BioMedical Admissions Test (BMAT)

Section 2: Mathematics

Topic M6: Statistics

Topic M7: Statistics

Sampling and bias

A representative sample needs to be both random and large enough to represent the population.

Collecting data

Data can be divided into two main types.

- **Discrete data** is data which can only take certain fixed values. For example, shoe size can be whole or half values, $8\frac{1}{2}$ etc.
- **Continuous data** is data which can take on any values in a range and not just particular values. For example, this could be height.

When designing a **questionnaire**, it must be:

- Clear and easy to understand
- Easy to answer
- Fair (not leading or biased)
- Easy to analyse

Cumulative frequency

Cumulative frequency is the **running totals of the data**. They can be used for discrete or continuous data.

Drawing **cumulative frequency graphs**:

1. Copy the first frequency into the cumulative column
2. Then add the next frequency on to the number from step 1
3. Continue until you have completed the table
4. Plot the cumulative frequency on the vertical axis of the graph, against its upper class limit on the horizontal axis.
5. Join the points with a fitting curve

You can calculate the median by dividing the total by 2 and finding the corresponding number on the horizontal axis.

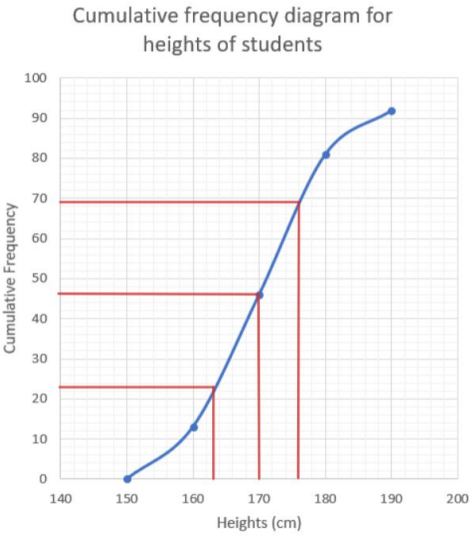
Example: Draw a cumulative frequency graph for the following data.

Heights, h (cm)	Frequency
$150 < h \leq 160$	13
$160 < h \leq 170$	33
$170 < h \leq 180$	35
$180 < h \leq 190$	11

Fill in the cumulative frequencies:

Heights, h (cm)	Frequency	Cumulative frequency
$150 < h \leq 160$	13	13
$160 < h \leq 170$	33	46
$170 < h \leq 180$	35	81
$180 < h \leq 190$	11	92

Plot the cumulative frequencies with the upper limit of each class interval, e.g. (160, 13), (170, 46) etc and join with a smooth curve.



Here we can see the red lines indicate the 25%, 50% and 75% lines. Therefore we can find the **median** and the **interquartile range** (75%-25%) from this graph.

Mean, median, mode and range

Mean, median and mode are all different forms of **averages**.

Ungrouped data

Mean $= \frac{sum\ of\ all\ values}{number\ of\ values}$

Median $=$ *the middle number*

Mode $=$ *most commonly occurring number*

Range $=$ *biggest number* $-$ *smallest number*

Interquartile range $= Q3 - Q1$

Example: find the mean, median, mode and range of the following numbers:
6, 2, 7, 8, 4, 7

Firstly, put these numbers in order: 2, 4, 6, 7, 7, 8
Mean = $\frac{2+4+6+7+7+8+}{6} = \frac{34}{6} = 5\frac{2}{3}$
Median = 6 and 7 are the middle numbers, so the median becomes 6.5
Mode = 7 is the most commonly repeated number
Range = 8 - 2 = 6

Grouped data

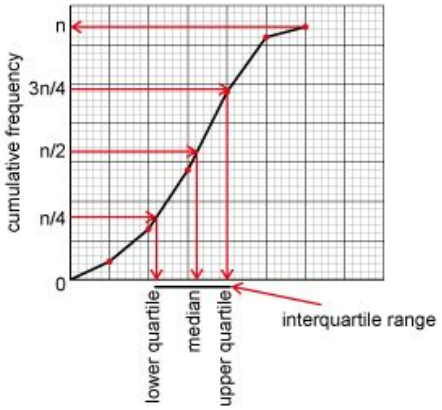
Calculating the estimated range = higher limit of the highest class interval - lower limit of the lowest class interval

Calculating the estimated median:

$$L + \frac{(\frac{n}{2} - C)}{f} \times w$$

Where L is the lower boundary of the class used, n is the total frequency, C is the cumulative frequency of the class before the median class, f is the frequency of the median class and w is the width of the class (so upper boundary - lower boundary)

It is also possible to find the median by using a graph.



Calculating the estimated mean:

1. Calculate the mid-interval value for each class interval
2. Multiply the mid-interval values by the frequency
3. Add the totals from step 2

This is only an estimate because we have to assume that the data within the class intervals are distributed evenly, which may not be the case at all.

Examples:

Class interval	Frequency (f)	Cumulative frequency	Mid-interval value (x)	$f \times x$
$1 \leq x \leq 5$	4	4	(1+5) ÷ 2 = 3	4 x 3 = 12
$6 \leq x \leq 10$	11	15	(6+10) ÷ 2 = 8	11 x 8 = 88
$11 \leq x \leq 15$	8	23	(11+15) ÷ 2 = 13	8 x 13 = 104
	Total = 23			Total = 204

Mode: the class with the highest frequency is $6 \leq x \leq 10$

Range: 15-1 = 14

Mean: $f \times x$ total is 204 and the sum of f is 23

$\frac{204}{23}$ = 8.87

Median: of the 23 figures, the median would be $\frac{23+1}{2}$ = 12th figure.

This falls within the class of $6 \leq x \leq 10$

Using $L + \frac{(\frac{n}{2} - C)}{f} \times w$

We turn this into $6 + \frac{12-4}{11} \times 4 = 8.9$

Averages and spread

Mean

- Can be used for any numerical data
- However, it is influenced by extreme values so may give false values

Mode

- Can be used for any data, even non-numerical data, such as favourite food
- The mode of a numerical data set may not be a central value

Median

- Not influenced by extremes - good indicator of central value
- Can only be used for data that can be ordered according to size

Range

- Gives a complete view of how spread out the data set is
- However, extreme values can skew the range greatly

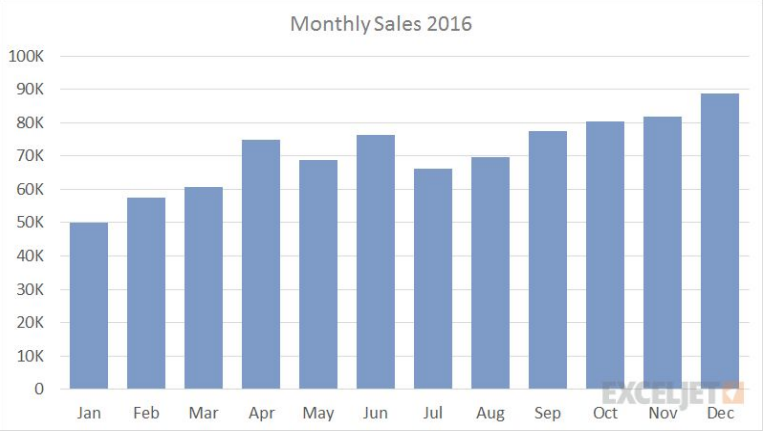
Interquartile range

- Shows spread of the middle 50% of the data
- Not affected by extremes
- However, does not give a complete picture of the range of data, as only looking at central 50%

Bar charts

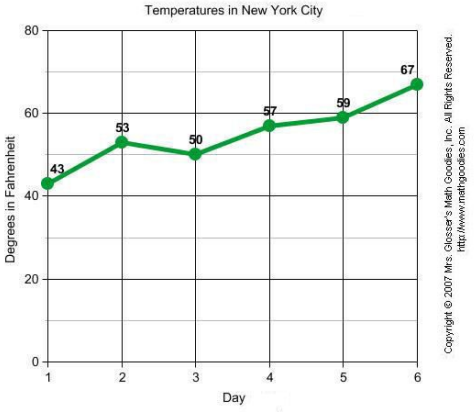
You should already be extremely familiar with bar charts. Each value is plotted as a bar.

Bar charts are useful when there are certain categories, such as here with months. Therefore, there should always be gaps between the bars.



Line graphs

These can be used to help visualise the change in one variable against time.

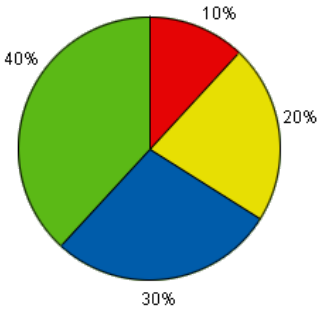


Pie charts

To draw a pie chart, it is important to first calculate the percentage that each section represents. We can then use this to calculate the proportion of the 360° that makes up a circle.

For example:

Color	Frequency	Percent	Angle
Red	5	10%	36 °
Yellow	10	20%	72 °
Blue	15	30%	108 °
Green	20	40%	144 °
Total	50	100%	360 °



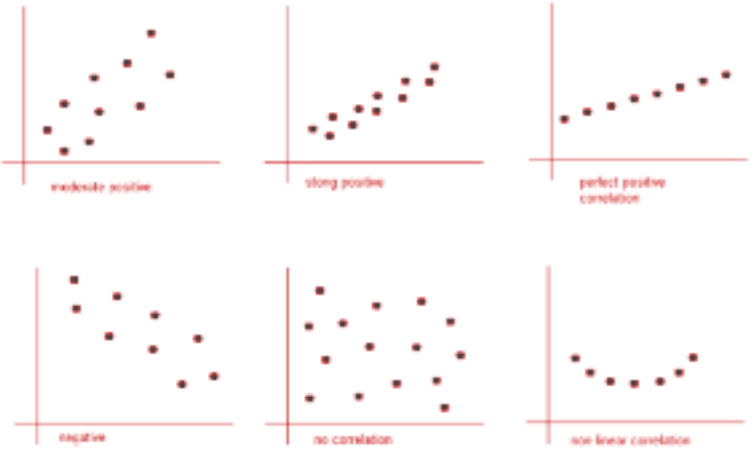
Scatter graphs

Bivariate data involves two variables and is used to see if there is any **correlation** (relationship) between the variables, for example the effect of height of a population on their weight.

The **explanatory variable** (i.e. the variable that is believed to be the cause of variation in the other factor) goes on the horizontal axis. The **response variable** (which changes in response to the explanatory variable) goes on the vertical axis. In our example, height would be on the horizontal axis.

When commenting on the correlation of variables, we can discuss whether it is a positive or negative correlation, and also the strength of this correlation.

You can see examples of these on the right.



Histograms

Histograms allow you to visualise patterns in data, e.g. is the data symmetrical? Is there an overall increase/decrease?

Rules for histograms:

- The vertical axis shows **frequency density**
- Bars are drawn on a **continuous**, horizontal scale
- There are **no spaces** between the bars unless a class interval contains no data
- The **area** of the bar is proportional to the frequency
- The bars are the same width as the **class intervals (class width)** and are bounded by the class intervals
- Class intervals do not have to be equal so the bars can be of **different widths**.
- **Frequency density** is defined as $\frac{frequency}{class\ width}$
- If the data is spread symmetrically then the **distribution is described as symmetrical**.