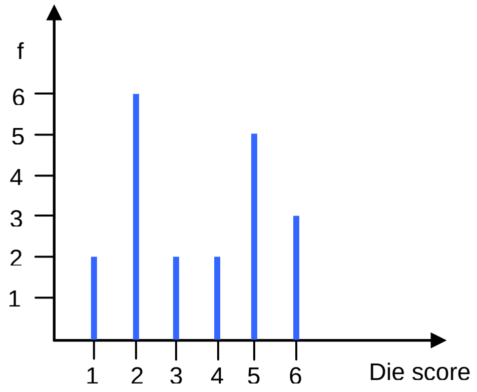
It is not easy to draw conclusions from the comparative bar chart, because the amount of data is different for each treatment group. However, the pie charts are automatically scaled to 360°, making it easier to compare. It looks like the drug is a little better than the placebo, although the data sets are on the small side.

Vertical line charts

Discrete data can be displayed using a pie chart or a bar chart. However, a bar chart can be replaced by a vertical line chart, which is perhaps more appropriate because the width of the bar on a bar chart can easily be misinterpreted as representing a range of values when, in fact, it only represents one discrete value.

E.g.

Die score	Frequency
1	2
2	6
3	2
4	2
5	5
6	3

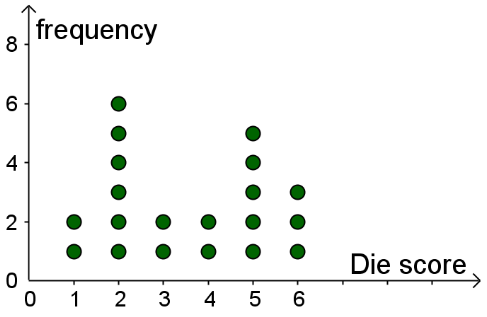


The vertical line chart emphasises that the scores are discrete and so can take no 'in between' values

Dot plots

A dot plot is similar to a vertical line chart but instead of lines you use a stack of dots. One dot might represent one data item, or with larger data sets one dot might represent, say, 10 items.

The dot plot below shows the same data as for the vertical line chart above.



Histograms

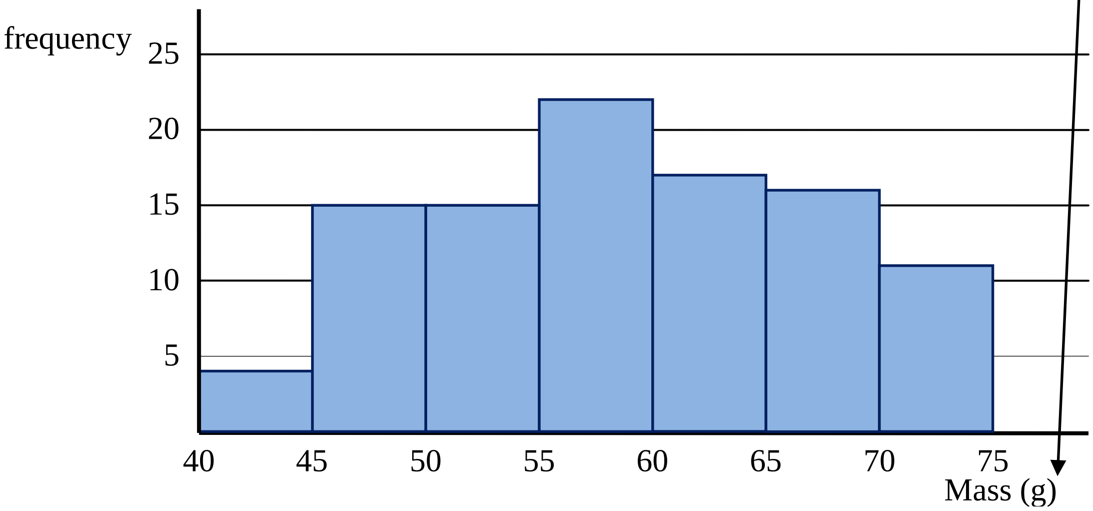
Bar charts, pie charts, dot plots and vertical line graphs are best used to display categorical and discrete data. When displaying sets of continuous data, a histogram is more suitable.

The table below shows data on the masses of 100 eggs, grouped into intervals of width 5 grams.

Class Int.(g)	Class Width (g)	Frequency
$40 \le m < 45$	5	4
$45 \le m < 50$	5	15
$50 \le m < 55$	5	15
$55 \le m < 60$	5	22
$60 \le m < 65$	5	17
$65 \le m < 70$	5	16
$70 \le m < 75$	5	11
$75 \le m < 80$	5	0

Remember to give units in tables and on axes.

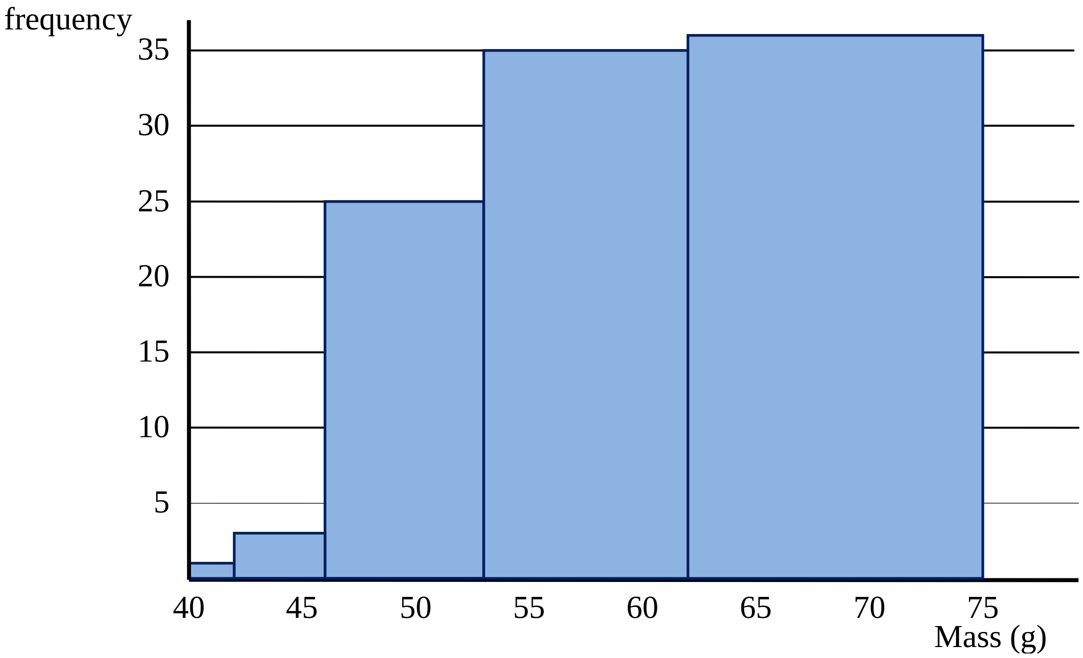
Plotting the frequency against the classes gives the following diagram:



Another way of classifying the eggs is given in the following table:

Size	Class Interval (g)	Class Width (g)	Frequency
Extra small	$40 \le m < 42$	2	1
Small	$42 \le m < 46$	4	3
Medium	$46 \le m < 53$	7	25
Standard	$53 \le m < 62$	9	35
Large	$62 \le m < 75$	13	36

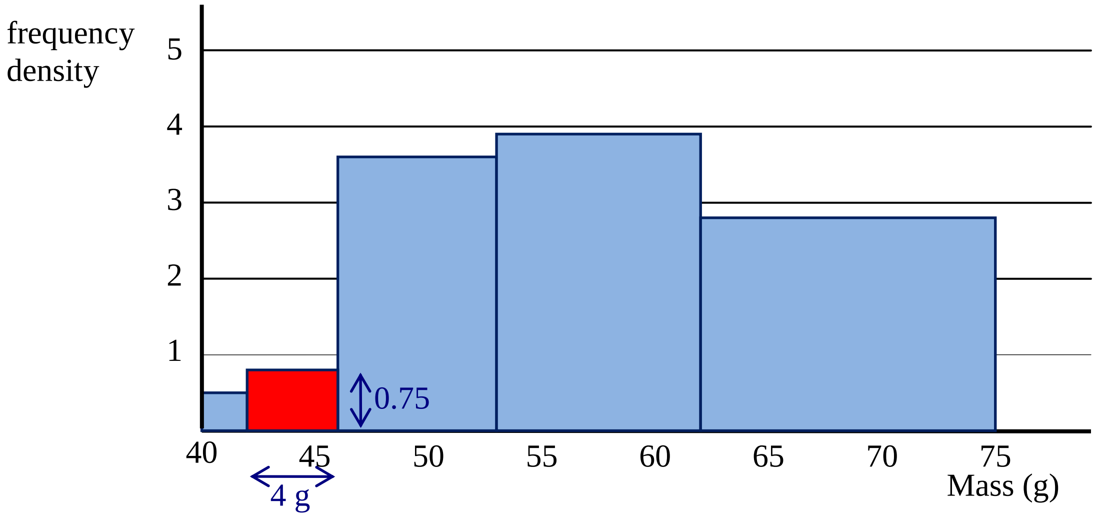
Plotting the frequency against the mass gives the following diagram:



The impression given by this diagram is quite different to the first one, because the data have been placed into unequal classes. It makes the distribution look negatively skewed, even though the first diagram shows it is distributed either side of a central modal group. **This diagram is not useful because it gives a distorted picture of the data**. In order to overcome this problem, we need to be able to take account of the unequal class widths, so that the diagram still gives an accurate impression of the overall distribution of the data.

When the class intervals are different, you compensate for the different widths of the intervals by making the ***area*** of the bar represent the frequency. To do this, you calculate the ***frequency density*** by dividing the frequency by the width of the interval to give the ***frequency density***. In this case this is the frequency per gram:

Size	Class Interval (g)	Class Width (g)	Frequency	Frequency density (frequency/g)
Extra small	$40 \le m < 42$	2	1	0.5
Small	$42 \le m < 46$	4	3	0.75
Medium	$46 \le m < 53$	7	25	3.57
Standard	$53 \le m < 62$	9	35	3.89
Large	$62 \le m < 75$	13	36	2.77



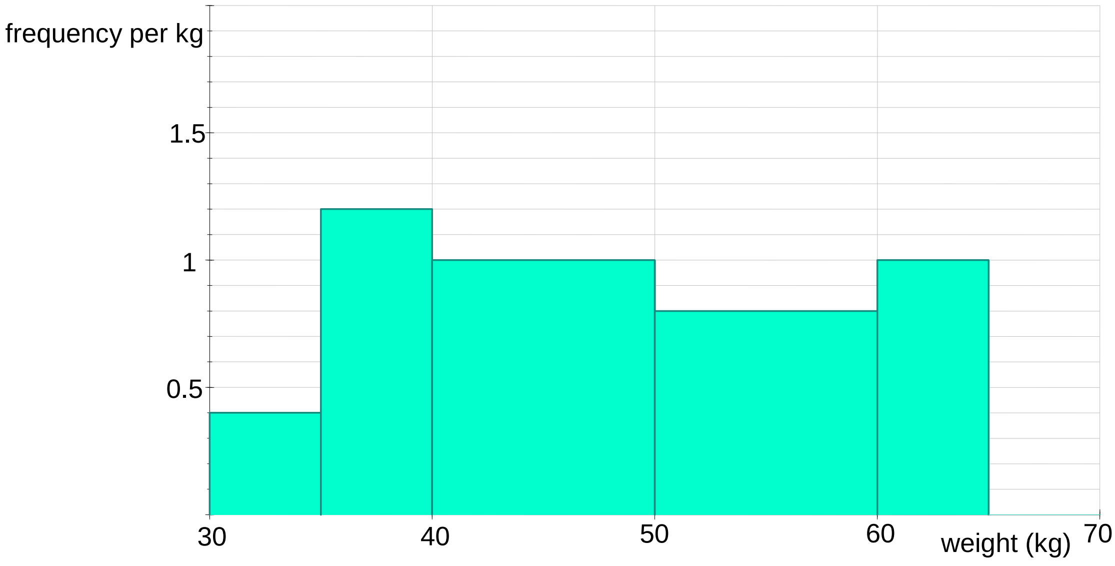
This diagram gives an impression of the overall distribution of the data which tallies with that given by the first diagram. The data is now fairly represented, even though it is grouped into intervals with different widths.

Look at the bar shown in red. The width is 4 grams and the frequency density is 0.75 per gram. So the area of the bar is 4 × 0.75 = 3, which equals the frequency. This diagram is now a ***histogram***.

If you divide the frequency by the class width, you get a frequency per unit interval and the area of the bar ***equals*** the frequency. However, the first diagram shown previously can also be regarded as a histogram, with the frequency density as frequency per 5 gram interval. In this case, the area of the bars is ***proportional*** to the frequency. The vertical axis should *not* be labelled frequency, however. On all histograms the vertical axis should either be labelled as frequency density, or with the units of the frequency density.

in this case units are frequency per gram

Example 3
The histogram below shows the weights of a group of 11 year old boys.



(i) How many boys weighed between 40 and 50 kg?
(ii) How many boys weighed between 30 and 35 kg?
(ii) How many boys were in the group altogether?

Solution
(i) The bar for 40-50 kg has a height of 1, which is the frequency per kg. Since the bar is 10 kg wide, the frequency is 10.

(ii) The bar for 30-35 has a height of 0.4. Since the bar is 5 kg wide, the frequency is 0.4 × 5 = 2.

(iii) The frequencies are as follows

Weight (kg)	Frequency
$30 \le w < 35$	2
$35 \le w < 40$	6
$40 \le w < 50$	10
$50 \le w < 60$	8
$60 \le w < 65$	5

So the total of the frequencies is 31. There were 31 boys in the group.