THE STATISTICAL PROCESS

Real-world investigation

Statistics can easily be misused or misrepresented, sometimes deliberately. For example, headlines such as 'Children with bigger feet read better' and 'Head injuries have soared since bike helmets were made compulsory' could lead you to conclude that stretching your feet will improve your reading or that not wearing a bike helmet will prevent head injury.

Both statements in the headlines may be partly true but neither represents the whole truth.

The reality is that older children have bigger feet and also better reading skills, simply because they are older. The introduction of compulsory helmets has meant that bike accidents which previously would cause death, now instead cause only injuries, leading to a higher rate of injuries overall but a lower rate of deaths. Removing the helmet would of course increase the death rate again.

Statistics are created by humans. The process of collection, analysis and interpretation of data must be carefully done to ensure the statistics are as accurate and 'fair' as possible. This may seem straightforward, but there are issues to be addressed at each stage if you want the result to be fair and unbiased.

One way of describing the statistical investigation process is as follows.

Step 1: Clarify the problem and formulate one or more questions that can be answered with data.

Write research questions in such a way that answers can be acted upon.

For example, 'Should the canteen sell sushi?' rather than, 'Do we need more options on the canteen menu?'

Step 2: Design and implement a plan to collect or obtain appropriate data.

The questions asked must be clear and easily understood.

The questions asked must be closed, i.e. have only a small number of alternative answers. Open-ended questions will usually produce data that is hard to categorise; it is best to design your questions so the categories are already there.

The questions must be neutral and not lead the person completing the questionnaire to a particular conclusion.

For example, both of the following are 'leading questions':

- Street violence is on the increase, are you in favour of teaching people to use guns?
- Unemployment is an issue, are you in favour of compulsory military training for all?

In statistics, a population is defined as every possible member of the relevant group. If you are unable to interview or survey the population then a sample must be selected without bias.

For example:

Your school is interested in finding out if all students support the idea of the Year 12 students having a special lounge to themselves. If you want a 'yes' from the survey then only ask the Year 12 students, but a more representative sample would be to take 10 students from each year level. This is called taking a stratified sample.

Phone surveys taken during the day may be biased towards the opinions of those who are not working; surveys taken before 11 am on weekends are unlikely to include the opinions of many teenagers, etc.

Step 3: Select and apply appropriate graphical or numerical techniques to analyse the data.

Step 4: Interpret the results of this analysis and relate the interpretation to the original question; communicate findings in a systematic and concise manner.

It is important to remember that in the following exercise you are only looking at bivariate data. Any investigations considered here are looking at the presence or absence of an association between two variables.

Using the statistical method: a case study

Joseph is establishing a new business across different countries and wants to set salary levels that will attract the best people, while keeping within his budget. Two of the positions he needs to fill are the positions of financial manager and human resources manager. Help Joseph do the research and investigation into the relative salaries he should offer for both positions.

Step 1: Clarify the problem and formulate one or more questions that can be answered with data.

Data needs to be found to answer the question 'What is the relationship between the salary of the financial manager and the human resources manager in a company?'

Step 2: Design and implement a plan to collect or obtain appropriate data.

The following data shows the average salary of the same two positions in cities across the world.

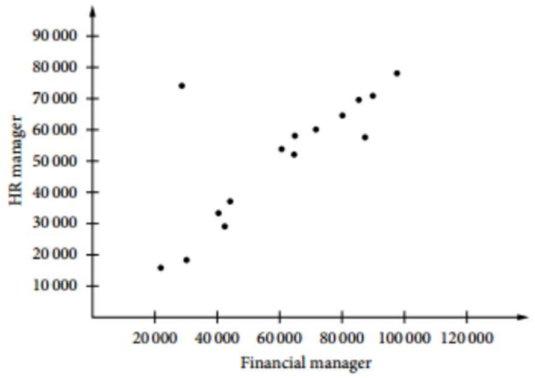
City	Financial manager ($AUD)	Human resources manager ($AUD)
Alor Setar	28 670	74 020
Amsterdam	87 150	57 760
Berwick	71 320	60 220
Christchurch	64 610	52 280
Glasgow	85 210	69 540
Kuala Lumpur	29 950	18 400
Luang Prabang	40 280	33 240

City	Financial ($AUD)	Human resources manager ($AUD)
New Dehli	42 130	29 020
Phakse	21 820	15 850
Singapore	89 560	70 960
Tripoli	64 950	58 000
Vienna	97 480	77 840
Wagga Wagga	79 790	64 690
Wellington	60 470	53 910

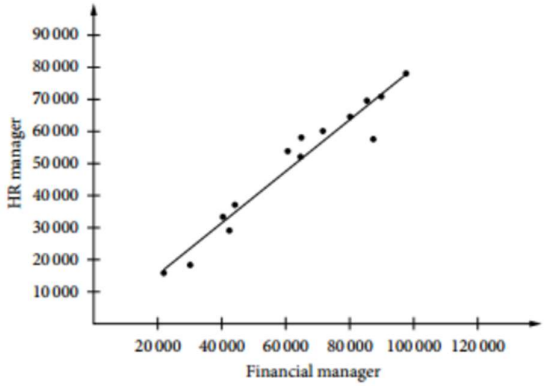
Step 3: Select and apply appropriate graphical or numerical techniques to analyse the data.

Check to see if there is an independent variable. In this case there is not an independent variable, as both variables seem to vary according to the relative wealth of the country.

Draw a scatterplot to visually show any trend.



The scatterplot picks up an outlier at Alor Setar (28 670, 74 020). This is removed from the data set before placing the line of best fit and calculating the regression analysis.



Line of best fit can be obtained using your calculator or approximating the gradient and the y-intercept from the graph.

From a calculator, the line of best fit is: $y = 0.81x - 632.7$

The correlation coefficient is: $r = 0.97$

Step 4: Interpret the results of this analysis and relate the interpretation.

The linear regression equation can be written as:

(Human Resources Manager Salary) = 0.81 × (Financial Manager Salary) − 632.7

There is a strong, positive, linear correlation, although in this case there is no obvious independent variable and the data is exhibiting a common response to the local situation.