

# HANDLING DATA COURSEWORK

## Discrete 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 discrete data and give some examples of representing and analysing such data. Most of the work will be familiar to the majority of students.<sup>1</sup>

**Discrete data:** Numerical data which can only take certain pre-determined values.  
E.g. the number of brothers each person has can only take values 0, 1, 2, ... etc.

Heights of adults could not be classed as discrete data since, in theory at least, someone's height would not be restricted to a predetermined set of values.

### An example of discrete data.

The following **frequency tables** show the marks obtained by two classes in the same test.

<i>Class 1</i>		<i>Class 2</i>	
Mark	Frequency	Mark	Frequency
0	3	0	1
1	8	1	2
2	7	2	1
3	7	3	5
4	5	4	3
5	7	5	2
6	3	6	7
		7	6
		8	3
		9	2

For example, in class 1, eight people obtained only 1 mark.

This data is discrete since each student could only obtain marks 0, 1, 2, ... etc.

Now in terms of analysing this data, I guess we would be concerned with which class performed the best **on average**.

---

<sup>1</sup> 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.

<b>Mean</b>	The ' <i>average</i> ' obtained by adding up all the data and dividing by the number of observations. The only average which uses every data item.
<b>Median</b>	The ' <i>average</i> ' obtained by selecting the 'middle' observation. <b>N.B.</b> the data must first be placed in order.
<b>Mode</b>	The 'most common' data item.

### Class 1.

Mode = 1 mark.

Median = 3 marks. (There are 40 students, so the median is obtained by *averaging* the marks obtained by the 20<sup>th</sup> and 21<sup>st</sup> students.)

$$\begin{aligned}\text{Mean} &= \frac{\text{total of the 40 marks}}{40} \\ &= \frac{3 \times 0 + 8 \times 1 + 7 \times 2 + 7 \times 3 + 5 \times 4 + 7 \times 5 + 3 \times 6}{40} \\ &= 2.9 \text{ marks.}\end{aligned}$$

### Class 2.

Mode = 6 marks.

Median = 6 marks. (There are 32 students, so the median is obtained by *averaging* the marks obtained by the 16<sup>th</sup> and 17<sup>th</sup> students.)

$$\begin{aligned}\text{Mean} &= \frac{\text{total of the 32 marks}}{32} \\ &= \frac{1 \times 0 + 2 \times 1 + 1 \times 2 + 5 \times 3 + 3 \times 4 + 2 \times 5 + 7 \times 6 + 6 \times 7 + 3 \times 8 + 2 \times 9}{32} \\ &= 5.21875 \text{ marks.}\end{aligned}$$

Hardly surprising in this case that every *average* suggests that class 2 have performed the better!

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

### Measures of dispersion (measure of 'spread').

With discrete data we can usually make do with either the **range** or the **inter-quartile range**. There is also the option of calculating the **standard deviation** of course (higher level GCSE only) but I prefer to leave this for continuous data.<sup>2</sup>

<b>Inter-quartile range</b>	The difference between the <i>upper-quartile</i> and the <i>lower-quartile</i> . The width of the interval between which the central 50% of observations lie.
<b>Lower-quartile</b>	Similar to the <i>median</i> . 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 <sup>th</sup> person from the bottom of the class.
<b>Upper-quartile</b>	Similar to the <i>median</i> . 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 <sup>th</sup> person from the bottom of the class.
<b>Range</b>	The difference between the largest data item and the smallest. Quite a crude measure of dispersion.

#### Class 1.

Range =  $6 - 0 = 6$  marks.

Lower-quartile = 1 mark. (There are 40 students, so the lower-quartile is the mark obtained by the 10<sup>th</sup> student.)

Upper-quartile = 4 marks. (There are 40 students, so the lower-quartile is the mark obtained by the 30<sup>th</sup> student.)

Inter-quartile range =  $4 - 1 = 3$  marks.

#### Class 2.

Range =  $9 - 0 = 9$  marks.

Lower-quartile = 3 marks. (There are 32 students, so the lower-quartile is the mark obtained by the 8<sup>th</sup> student.)

Upper-quartile = 7 marks. (There are 32 students, so the lower-quartile is the mark obtained by the 24<sup>th</sup> student.)

Inter-quartile range =  $7 - 3 = 4$  marks.

Range aside, there appears little difference between the two classes in terms of how spread out the test marks are.

---

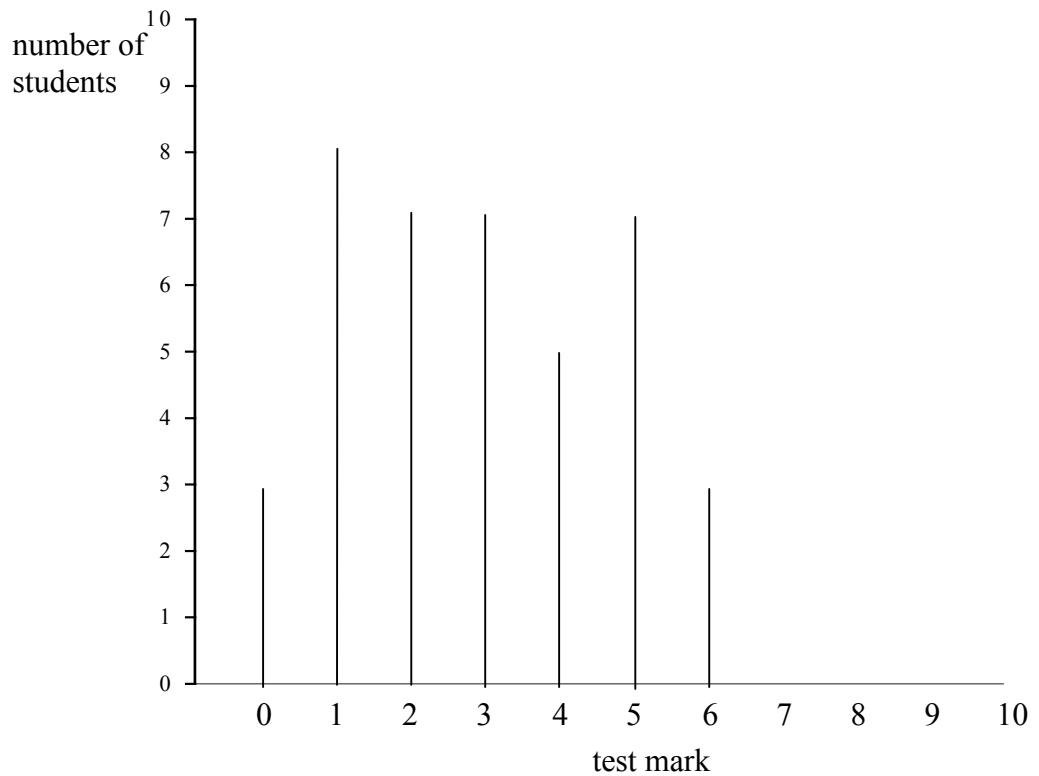
<sup>2</sup> See the work on **continuous data** for an example of calculating **standard deviations**.

## Charts and graphs.

Aside from basic pie-charts and bar-charts, both of which are really only useful for discrete data which is **qualitative** (such as favourite colour, favourite football team etc.) we are left with only a few viable charts to represent discrete data.

### i) Line chart.

The chart below shows the marks obtained by class 1.

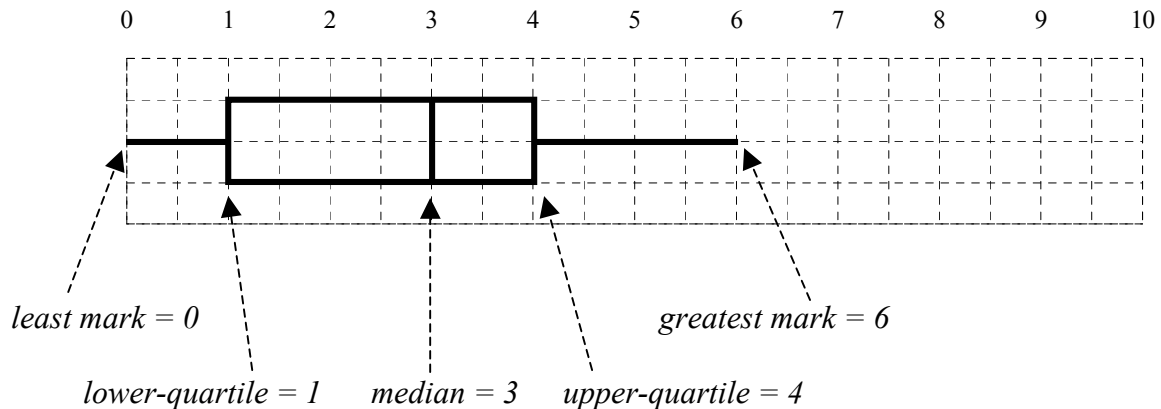


## ii) Box and Whisker plot.

Box and whisker plots allow us to readily compare two or more sets of data by examining the medians, the quartiles and the extreme values within our data sets.

The idea is simple and can be very effective.

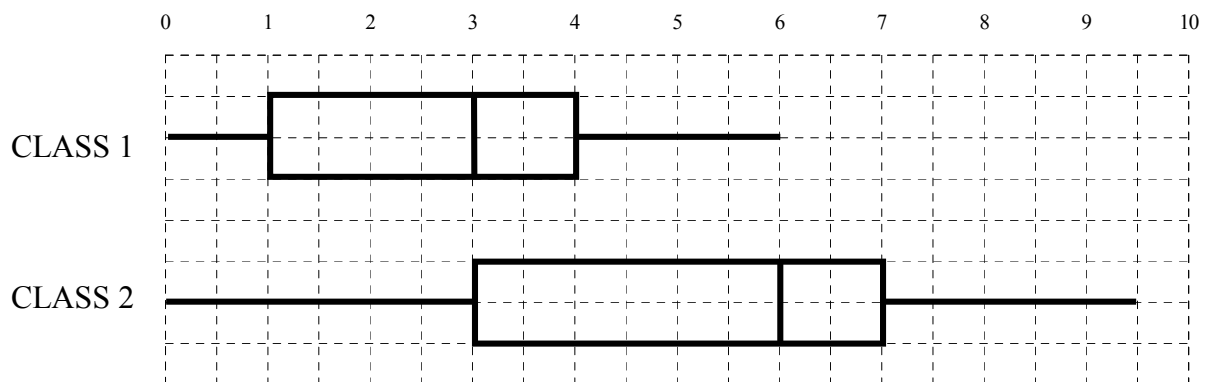
The diagram below shows a box and whisker plot for the marks obtained by class 1 together with a breakdown of the individual components.



The 'whiskers' give the range of the data; in this case from 0 marks to 6 marks.

The 'box' shows where the central 50% of test marks lie; in this case between the quartiles of 1 mark and 4 marks.

Of course, such a chart is only useful when used to compare different sets of data.



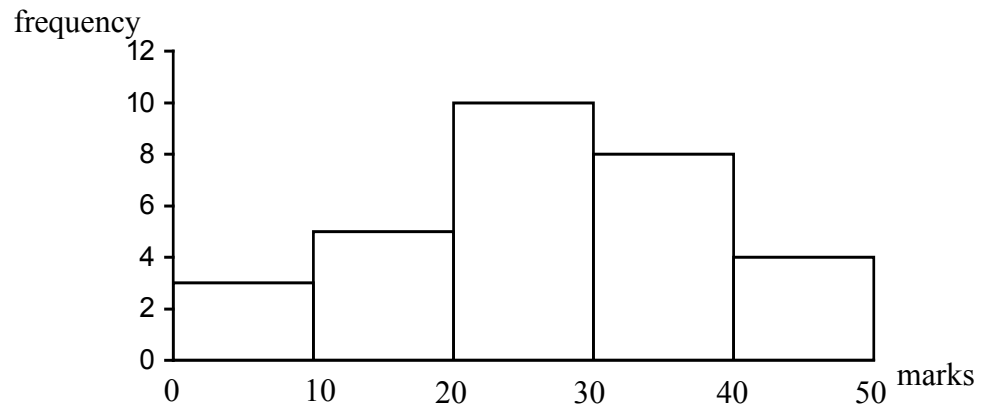
We see straight away from the plots that class 2 had a higher average (median) mark than class 1 and also had a greater range of marks.

It also confirms that the inter-quartile ranges are very similar.

### **iii) Grouped frequency diagram.**

A grouped frequency diagram (such as a bar-chart or a frequency polygon) is useful for comparing sets of data which have been **grouped** appropriately. Particularly useful for **continuous data**.

The grouped frequency diagram below shows the results of 30 students in a test.



For example, 5 students scored between 10 and 20 marks.

### **iv) Other charts.**

We leave a discussion of the remaining types of commonly used charts (such as **scatter graphs** and **cumulative frequency graphs**), until our discussion of **continuous data**.<sup>3</sup>

---

<sup>3</sup> See the coursework section at <http://www.mathsguru.co.uk/> for the document on continuous data.  
©www.mathsguru.co.uk