

HANDLING DATA COURSEWORK

Continuous data

Preamble.

All statistical investigations will involve collecting, representing and analysing data and most will involve a mixture of what are termed **continuous** data and **discrete** data.

In this short document I discuss what is meant by continuous data and give some examples of representing and analysing such data. Most of the work will be familiar to the majority of students.¹

| | |
|--------------------------------|--|
| <u>Continuous data:</u> | Numerical data which can take ANY value within a given range. Such data is usually obtained by measurement and usually has to be rounded up in some way. E.g. a persons height can, theoretically, take any value within a certain range and would be rounded up to, say, the nearest cm etc. |
|--------------------------------|--|

An example of continuous data.

Since continuous data has necessarily been rounded in some way, it is usual to present such data in the form of a **grouped frequency table**.

The following table shows the heights of a sample of 30 shrubs.

| Height (cm) | Frequency |
|-------------|-----------|
| 0–5 | 12 |
| 5–10 | 4 |
| 10–15 | 5 |
| 15–20 | 6 |
| 20–30 | 3 |

For example, 4 shrubs had heights between 5 cm and 10 cm etc.

Obviously such a table does not give us access to the original data and so any work we undertake will consequently not be as accurate as it would be were we to use the original data.

¹ This document does not give any advice regarding coursework tasks. It does not, for example, discuss which *average* might be the most appropriate for a particular set of circumstances etc. See the coursework section at <http://www.mathsguru.co.uk/> for such guidance.

Averages.

There are three basic ways of averaging a set of data.

i) Mode.

The 'most common' data item.

In the case of the data above, the best we can do is to say that the **modal** (most common) group is the 0 to 5 cm group.

ii) Mean.

The '*average*' obtained by totalling up all the data and dividing by the number of observations. The only average which uses every data item.

To estimate the mean of grouped data, we approximate by saying that each interval is represented by its mid-point. That is we assume that, for example, the 4 shrubs with heights between 5 and 10 cm were actually $\frac{5 + 10}{2} = 7.5$ cm each.

We place our working in a table.

| Height (cm) | Frequency | Mid-point | Total |
|-------------|-----------|-----------|------------------------|
| 0-5 | 12 | 2.5 | $12 \times 2.5 = 30$ |
| 5-10 | 4 | 7.5 | $4 \times 7.5 = 30$ |
| 10-15 | 5 | 12.5 | $5 \times 12.5 = 62.5$ |
| 15-20 | 6 | 17.5 | $6 \times 17.5 = 105$ |
| 20-30 | 3 | 25 | $3 \times 25 = 75$ |

E.g. 4 shrubs each with a height of 7.5 cm.

Therefore the mean height is given by

$$\begin{aligned} & \frac{\text{total of the 30 heights}}{30} \\ &= \frac{30 + 30 + 62.5 + 105 + 75}{30} \\ &= 10.08333... \text{ cm.} \end{aligned}$$

iii) Median.

The '*average*' obtained by selecting the 'middle' observation.

In the case of the above data, as with all grouped data, we can **estimate** the median by using a **cumulative frequency graph**.

| Height (cm) | Frequency |
|-------------|-----------|
| 0-5 | 12 |
| 5-10 | 4 |
| 10-15 | 5 |
| 15-20 | 6 |
| 20-30 | 3 |

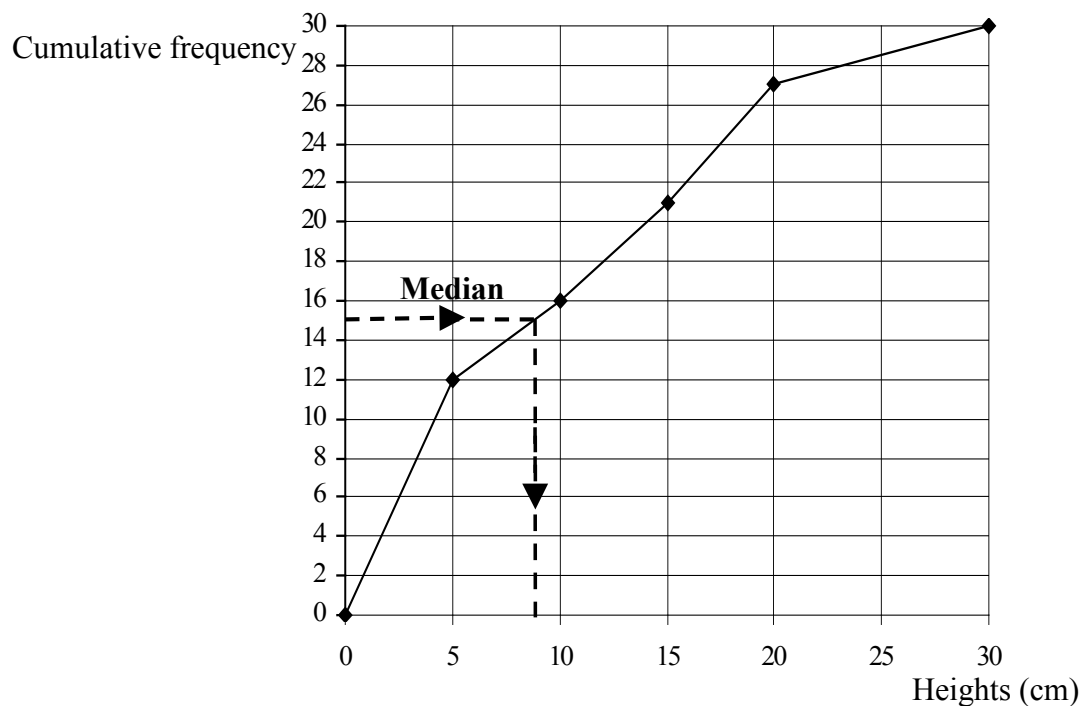
Since there are 30 shrubs, the median height is given (approximately) by the height of the 15th shrub which we quickly see is between 5 and 10 cm.

To get a better picture we add a column of *Cumulative Frequencies* to the table.

| Height (cm) | Frequency | Cumulative frequency |
|-------------|-----------|----------------------|
| 0–5 | 12 | 12 |
| 5–10 | 4 | 16 |
| 10–15 | 5 | 21 |
| 15–20 | 6 | 27 |
| 20–30 | 3 | 30 |

E.g. 16 shrubs each had heights **less than** 10 cm.

The following shows the **cumulative frequency graph** of the heights of the 30 shrubs.



From the graph we see that the median height is approximately 8 cm.

Hand in hand with calculating an *average*, is the need to determine how '*spread out*' the data is.

Measures of dispersion (measure of 'spread').

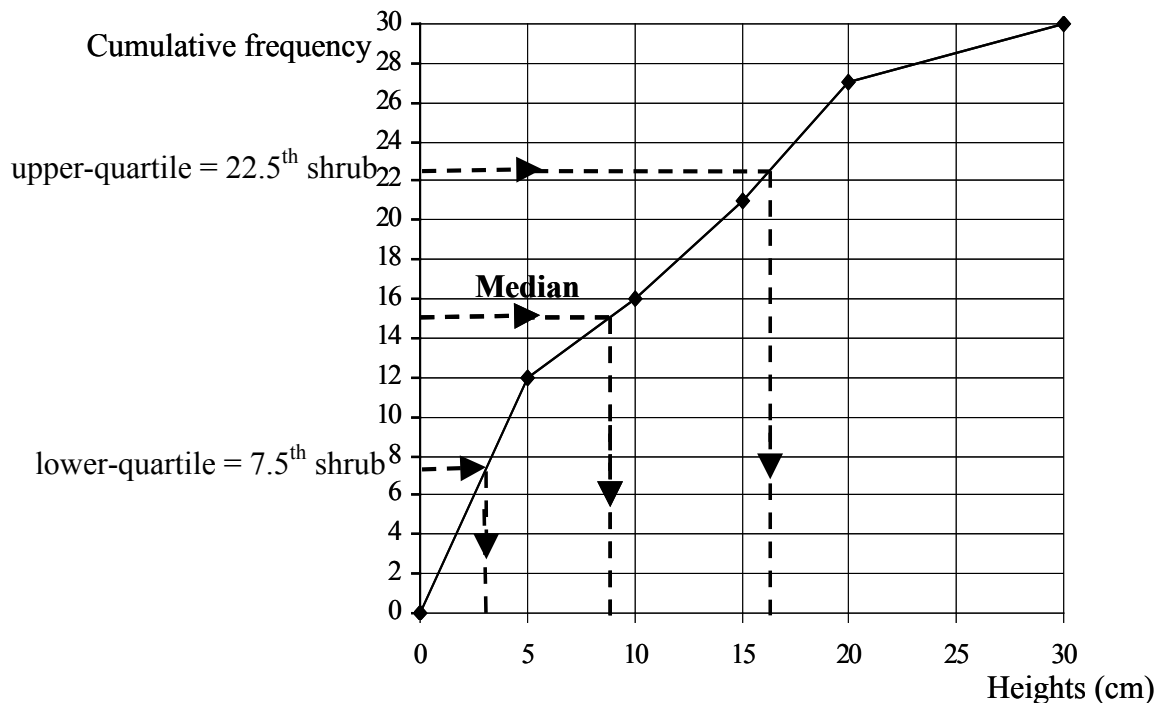
With continuous data we usually utilise either the **inter-quartile range** or the **standard deviation** (higher level GCSE only).

| | |
|-----------------------------|---|
| Inter-quartile range | The difference between the upper-quartile and the lower-quartile . The width of the interval between which the central 50% of observations lie. |
| Lower-quartile | Similar to the median . The observation lying one-quarter of the way into the sample. E.g. in a class of 20 students, the lower-quartile test score would be the score of the 5 th person from the bottom of the class. |
| Upper-quartile | Similar to the median . The observation lying three-quarters of the way into the sample. E.g. in a class of 20 students, the upper-quartile test score would be the score of the 15 th person from the bottom of the class. |
| Standard deviation | A numerical quantity obtained from a set of data, related to the mean, which measures how 'spread out' the data items are. |

Inter-quartile range.

With grouped data, estimating the quartiles and thus the interquartile range is best achieved with the aid of a cumulative frequency graph.

The following cumulative frequency graph (which we detailed above) shows the heights of the 30 shrubs and also details how we determine the quartiles.



From the graph we arrive at the following:

lower-quartile = height of the 7.5th shrub \approx 3 cm

upper-quartile = height of the 22.5th shrub \approx 16.5 cm

Inter-quartile range \approx 16.5 – 3 = 13.5 cm.

Standard deviation. (Higher level GCSE only)

This is a very sophisticated measure of ‘*spread*’ and is essentially the only measure which uses all of the data.

It is a measure of the ‘*average spread*’ of the data about the mean.

The formula for the standard deviation is given by $s = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$, where \bar{x} denotes the mean and \sum means “the sum of” etc.

Let us return (yet again!) to the shrub data:

We already know that the mean height is given by $\bar{x} = 10.08333\dots$ cm.

| Height (cm) | Frequency (<i>f</i>) | Mid-point (<i>x</i>) | x^2 | $f \times x^2$ |
|---------------|---------------------------|---------------------------|----------------|-----------------------|
| 0–5 | 12 | 2.5 | $2.5^2 = 6.25$ | $12 \times 6.25 = 75$ |
| 5–10 | 4 | 7.5 | 56.25 | 225 |
| 10–15 | 5 | 12.5 | 156.25 | 781.25 |
| 15–20 | 6 | 17.5 | 306.25 | 1837.5 |
| 20–30 | 3 | 25 | 625 | 1875 |
| Totals | 30 | | | 4793.75 |

Therefore,
$$s = \sqrt{\frac{4793.75}{30} - 10.08333\dots^2}$$
$$= 7.6235\dots \text{ cm.}$$

Charts and graphs.

All of the graphs discussed within the discrete data document² (except line-charts) apply to continuous data and so we do not detail them again here.

i) Cumulative frequency graphs.

We have already seen (above) how these can be used to estimate the median and the inter-quartile range of grouped data.

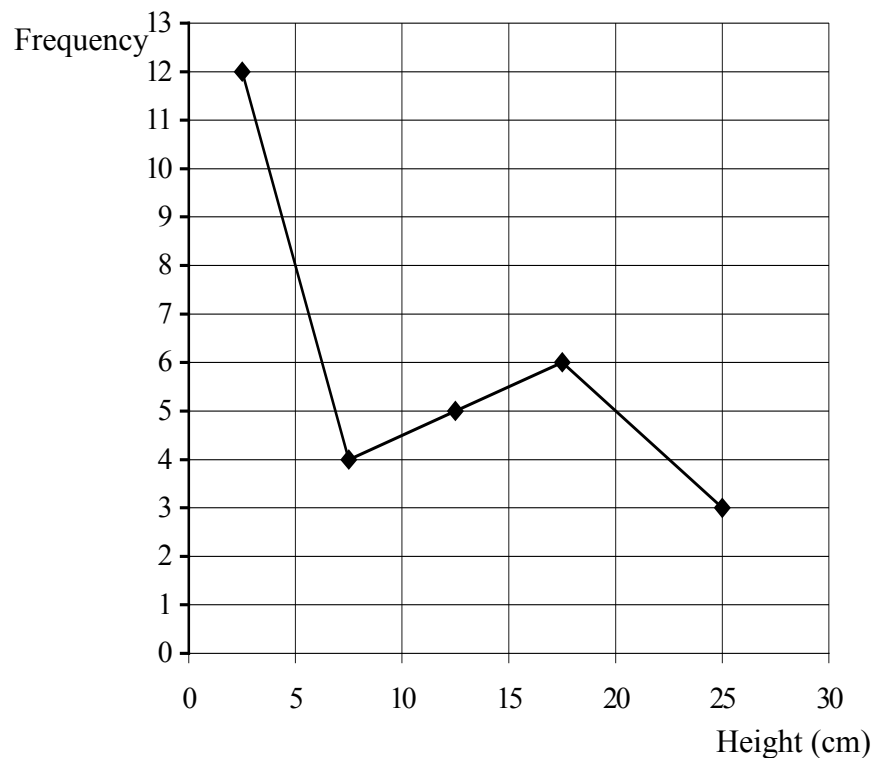
ii) Frequency polygon.

A form of grouped frequency diagram which is useful for comparing sets of data which have been **grouped** appropriately.

Let us illustrate with the shrub data again.

| Height (cm) | Frequency |
|-------------|-----------|
| 0–5 | 12 |
| 5–10 | 4 |
| 10–15 | 5 |
| 15–20 | 6 |
| 20–30 | 3 |

A frequency polygon is obtained by simply plotting the mid-points of each group together with the frequencies.



Frequency polygons are particularly useful when two or more sets of data are plotted on the same diagram.

² See the coursework section at <http://www.mathsguru.co.uk/> for the discrete data document.

iii) Histograms. (Higher level GCSE only)

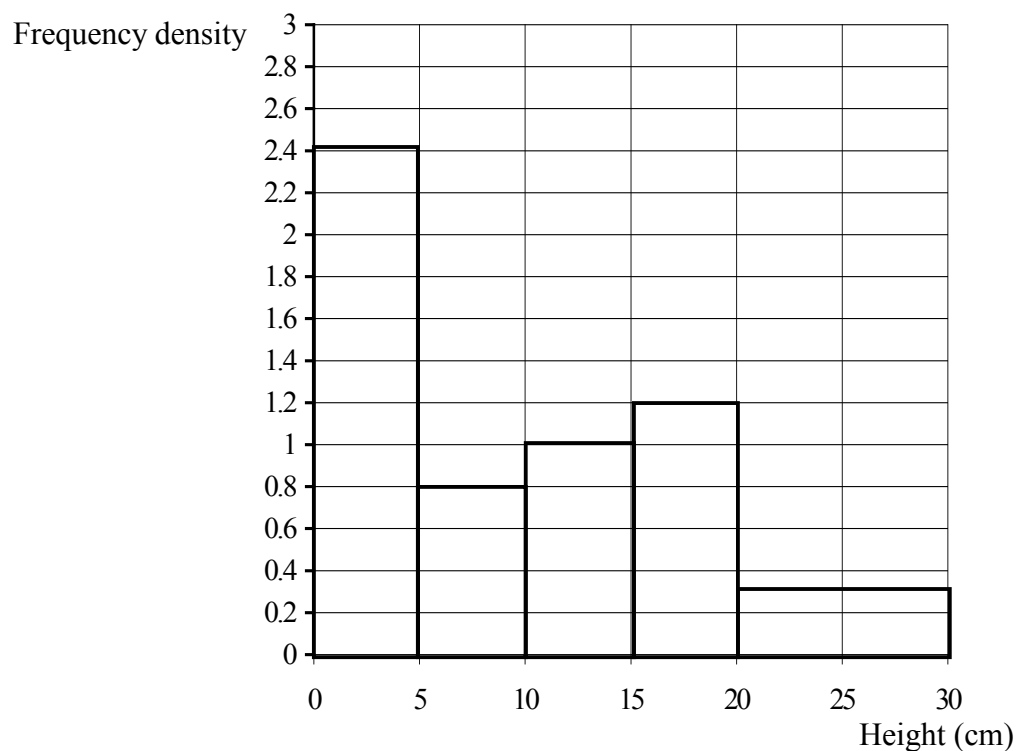
Used when data is placed into groups of unequal width. In such cases a simple bar-chart is misleading and a histogram is then seen to be a *corrected* bar-chart.

A histogram is constructed so that the area of each bar represents the frequency (as opposed to the height of the bar).

For the shrub data we need to first calculate what are termed **frequency densities** which are obtained by dividing the class frequencies by the relevant class width.

| Height (cm) | Frequency | Frequency density |
|-------------|-----------|-------------------|
| 0–5 | 12 | $12 \div 5 = 2.4$ |
| 5–10 | 4 | $4 \div 5 = 0.8$ |
| 10–15 | 5 | $5 \div 5 = 1$ |
| 15–20 | 6 | $6 \div 5 = 1.2$ |
| 20–30 | 3 | $3 \div 10 = 0.3$ |

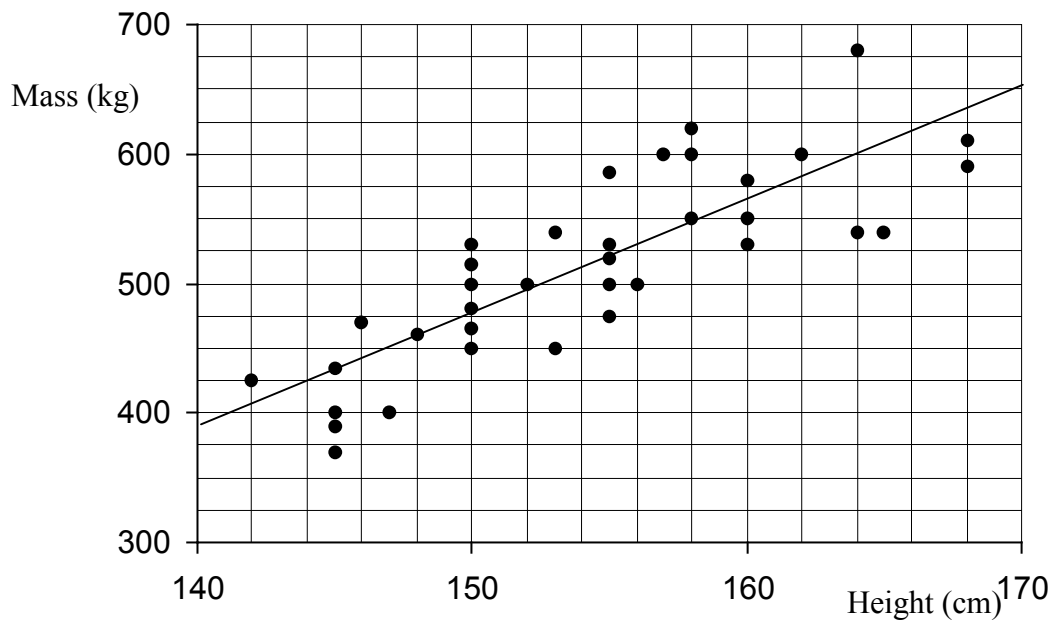
We can now draw the histogram.



iv) Scatter graphs.

Used when attempting to determine if there is any **correlation** between two sets of measurements obtained from the same sample.

Consider the following scatter diagram which shows the heights and masses of some horses.



Each point represents a single horse and gives its' height in cm and it's mass in kg.

This graph shows that there is strong **positive correlation** between the horses heights and masses and because of this we are able to draw a **line of best fit** through the plotted points.

Such lines allow us to estimate one of the quantities given the other.

NOTE that the line of best fit should pass through the '*average point*'.