Statistics Year 10

Statistics is the process in which raw information – data – is processed to give us some extra meaning. The process is of no value if no extra information is found, regardless of how pretty the results look.

In Year 10 students should start moving from skills, such as finding means and drawing graphs, towards what those things can show us about the data. While being able to find the mean and median and draw graphs are important, students should not focus so hard on them that they forget to study the purpose of those skills.

Terminology

Students who have learnt the correct terms have a huge advantage over those that do not. The terms used are defined in the set of flash cards on the site.

PPDAC

The process of gathering and processing statistics in the NCEA system assumes the use of the PPDAC cycle.

Problem – in general terms, writing a question which can be answered.

Plan – how to go about answering the question, including recording results.

Data – collecting, cleaning and managing the information.

Analysis – sorting, finding means etc and then presenting that in tables and graphs.

Conclusion – answering the question posed at the beginning.

Typically students are happy with the "Plan", "Data" and "Analyse" (that is, the middle portion). However, they must not assume that ability in those areas will cover any flaws with the "Problem" and "Conclusion" stages: it will not.

While many students initially struggle with the starting "Problem" stage, a bit of practice in later years can usually sort that out. Generally the questions asked do not need to be particularly high level.

The "Conclusion" stage is the key for getting excellent results in later years, as it requires the highest level thinking. The best practice for developing conclusion writing skills is **comparison** questions (where one set of data is compared to another set). Students should work on these as much as possible.

Students do **not** need to memorise the PPDAC terminology, just understand the process.

Sampling

Sampling is the process whereby a small portion of the population (the whole set of data) is measured.

Sampling is done when:

- It is impossible to measure the whole population.
- It is too expensive or time consuming to test every item.
- Testing the product destroys it.

Sometimes it is technically possible to test the whole population, but by the time you have it has already changed, so the information is immediately out of date. The NZ census takes so long to gather that much of the information is already old by the time it is published. (A full population sample is called a census.)

Examples of testing destroying the product include testing fruit for residues, testing air bags to see if they inflate in time, and tasting wine to see how good it is.

Data sampling must be:

- Random – the selection process must give every member of the population equal chance.
- Large enough.
- Independent – the results of one sampling must not affect the results of any other sampling.

There is also the following concepts, which students are not yet expected to understand in any detail:

- Representative – divide the sample up in such a way as to mirror the population
- Free of Bias – must not introduce systematic errors as a result of the sampling process.

Year 10 students should be able to recognise when data is not being gathered randomly and lacks independence. They do not need to be able to design sophisticated systems to ensure randomness.

> e.g. a political poll samples people by randomly selecting families to question. That gives random selection, but not independence, as families often have similar political view.
>
> e.g. a political poll samples people by questioning everyone at a rock concert. That is not random, because the people selected will tend to be of a certain type, but is independent, because the views of one person going will not affect the next person.

"Large enough" is mostly about not being lazy (as some students choose ridiculously small samples to minimize the amount of work they have to do).

Frequency Tables

The frequency of something is how often that value (or range of values) occurs.

While a frequency table is easy to draw up, they can often lose information. It is recommended that they not be used with ranges to record data. A stem-and-leaf is better for that.

e.g. 25, 28, 30, 33, 33, 34, 34, 35, 36, 37, 37, 38, 40, 41, 41, 42, 42, 44, 46, 48, 50, 53

Value	Frequency	
20–29	2	There are 2 values somewhere between 20 and 29.
30–39	10	There are 9 values somewhere between 30 and 39.
40–49	8	There are 8 values somewhere between 40 and 49.
50+	2	There are 2 values of 50 or more.

Generally students need to be able to recognise what a frequency table shows, but will normally only need to use them in the form of a tally chart when recording data.

(A tally chart is when each data point is marked with a vertical line, with five being ~~𝍸~~, used to quickly count items. Otherwise a tally chart is just a form of a frequency table.)

Data Cleaning

Cleaning is the process of ensuring the data has all obvious errors corrected or removed.

Data may need to be corrected because
- the data entered uses the wrong units (e.g. height is in centimetres not metres),
- the data is entered in the wrong area,
- the value is typed in wrongly (e.g. 1779 is typed instead of 179),
- it is impossible to read (or the wrong format for the machine to read).

Data that cannot be corrected must be removed. You must say you are doing this, and explain why.

Only incorrect data can be removed. It is not acceptable to remove data because you do not like it, or because you can't be bothered working with so many points or any similar non-statistical reasons.

While it is good procedure to repair data, this can only be done if you are sure that the correction is certain. You may not guess what the correct answer was. It is a good idea to clearly indicate any corrections made to the data in your analysis.

Students are often confused what to do when comparing samples of different sizes and chop their data so that each set is the same size. You must not do this. All the higher level methods of analysing data are not affected by sample size, whether statistical (mean, median, range) or graphic (box-and-whisker, percentage bars).

Outliers

An outlier is a data point that is a long way from the others in the set.

An outlier is sometimes the result of an error, and it can be removed if this is the case.

If an outlier is correct then the options are:
- include it in all analysis, but state where it causes a value to be misleading,
- ignore it for some or all analysis, but state that you are doing so and why, or
- do the analysis twice: once with the outlier and once without it.

Not every value at the extremes is an outlier (there will always be some values at the top and the bottom end). An outlier has to be separated by some large distance from the other values.

By far the best option for dealing with outliers at Year 10 is to include the outlier in all calculations, but indicate what effect it has. This is because if the student has incorrectly called a value and outlier and ignored it, then any analysis without it is wrong.

In general students are too keen to call any value outside the normal range an outlier. In particular a group of values a long way from the normal range is most unlikely to be a group of outliers, and should probably be referred to as a cluster.

Clusters

A cluster is a number of data points that are close together, but separated from others.

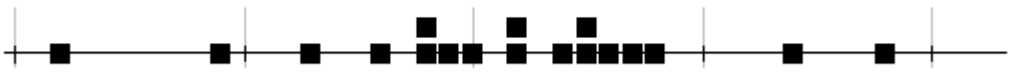
Many clusters occur naturally by random. Others can indicate something important about the data. Sometimes the data is evenly spread, but for one cluster, and sometimes the data is all clumpy.

e.g. the data below shows clustering

Students should note when clusters occur, and what their effect is. An educated guess as to whether it is a real effect or merely random can be made, but students should try not to read too much into their data.

It is usual for most data to be about a central value, with fewer and fewer values further away. This is not clustering.

e.g. this data is not clustered, the majority of values are usually in the middle.



Statistical Analysis

Year 10 students are expected to be able to find from a set of data the:

- Mean,
- Median,
- Mode,
- Quartiles,
- Range, and
- Inter-quartile range.

They should also know what these statistics show about the sample.

The mean (which should not be called the average) is a measure of the typical value. It is found by

$$\text{mean} = \frac{\text{all values added together}}{\text{number of values}}$$

The mean will almost always be a decimal, and students must resist the urge to round it to the nearest whole number at the calculation stage (but neither should they quote it to a ridiculous number of decimal places). However, when discussing its meaning it is acceptable to round it, provided this is indicated with an "about", "approximately" or similar.

The median is another measure of typical value. It is the middle value when the data is sorted into numerical order. If there are an even number of samples the median is halfway between the two middle values.

e.g. for: 3, 3, 4, 5, 6, 6 the median is 4.5, being halfway between the middle two values.

The lower quartile is the value 25% of the way along the ordered data. It is found by taking the middle value of all values below the median. The upper quartile is the 75% value, found by taking the middle value of all values above the median. As with the median calculations, if this falls between two values the middle of them is taken.

e.g. 30, 31, 31, 31. 35, 36, 39, 40, 40, 41, 43, 44, 49

The median is 39, being the 7th value out of 13. The lower quartile is 31, and the upper quartile is 42 (shown with arrows).

The process of finding quartiles should divide the data into four equal quarters.

The range is the difference between the top value and the bottom value. It measures the spread of data (but is very susceptible to outliers).

The inter-quartile range is between the upper quartile and lower quartile. It measures the spread of the middle 50% of data, so the range of typical values.

The mode is the most common value. It has little use in data analysis.

Graphical Analysis

Year 10 students are expected to be able to draw and analyse, from previous years:

- Column and bar graphs,
- Line Graphs,
- Histograms,
- Box-and-whisker charts,
- Stem-and-leaf plots,
- Percentage bars, and
- Pie charts

New to this year are

- Scatter plots.

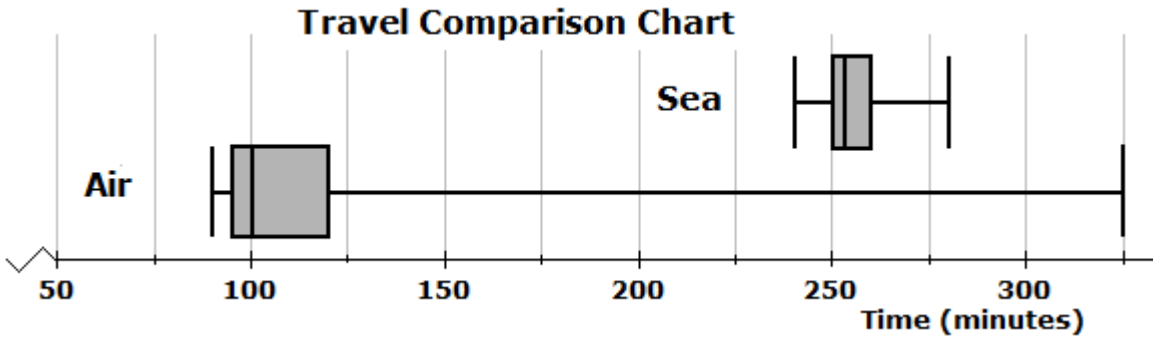
In NCEA assessments it is usual for students to be given a choice in which form of graph to use. By Year 10 students should be past drawing stem-and-leaf, bar graphs or dot plots as the primary graph. While easy to draw, they convey little information, no matter how pretty.

In general data showing a change over time will be plotted using a line graph. Bivariate data (where each data point has two measurements) will need a scatter graph. Other data will usually be shown by way of box-and-whisker – especially when comparing data sets.

The following general rules for graphing apply. All graphs need:

Title – choose something pretty simple (there are no extras for a "better" title).
Labelled axes – saying what both represent, and the units used to measure them, or
Key or clear indication of the meaning of colours.
An even scale all the way along.

Students usually know these rules, but forget to apply them to the new graphs, such as box-and-whisker graphs. They still need a title and labelled axis with units.

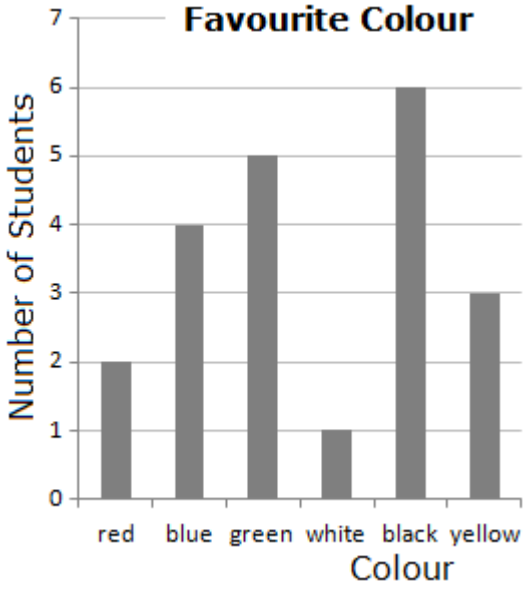


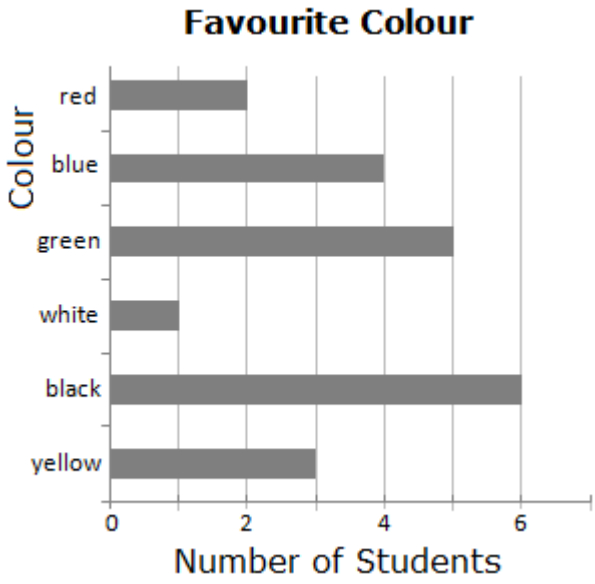
Statistics axes do not need to start at zero, and it is often easier to see the information if they do not. (Note this is different from drawing graphs in any other topic of Maths – Algebra, Graphing or Patterns – where the standard is to always start at zero.)

The axis should have a zig-zag or clearly marked gap if it does not start at zero.

Column and Bar Graphs

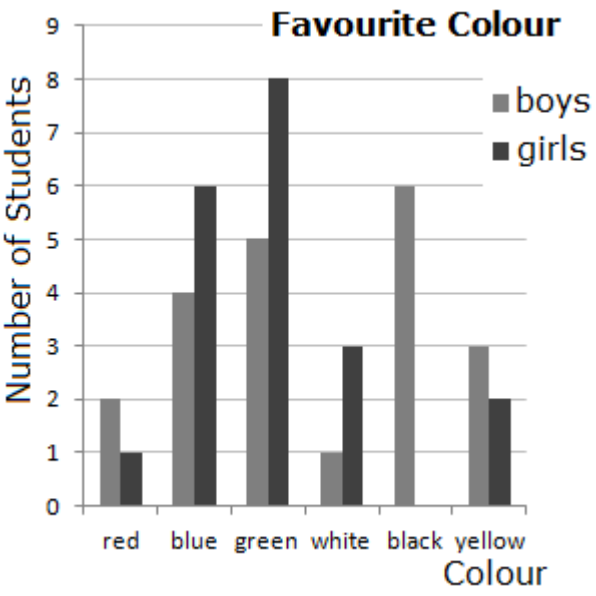
Column or bar graphs are mostly used for categorical data – data which has no order: eye colour, males vs females etc.

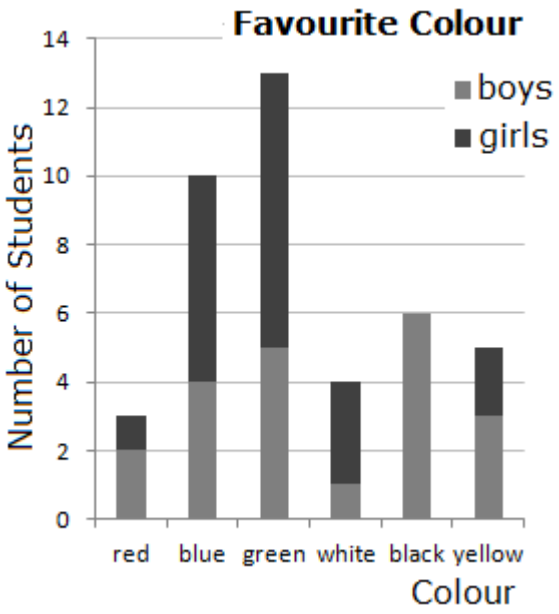




There is basically no reason to prefer column over bar graphs, although it is usually easier for students to draw with vertical bars.

While students need to be able to draw and interpret column and bar graphs, they are usually not at a sufficiently high level for NCEA purposes. They are most acceptable when they are used for comparison purposes (as left below) or to show breakdowns inside each sub-total (as right below).





Histograms

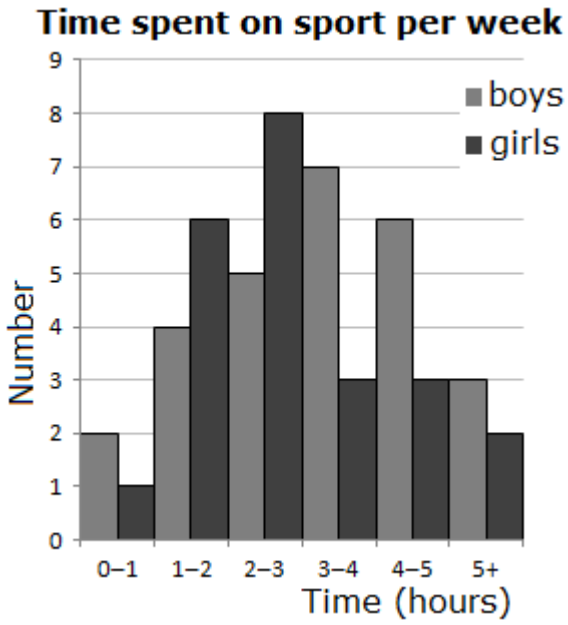
Histograms are like bar charts, except they are used for continuous data – data which can take any value not just exact numbers – or data which is being treated as if it was continuous.

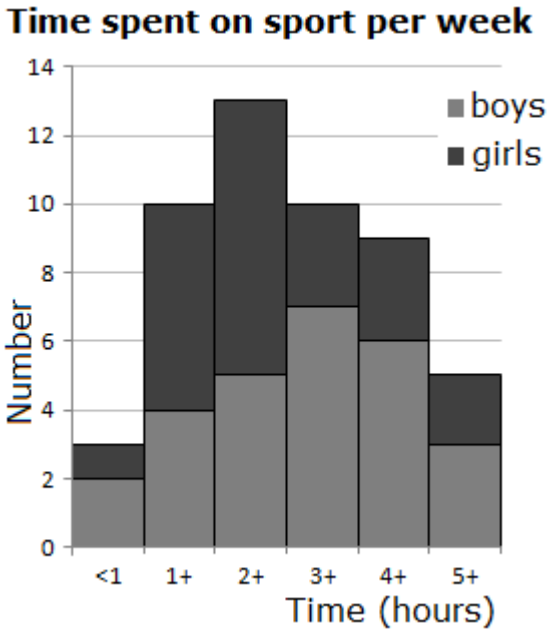
The key to histograms is that the bars represent a range, even if only because the values have been rounded. Because their are no gaps between the values represented by each column, they are shown with no gaps between them.

Histograms allow a look at the overall shape of a distribution, but are normally far from ideal. When they have comparison data they can be very hard to read.

The graph on the left below shows the peak for girls is at 2-3 hours, whereas for boys it is at 3–4 hours. The difficulty is in trying to see where the overall peak is.

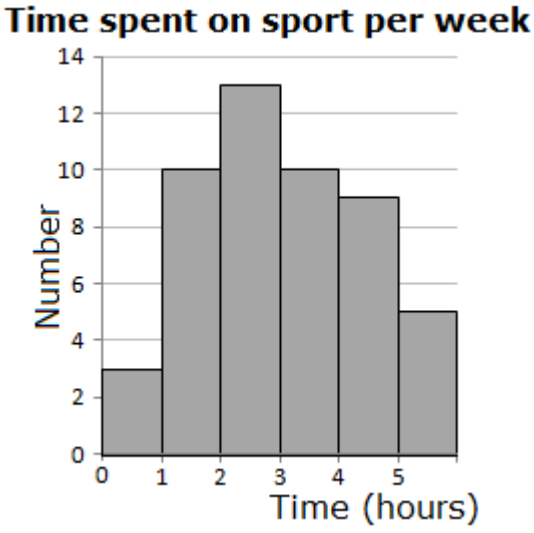
The graph on the right shows the overall peak clearly, but while the distribution for boys is obvious, that for girls is very difficult to make out without a base line.





In the examples above the individual bars are labelled fully, but the common alternative is to label the points **between** the bars.

We see to the right that system in operation. The second column is from 1 to 2 hours, marked on the joins.



Histograms are also very dependent on the size of each column. The same data can be made to look very different by choosing different ranges. In general they are not good for higher level analysis.

Stem-and-Leaf Plots

Stem-and-leaf plots are used to organise data into order, either to record it quickly or so that the quartiles can be found.

The stem, on the left, indicates the range and the leaf the actual values. Data is put into the proper range, preferably in such a way as to leave the data in numerical order (an "ordered" stem-and-leaf).

e.g. 22, 36, 45, 26, 44, 51, 33, 36, 28, 47, 33, 54, 23, 39, 42, 40, 29

2	2, 3, 6, 8, 9
3	3, 3, 6, 6, 9
4	0, 2, 4, 5, 7
5	1, 4

The numbers to the left side signify the range, with "2" standing for anything in the 20s, so that the numbers to the right don't need to include the first 2 when written down.

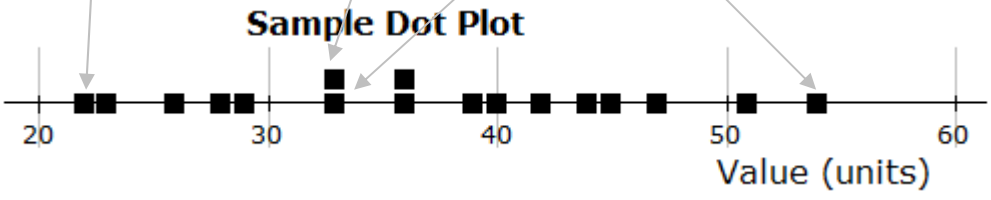
Stem-and-leaf plots merely repeat all the information already given. Because of this they cannot be considered a suitable graph for analysis. They are for recording or ordering only.

Dot Plots

Dot plots give a visualisation of the spread of data, with each data point being shown on a scale.

A scale is drawn in one direction – usually horizontally, but it can be vertically too. Each point is plotted where is appears on the line as a dot, cross, box etc. Values are placed above previous entries of the same value or when there is not enough room otherwise. It is important that each value be clearly separated from the others for the correct visual effect.

e.g. 22, 36, 45, 26, 44, 51, 33, 36, 28, 47, 33, 54, 23, 39, 42, 40, 29



Dot plots can show valuable information, such as clusters or peaks of data, but are rarely sufficient value on their own. If drawn they are generally hugely improved by having a box-and-whisker added (since the dot plot arranges the data in order and already has the right scale, this generally doesn't take long). They can take a long time to draw if there are many data points.

Scatter Plots

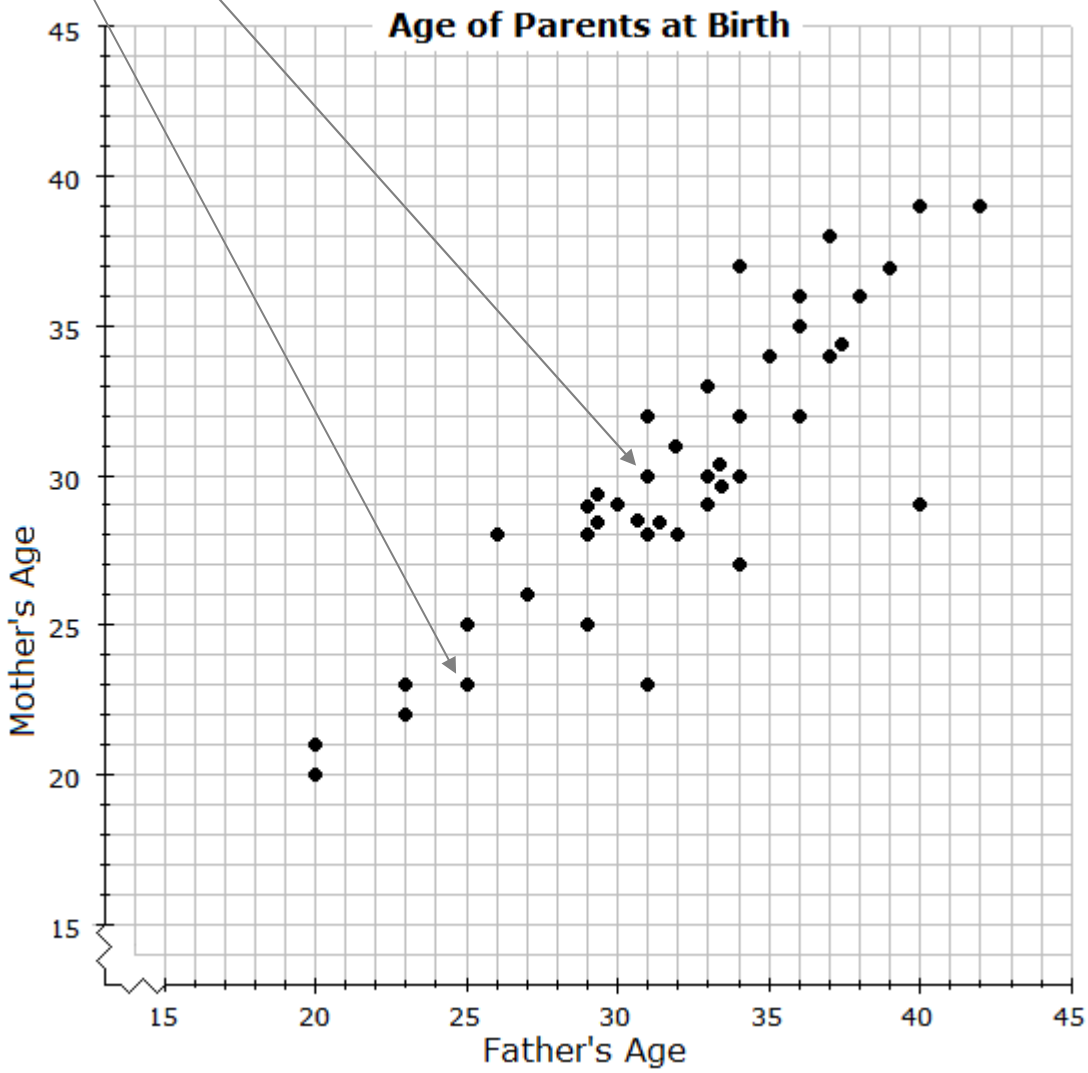
Scatter plots give a visualisation of how one measurement is related to another.

For each data point, there are two measurements. The distance along one scale is crossed with the distance along the other, and the intersection plotted with a dot, cross etc. It is usual to put the variable that is thought to be the cause of the other – if there is one – along the *x* axis.

Below is plotted a graph showing the age of the mother and the age of the father for some births. Each point represents one birth.

This birth is for a mother aged 23 and a father aged 25.

This one for a mother aged 30 and a father aged 31



If two points would be on the same spot, it is usual to mark the second as close as possible to the first, but so that it is still visible. (Some graphing programs make the dot bigger instead of marking it twice, to indicate the extra weight.)

The scales on the two axes do not need to be the same, but in order to make the data's pattern as obvious as possible it is common to start the scales already partway along, marked with a zigzag line or similar.

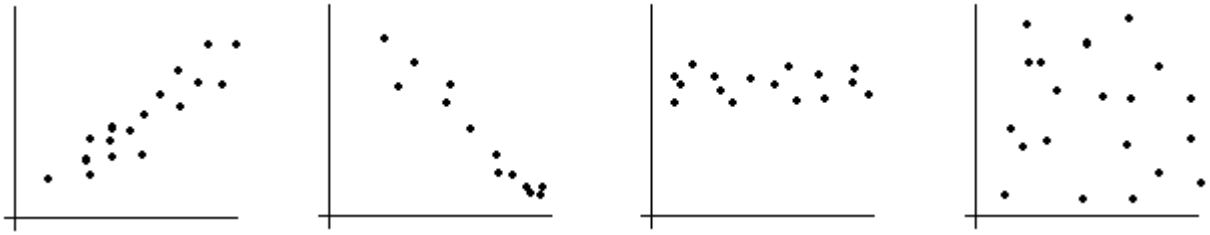
Correlation

If the dots form a more or less linear pattern, then there is said to be a correlation between the two measurements. A "line of best fit" is the line that best describes the pattern.

A correlation is positive if as the *x* variable increases the *y* variable increases. A correlation is negative if, as the *x* variable increases, the *y* variable decreases.

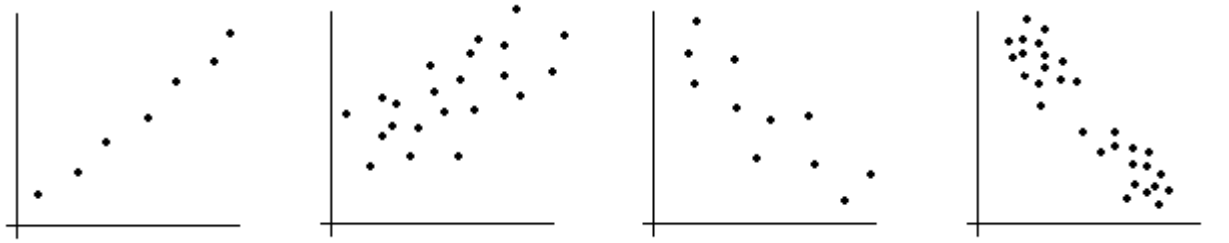
A correlation is said to be strong if the points are mostly close to the line of best fit. It is weak if the points are scattered further away.

There is no correlation if the *x* variable or *y* variable are more or less constant regardless of the other – that is, a basically horizontal or vertical line – even if the line is very obvious.



Positive Correlation | Negative Correlation | No Correlation | No Correlation

A strong correlation indicates that the relationship is reliable, and so one variable can be confidently used to predict the other. A weak correlation means – while there is some relationship between the measurements – that one cannot use one to predict the other with any accuracy. Note that "strong" and "positive" are entirely separate concepts.



Strong Positive | Weak Positive | Moderate Negative | Clustered Negative

To draw the line of best fit, place a ruler so that generally there are be as many points above the line as below, and also so that the scatter on either side will be basically even. Do not try to make the line go through (0, 0) unless there is a reason to believe it should. Students are not expected to get the line of best fit perfect. It may be necessary to ignore outliers.

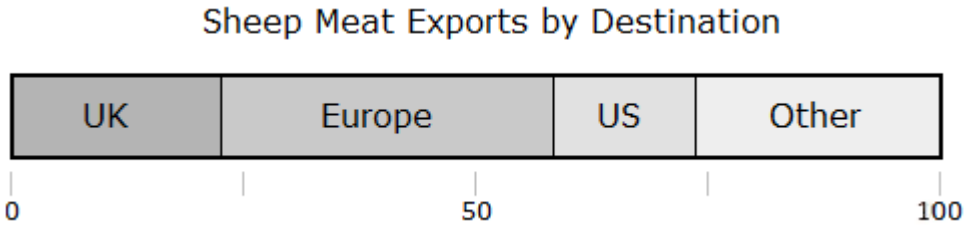
The line of best fit can be used to make predictions (assuming the correlation is reasonably strong) and its position may also give some information (via slope, intercept etc) about the relationship.

Correlations say **nothing** at all about causes. If measurement A and measurement B are correlated, then there are four main options: (1) A causes B, (2) B causes A, (3) something else causes both A and B, or (4) it is a fluke. Do not assume because there is a correlation that there is a direct cause.

Percentage Bars

Percentage bars are used to show the fraction of a total divided into different categories.

A long fat bar is drawn, then divided up into sections in the appropriate percentages.



These can be useful to compare different proportions (between years, countries, male/female etc).

A scale is often useful, either as a percentage (as shown above) or a measure of the actual value the bar represents.

The method for calculating the length of each segment is: fraction represented × whole length.

e.g. 30 out of 75 on a 12 cm long % bar, is represented by $\frac{30}{75} \times 12 = 4.8$ cm.

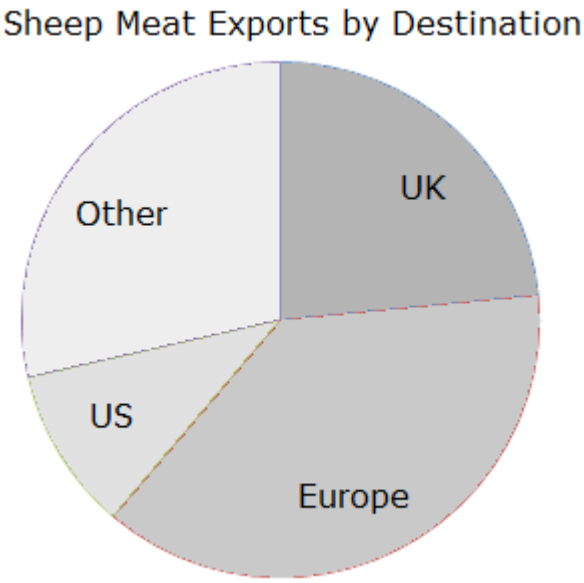
Pie Charts

Pie charts are also used to show the fraction of a total divided into different categories.

A circle is drawn up, then divided up into sections in the appropriate percentages.

There are few, if any, situations where a pie chart is preferable to a percentage bar.

Pie charts are
- hard to read properly,
- cannot have a scale added,
- are awkward to draw, and
- make it hard to do side by side comparisons.



The method for calculating the angle in a pie chart is to take the fraction of the amount represented and multiply by 360°.

e.g. the wedge for 22 out of a total of 50 is found by: $\frac{22}{50} \times 360° = 158.4°$

Box-and-whisker Graphs

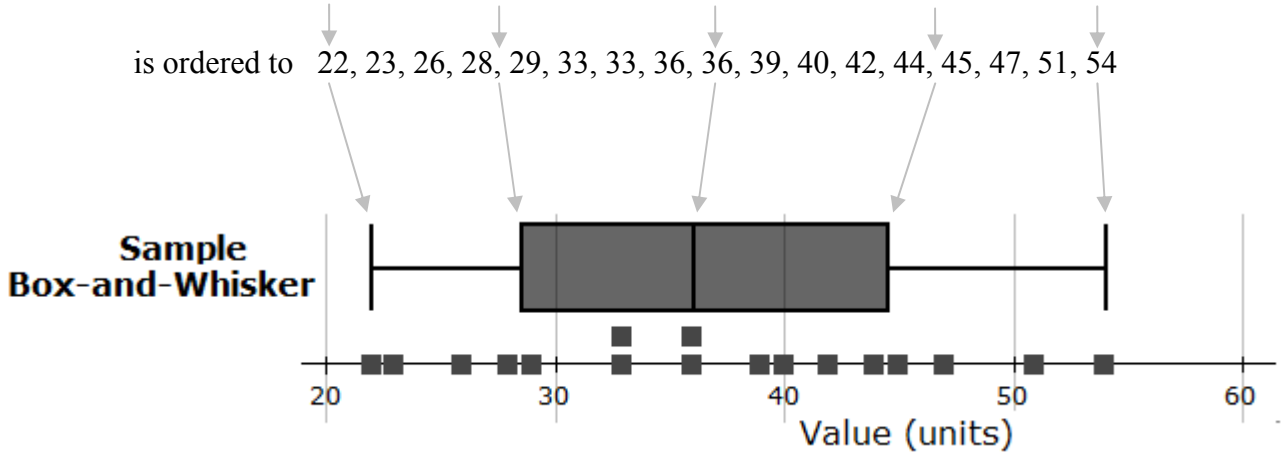
Box-and-whisker graphs show the spread of data, and especially the typical range of values.

They are particularly useful when comparing distributions of data.

The data is ordered, and the lowest value, lower quartile, median, upper quartile and highest value are calculated. These five values are then marked on a scale – whether horizontal or vertical – and then connected with a box in the middle and whiskers to the ends.

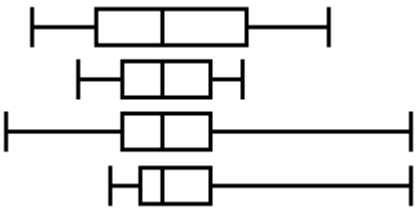
e.g. 22, 36, 45, 26, 44, 51, 33, 36, 28, 47, 33, 54, 23, 39, 42, 40, 29

is ordered to 22, 23, 26, 28, 29, 33, 33, 36, 36, 39, 40, 42, 44, 45, 47, 51, 54



It is vital to remember that each segment represents exactly 25% of the values. The span of the inner box (the inter-quartile range) then 50% of all the data and is a good measure of the spread of typical values.

Normal smoothly spread values will give box-and-whisker graphs where each quarter is more or less the same length. One of the strengths of box-and-whisker graphs is that deviations from that are clearly visible.



Symmetrical data, with a fairly even spread

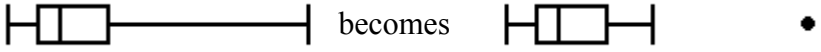
As above, but more consistent (because tighter)

Data with a large amount of extreme values

Asymmetrical data (skewed right)

The box-and-whisker explained above is the standard expected for Year 10 students. They can be seen in slightly different forms in the real world (especially medicine):

◦ Because outliers badly affect the length of the whiskers, in extreme cases those points are marked separately with a dot or cross and the whisker only shown to the next value.



◦ The inner box is sometimes different from the middle 50% (if so, it must be clearly noted).

◦ The mean is sometimes shown with a small cross. Very rarely it replaces the median.

Line Graphs

Line graphs are used to show patterns (trends) in the data, especially over time.

Time, or its equivalent, is always on the horizontal axis (*x*-axis. The scale should be evenly spaced, even if the information is uneven.

The vertical axis is then the one that varies. Often it will not start at zero, in order to show the changes better, but this must be clearly marked (normally with a zig-zag line).

The information starts from the first date and goes to the last date, even if this means it does not reach the side of the graph. Points are connected by straight lines. Use a ruler. Individual points are normally shown, but are often not shown if it makes the graph hard to read with them.

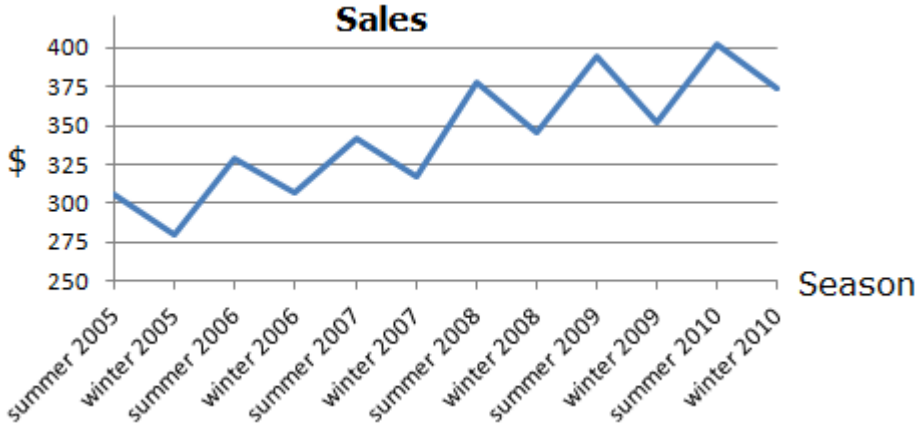
e.g.

Year	Price
1999	36
2001	37
2002	42
2004	41
2006	47
2007	51
2008	50

Missing information must be noted clearly. In the case above the dots indicate the actual data points, so that it is clear when years are missing. Other options include leaving gaps and using dotted lines to connect points with a gap in between. Never put the graph to zero for missing information, as that destroys any pattern.

Line graphs are most useful to see patterns in the data, which are called trends, especially when comparing two sets of data. There can be both long term and short term trends.

e.g. this graph shows a seasonal pattern, with a long term trend to higher sales.



Hints for Good Analysis and Conclusions

The conclusion is the most important part of the PPDAC cycle when attempting to get Merit or Excellence. It should follow logically from the steps in the analysis.

Mistakes can be overlooked in other sections if the conclusion is good.

First decide if your analysis is of data taken at one time (a "snapshot"), is bivariate (has two values for each data point), or is of values that change with time. They need to be treated very differently.

By Year 10 you should be moving away from just providing means, medians etc and graphs, and more towards what they show about data.

Snapshot Data

The starting point with static data is calculating the mean, median, quartiles etc and drawing a box-and-whisker graph. Other graphs can be useful, as further evidence.

Start your analysis and conclusion with a discussion about the "typical" values of your data. The mean is the most important statistic for that, unless the data is quite asymmetric in which case the median might be considered more "typical".

You can expand your discussion of the typical values by referring to the inter-quartile range. Be careful not to refer to it as containing "most" values, as the IQR covers exactly half of all data points.

From there you can discuss the overall spread of the data, discussing the range and any outliers, clusters or lack of overall symmetry.

Bivariate Data

Bivariate data is plotted with a scatter graph.

Start your analysis with any evidence of a trend line – a correlation. Say whether it is positive or negative, and whether it is weak or strong. Note any outliers or clusters.

Finish with any conclusions about the trend. Do not overdo what causes the correlation, since data analysis can only point towards that, not prove it.

Note: there is no point drawing a scatter graph and then discussing the two different measurements separately. The whole point of such a graph is to find the relationship between one measurement and another. If you want to discuss each measurement separately, draw another sort of graph.

Time Series Data

The starting point with time data is a line graph. While a mean, median etc can be calculated for time series, they often have no meaning in terms of the analysis.

Start your analysis and conclusion with the long term pattern(s). Ideally, put an approximate number to any slope of the trend. You can briefly suggest what that means for the future.

Then note any short term patterns that exist – seasonal or otherwise.

Finish with any notes about the amount of variation in the data (is the line fairly straight, or does it move around a lot) and how your analysis was affected by any unusual values or outliers.

In all Cases

Wherever possible use the numbers you have calculated and the graph(s) you have drawn to explain why you come to your conclusion(s).

Stick mostly to the data and statistics. While a quick indication as to what causes a pattern or trend can be useful, do not wander off into any discussion of more than a sentence. In particular do not make wild guesses about what the data means. It is not Social Studies: stick to what the data shows.

You should try as much as possible to compare two data sets or two measurements, rather than analyse them separately.

Excellence is often obtained by comments relating to any problems (at any point in the PPDAC cycle) and any improvements that might be made. Remember that you are being marked on your statistical knowledge, so improvements that are not Mathematical are of no interest.