# Bivariate Analysis

## Data Cleaning

Cleaning is the process of ensuring the data has all obvious errors corrected or removed. Data that cannot be corrected must be removed. You must say you are doing this, and explain why.

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).

**Only data known to be wrong 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.

## Outliers

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

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 a very long distance from the other values.

In general students are far 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.

Include the outlier in all graphs, but indicate what effect it has. You may draw the line of best fit to exclude it, if you are certain you want to call it an outlier.

## 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.

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.

## Scatter Plots

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

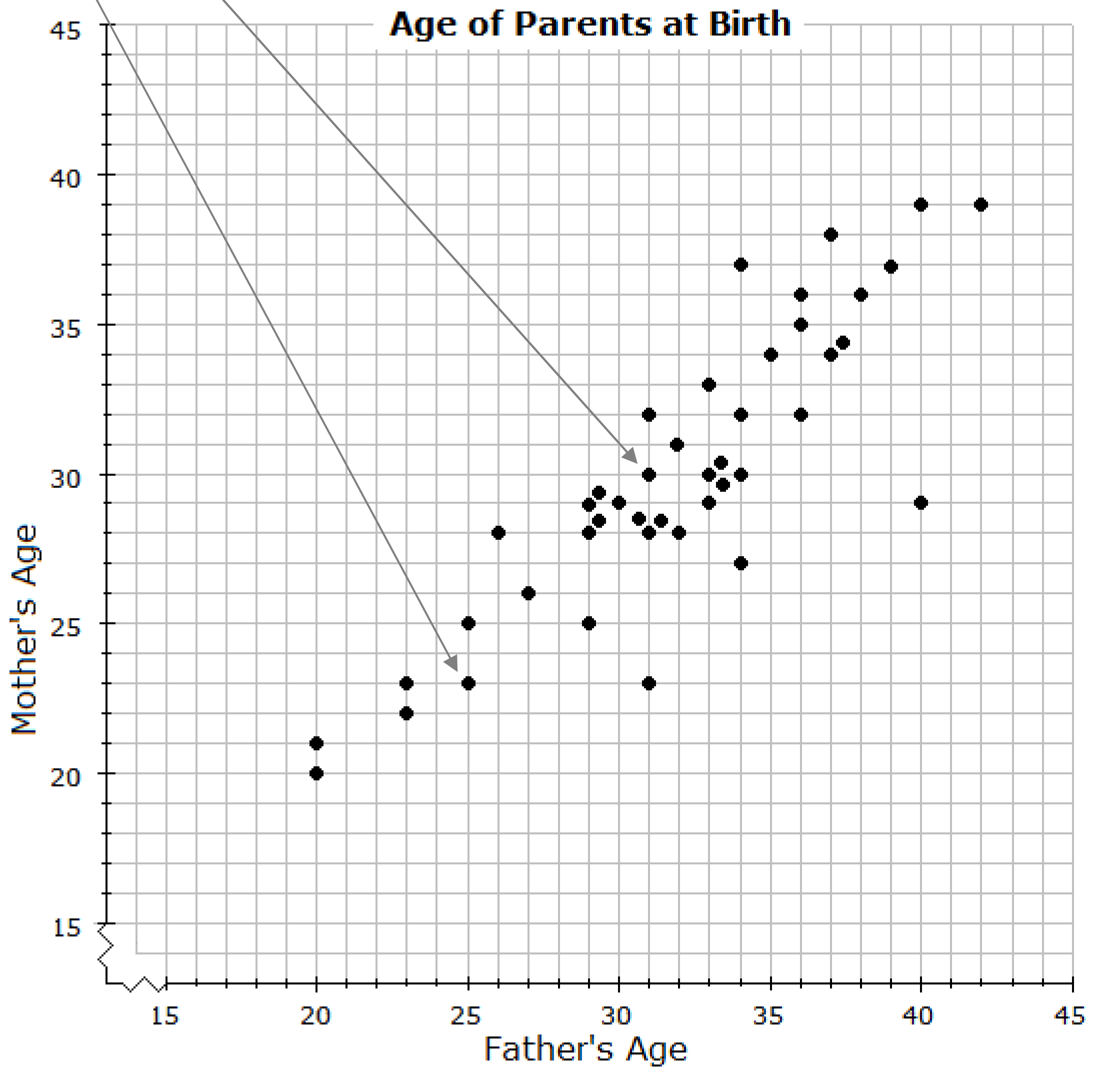
A scatter plot is used when there are two variables measured for each person or thing.

The distance along one scale is crossed with the distance along the other, and their 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 select the scale to spread the dots as widely as possible. If the axes do not start at zero, they are marked with a zigzag line or similar.

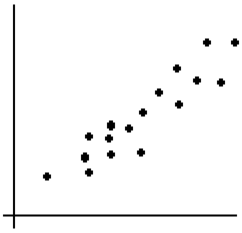
## Correlation

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.

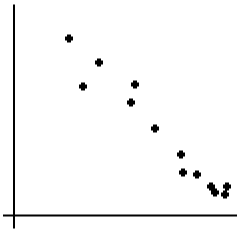
A correlation is positive if as the *x* variable increases the *y* variable increases. A correlation is negative if, as the *x* variable increases, the *y* variable decreases.

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.

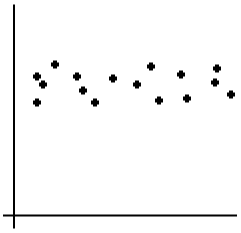
There is no correlation if the *x* variable or *y* variable are more or less constant regardless of the other – that is, a basically horizontal or vertical line – even if the line is very obvious.



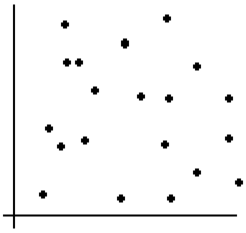
Positive Correlation



Negative Correlation



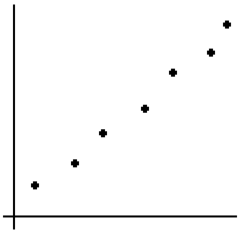
No Correlation



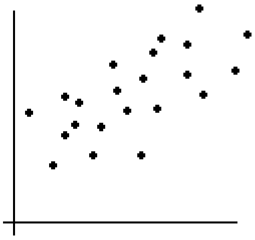
No Correlation

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.

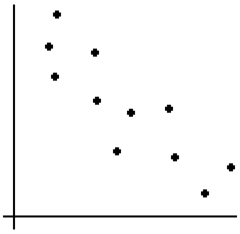
Note that “strong” and “positive” are not related concepts.



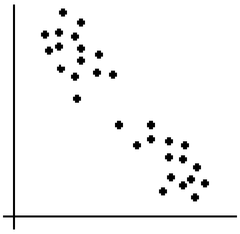
Strong Positive



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.  Students are not expected to get the line of best fit perfect. It may be necessary to ignore outliers.

Do not try to make the line go through (0, 0) unless there is a reason to believe it should.

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.

## Analysis and Conclusions

Every analysis must have a conclusion. At Achieved level the conclusion need only be as simple as noting a relationship (correlation) appears to exist, and describing it briefly.

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.

### Analysis

Bivariate analysis with scatter graphs is for comparing the two different measurements taken. The point is to find the **relationship** between one measurement and another. There is no point discussing one measurement by itself without its relationship to the other. (For this reason we do not find means, medians, quartiles etc.)

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.

Give its meaning in general terms first – “we see that the heavier students eat the most lollies”. Give it in terms of numbers if you can – “for each 10 kg heavier the student is, the line of best fits shows on average they eat 5 more lollies a day”.

You can use your line of best fit to make predictions. If so, discuss how accurate that is likely to be.

If you think it makes sense in the real world that there should be a correlation, explain that briefly, but do not assume because there is a correlation that there must be a direct cause.

Remember that you can only really talk about the data you have – do not try to make it represent groups other than that sampled. For example, a relationship between lack of sleep and amount of texting found in a sampling of 14 year old boys should not be stretched to say “quantity of texting is correlated with lack of sleep”, without adding “in 14 year old boys”. (You might suggest that this is likely to be true in similar groups, but only testing will determine if the suggestion is correct.)

Note any outliers or clusters. Note any other patterns of interest. Say what their effect is.

Make any comments about problems you faced, any limitations of your analysis, and any improvements that could be made.

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.

### Conclusion

Should be a brief repeat about what trend you see in mathematical terms (correlation which is positive or negative, strong or weak).

Discuss briefly what that might mean in the real world.