

# Presenting statistical information – Graphs

## Introduction

This document provides guidelines on how to create meaningful, easy to read and well-formatted graphs for use in statistical reporting. It contains a brief overview of the most common types of graphs, and hints and tips on presentation.

Well-presented graphs can greatly enhance the readability of a research report. Graphs simplify data and display key findings in pictorial form. A graph is a visual representation of data that can present complex information quickly and clearly, and assist the reader to see patterns and trends in data.

Generally speaking, a graph is an effective way of communicating data when:

- precise numeric details are not required
- a trend or comparison can be demonstrated
- there are relationships between data values.

## Key terms

The following terms are used in this paper. Fuller definitions and further information follow in the document.

Term	Brief definition
Axes	a graph consists of two axes called the x-axis (horizontal) and y-axis (vertical)
Categorical data	data which fall into one of two or more discrete categories, but with no intrinsic ordering to the categories
Confidence interval	provides a range of values around the estimate, within which the true value can be expected to fall
Continuous data	data on a numeric scale
Interval data	data that have both an order, and an equal spacing between categories
Legend	provides a key to the various data plotted on a graph
Ordinal data	similar to categorical data, except there is a clear ordering to the variables
Statistical significance	a statement about the likelihood of findings being due to chance

## Graph formatting

Appropriate formatting and presentation of graphs enhance the reader's capacity to understand the information being presented and should include:

- titles, axis labels, legends and footnotes
- appropriate representation of axes, scale and error
- a visual style that is clutter-free, easy to interpret, and appropriately represents trends or differences in the data.

### Titles, axis labels, legends and footnotes

Place a title on the graph, summarizing what the graph is showing. The title is usually placed in the centre, either above or below the graph.

Apply a numbering system to the titles of all graphs so they can then be easily referenced in a summary list or in the appendices of the report. For example, Figure 1: [title], Figure 2: [title] etc.

The title of the graph should explain what the x and y axes represent. For example, if comparing the amount of fertilizer used on rose bushes with the number of roses yielded on average per bush, the amount of fertilizer would be the independent variable and the number of roses would be the dependent variable. Therefore, the title should be "Average number of roses per plant by amount of fertilizer."

Include labels for both the x-axis (horizontal) and y-axis (vertical): labels should be brief and explain exactly what each aspect of the graph is showing. The units of measurement (dollars, metres etc.) should also be included.

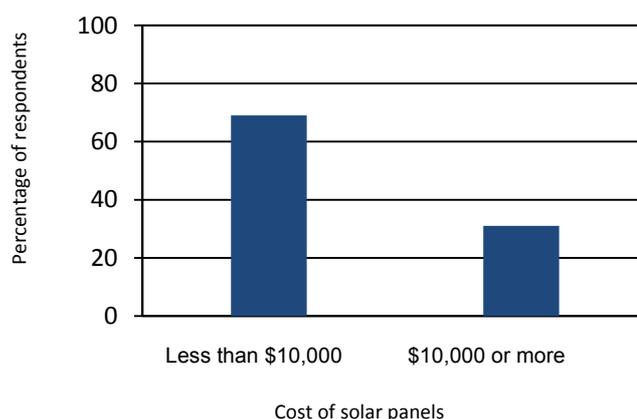
Include a legend. This provides a key to the various data plotted on a graph. For example, if you have used colours or shading, the legend should explain what the colours and/or shading represent.

Add footnotes to further explain the data. For example, for a sample survey, include a footnote describing the sample that is being represented in the graph, and the number of respondents in the sample (n). If preparing a graph showing the percentage of people who have a solar panel on their roof, where 1,000 people were interviewed in total, an appropriate footer would be "Base: All respondents (n=1,000)". If 30% of respondents reported having a solar panel, and they were asked how much they paid for their solar panels, an appropriate footer for the graph would be "Base: Respondents who have a solar panel on their roof (n= 300)". Including an appropriate base on the graph allows the reader to see how many people answered the question, and make a quick assessment of the likely accuracy of the results based on the sample size.

A footnote should mention the source of the data.

An example of appropriate title, axes labels and footnote is shown below.

Figure 1: Cost of solar panels



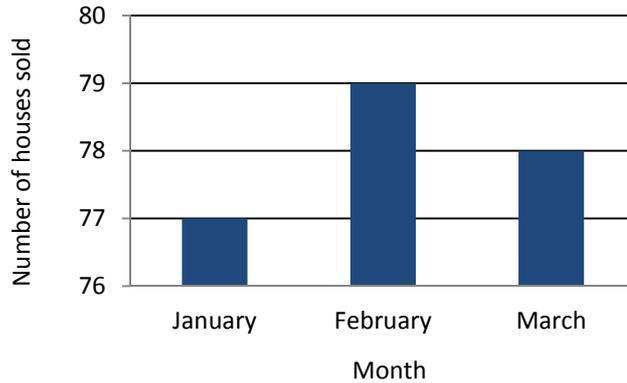
Base: All respondents who have a solar panel(s) installed on their roof (n=300)

**Please note:** the data contained within all of the example graphs throughout this document are entirely fictitious and are intended simply for illustration purposes.

**Axes and scales**

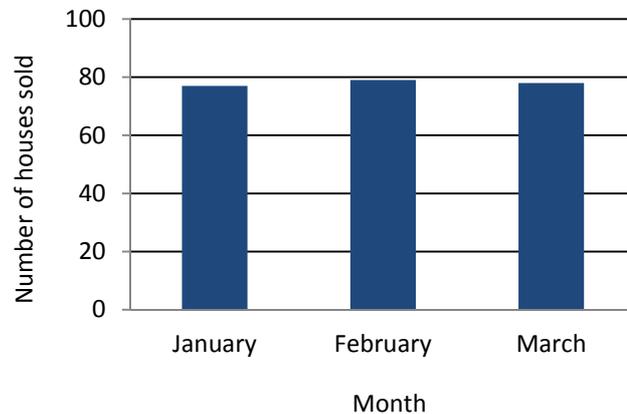
Start the vertical axis at zero (0), so that the range of the axes does not distort the data and allow for misinterpretation. The graphs below present exactly the same data, but with different scales:

**Figure 2a: Houses sold in Kedron, 1<sup>st</sup> Quarter 2014**



Source: Sales database, xyz Real Estate, 1<sup>st</sup> Quarter 2014

**Figure 2b: Houses sold in Kedron, 1<sup>st</sup> Quarter 2014**



Source: Sales database, xyz Real Estate, 1<sup>st</sup> Quarter 2014

In the top graph, the scale distorts the data by making February’s sales appear much higher than those in January and March. It is only upon close inspection of the y-axis that a reader sees the difference was only two sales. In this case, the bottom graph is more appropriate. It shows the true result — that sales were fairly steady over the three months.

An exception to this rule is when there are negative values, in which case the scale would start at less than zero.

## Indicating range of error

Results represented in graphs are often estimates. They may not be precise figures, but may fall within a range with a certain level of confidence. A confidence interval gives an estimated range of values taken from a set of sample data. The range of values is likely to include what the 'true' value would be if the entire population were to be surveyed. Confidence intervals indicate the range of error that may be contained in the results.

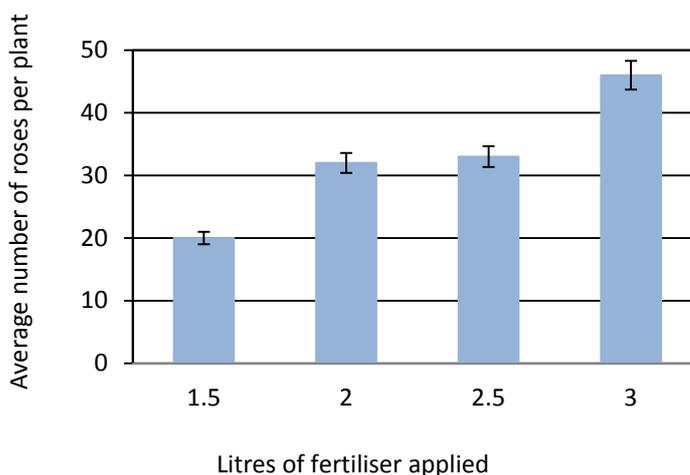
When reporting exact known figures, such as actual number of sales made by an individual company in different months of the year, taken from the company's sales reports, then confidence intervals will not be necessary. Similarly, if a graph shows results from a census (where an entire population is surveyed), then confidence intervals will not be necessary. The use of confidence intervals is only necessary when a sample of a population is surveyed, and the results are an estimate of the true value that would result if the whole population were to be surveyed.

Statistical results are often presented using the 95% confidence interval. The 95% confidence interval is the range of values within which there is a 95% chance the true population value lies. The error bars show the upper and lower limits for the estimate, i.e. the range within which we can be 95% confident that the true value lies.

In graphs, confidence intervals are typically displayed through the use of error bars. Including error bars allows the reader to easily determine whether there is a statistically significant difference between any estimates shown in the graph. If the error bars for any response categories do not overlap, then there is usually a statistically significant difference in the estimates for those response categories. If the error bars do overlap, then there is no statistically significant difference in the results. In the example shown below, plants given 1.5 litres of fertiliser yielded significantly fewer roses on average than all the other plants. There was no statistically significant difference in the number of roses on the plants given 2 litres of fertiliser or 2.5 litres of fertiliser. Plants given 3 litres of fertiliser yielded significantly more roses on average than all the other plants.

Error bars are easy to include on bar graphs, but can add clutter or even be impossible to include on other types of graphs.

Figure 3: Average number of roses per plant by litres of fertiliser



Base: All rose plants in xyz Nursery, January 2014 (n=400, 100 rose plants per group)

## Visual style

Reduce clutter — less is more: graphs should be as simple as possible and only communicate the required information.

Highlight what is important, and leave out what is not. An efficient graph allows readers to focus on what is really important, whether that be numbers, relationships, trends and/or anomalies. It's easy to crowd a graph with patterns, formatting and numbers. Superfluous decorations (such as three dimensional effects, colour gradients, shadows etc.) can distract the reader or make a graph more difficult to interpret. However, the inclusion of all of these features can detract from the clarity and readability of the graph.

Data ink consists of numbers (the scale) and the vital points representing the data — for example, the points on a scatter plot. Conversely, non-data ink includes any titles, headings, legends, grids and borders, and while these are still important, they should not be overused. Generally speaking, most ink used to present a graph should be data ink. Excessive non-data ink can create clutter and distract readers from key information. This relates to the 'less is more' principle.

Colours and patterns can be a useful way of illustrating information. For example, it is sometimes necessary to use colour to differentiate between categories. On the other hand, they should not be overused, as unnecessary use of colours and patterns may distract readers from the important information being presented. Choose colours with care, and apply them consistently.

When deciding whether to use colour or black and white, consider how the graph will be displayed. When coloured graphs are printed in black and white, they can appear with various (often unpredictable) shades that may make the graph difficult to interpret. For this reason, if a graph will be printed in black and white, it should ideally be created in black and white.

Dimension - graphs should be two-dimensional (2D) wherever possible. Three-dimensional (3D) graphs can distort scale as information toward the front of 3D graphs appears larger than information towards the back. This distortion is misleading and can confuse the reader.

## Types of data and choice of graph

There are many different types of graph, each suitable for different types of data. The type of graph selected for use depends on the type of data being represented.

**Categorical data** are data which fall into one of two or more discrete categories, but with no intrinsic ordering to the categories. For example, sex is a categorical variable with two categories (male, female). Hair colour is another categorical variable, but with a number of categories (blonde, brown, auburn, black etc). For purely categorical data, the variables do not have a clear order. Graphs suitable for categorical data include *bar graphs* (both horizontal and vertical), *clustered bar graphs* and *stacked column charts*.

**Ordinal data** are similar to categorical data, except there is a clear ordering to the variables. While ordinal data has a definite ordering, the degree of difference between categories is not always consistent or measurable. For example, highest level of education might be (very simplistically) classified into primary school, secondary school, some tertiary study, completed tertiary study. There is likely to be a bigger difference between respondents having educational experience at the primary school level versus the secondary school level, than there is between those undertaking some tertiary study, and those who have completed tertiary study. *Bar graphs* and *histograms* are suitable for presenting ordinal data.

Data on a numeric scale are usually thought of, and treated, as **continuous data**. For example, numbers of people scoring a grade between 1 and 100 on an examination would constitute continuous data. Other examples of continuous data include height, time, mass, distance and dollar values. Graphs suitable for continuous data include *line graphs* and *scatter plots*.

**Interval data** are data that have both an order, and an equal spacing between categories. For example, if a survey asks respondents to nominate how much money they have saved in the bank for an emergency, and the available response categories are '\$0-\$4,999', '\$5,000 to \$9,999', '\$10,000 to \$14,999' and so on, then the data are interval data. Sometimes, continuous data are converted to interval data for reporting purposes. Graphs suitable for interval data include *bar graphs*, *histograms* and *box-whisker plots*.

## Examples of graphs

Displaying data in a manner that is both professional and easy to understand is a crucial skill. Choosing the right graph for the data helps to ensure that the results are presented in a clear and concise format. Some examples of recommended graphs and the type of data they are suitable for are presented below.

### Bar graphs

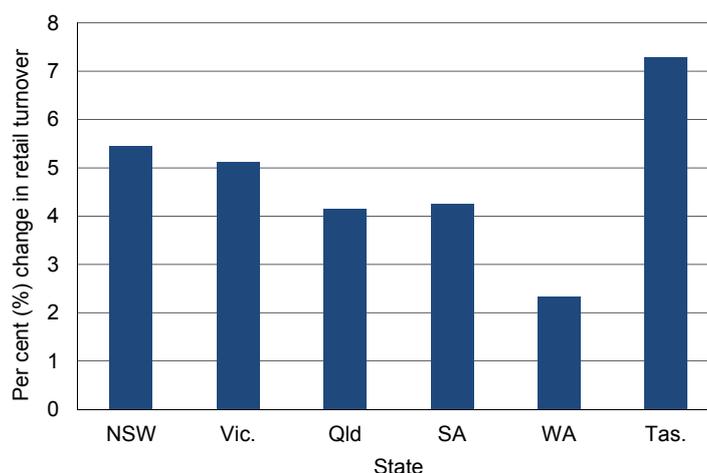
Bar graphs are ideal for presenting categorical data. Bar graphs use rectangular bars to visually display each value and how it compares to other values in the graph — the greater the length of the bar, the greater the value. This provides a simple and easy way to interpret the data. The bars on a bar graph can be horizontal or vertical. It is easy to include error bars on this type of graph. Bar graphs are therefore popular and commonly used in survey reporting.

#### Vertical bar graphs

Vertical bar graphs are best for comparing estimates (means or percentages) for between two and seven groups. Any more than seven groups can tend to make the graph crowded, and result in difficulties labelling the x-axis. When dealing with more than seven groups, a horizontal bar graph is generally a better choice.

The graph below shows annual change in retail turnover for convenience stores in December 2013 for six Australian states.

Figure 4: Annual percentage change in retail turnover for convenience stores by state  
July 2012 – July 2013



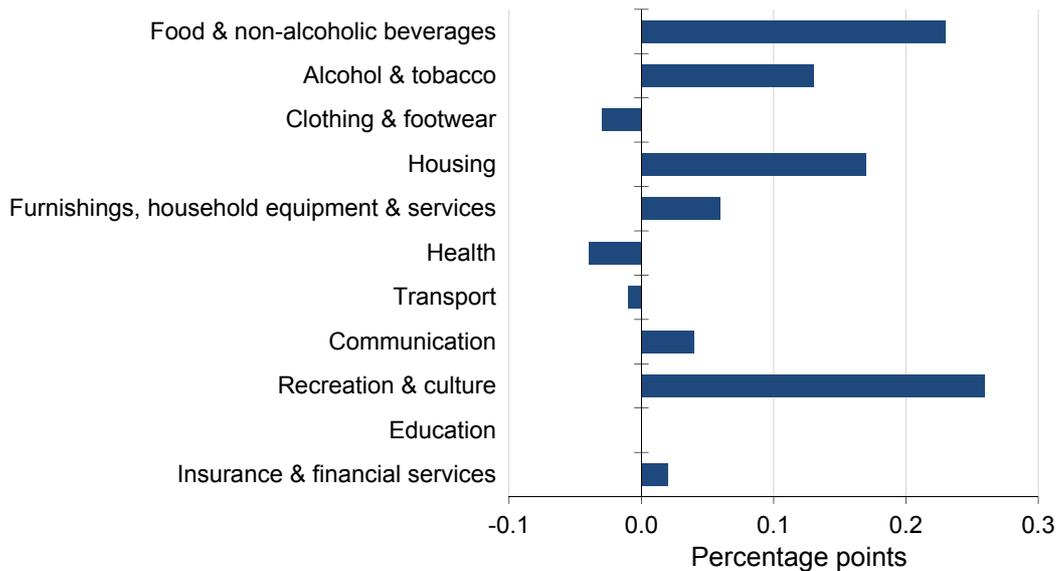
Base: All convenience stores in Australia (n=42,200)

**Horizontal bar graphs**

These are the same as vertical bar graphs, but turned on their side. Horizontal bar graphs are best for showing categorical data when comparing estimates (means or percentages) for eight or more different groups. Horizontal bar graphs are also appropriate to use when the category labels are too long to appear neatly on the x-axis.

The graph below shows quarterly Consumer Price Index (CPI) contributions by retail group for the December quarter, 2013. Note the presence of a larger number of categories with longer labels than those of the previous example.

**Figure 5: Quarterly CPI contributions, by retail group, December quarter 2013**



Source: XYZ retail group census

**Clustered bar graphs**

Clustered or grouped bar graphs are bar graphs that show two or more categories on one graph. Plotting multiple categories on one graph increases the amount of information that can be shown, although care must be taken to avoid over-complicating the graph.

The graph below shows migration between New Zealand and Queensland across three financial years, including three categories: arrivals, departure and net migration.

**Figure 6: Migration between New Zealand and Queensland, 1 July to 30 June**



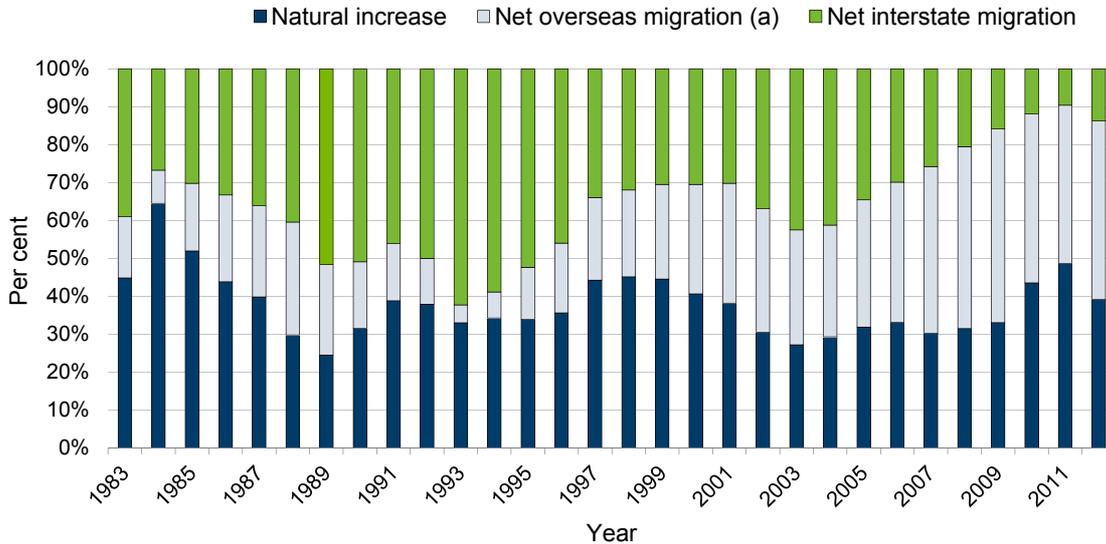
Source: XYZ arrivals and departures data

### 100% stacked column graphs

One hundred percent (100%) stacked column graphs compare the percentage that each value contributes to a total of 100% across categories. These graphs are best used for categorical data when each column is comprised of no more than three components. Otherwise, trying to compare across categories can become difficult.

The example below shows how components of population change differ across years.

Figure 7: Components of population change, Queensland, year to 30 June



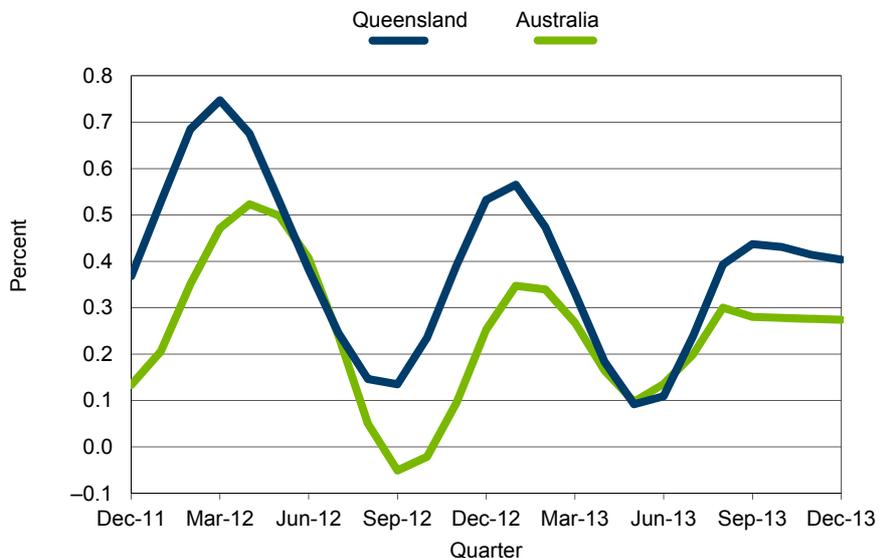
(a) Net overseas migration estimates for calculating population growth use an improved methodology from June quarter 2006 onwards and are not directly comparable with net overseas migration estimates from earlier periods.

Source: XYZ arrivals and departures data, June 2014

### Line graphs

Line graphs are used to illustrate trends over time for continuous data. They can also be used to compare two different variables over time. In the example below, monthly changes in retail turnover are plotted for both Queensland and Australia, and show that the trends behave similarly with a slight time delay.

Figure 8: Monthly change in retail turnover for Queensland and Australia



Source: XYZ retail group census 2011, 2012 and 2013

## Histograms

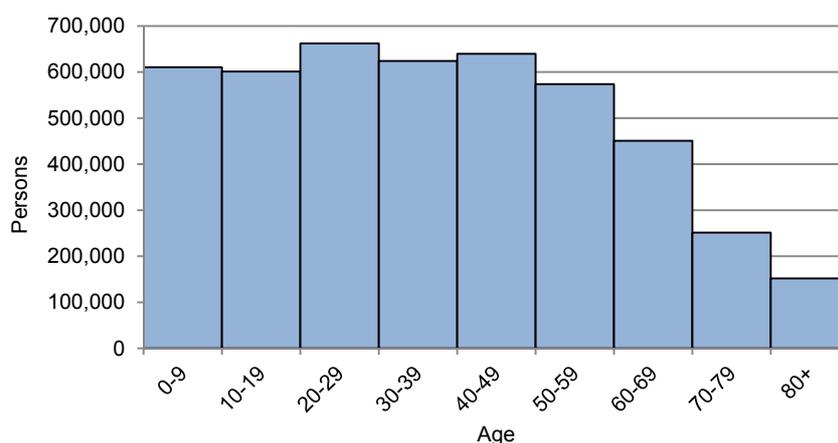
In a histogram, the data are grouped into ranges (e.g. 10–19, 20–29) and then plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category, and the height is proportional to the frequency or percentage of that category. Bars or categories in a histogram are presented in ascending or descending order. Histograms are used for data that are at least at the ordinal level of measurement, and most often for plotting continuous data.

A histogram looks very similar to a bar chart except that the bars are touching and may not be of equal width. In a bar chart, the spaces between the bars visually indicate that the categories are separate. Examples of variables with separate categories are *marital status* (married, single) and *sex* (male, female). In a histogram, the touching bars indicate that the categories or intervals are ordered from low to high in a meaningful way. For example, the categories of the variables *hours spent studying* or *age*, are ordered intervals.

A histogram shows the underlying frequency distribution (shape) of a set of continuous data. Histograms clearly show outliers (a value that lies outside most of the other values in a set of data) and skewness (any lack of symmetry in the data).

The example below illustrates the distribution of the population across age ranges in Queensland.

Figure 9: Age distribution for Queensland population, 30 June 2012



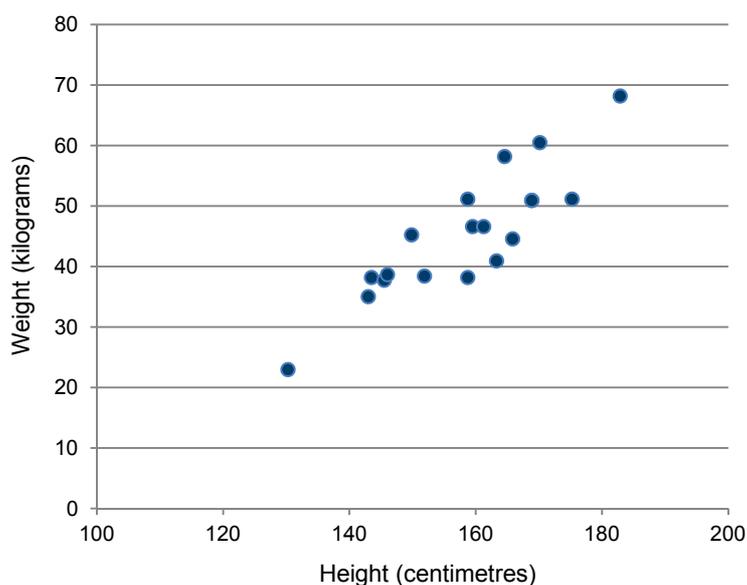
Source: XYZ census data, June 2012

## Scatter plots

Scatter plots are used to plot data points on a horizontal and a vertical axis to show relationships between two variables (or how much one variable is affected by another). Scatter plots are used for plotting continuous data. They are particularly useful when comparing two variables in situations where there are many data points, the measurement intervals on the x and/or y-axis may be uneven, and/or the reader is looking for trends and groupings in the data.

The following example demonstrates what appears to be a linear relationship between heights and weights of a group of high school students.

Figure 10: Weight by height, Grade 8 school students, Class 8C. Kenmore High



Base: All students in class 8c at Kenmore High School (n=18)

## Box whisker plot

A box whisker plot is a graphical method of displaying variation in a set of data. This type of graph is often used in exploratory data analysis and is suitable for data measured on an interval scale. It shows the shape of the distribution, its central value, and its variability. Specifically, a box whisker plot shows the following:

- the minimum value — smallest number in the dataset
- the lower quartile — the value below which the lower 25% of the data are contained
- the median value — or middle number in a set of numbers
- the upper quartile — the value above which the upper 25% of the data are contained
- the maximum value — largest number in the dataset.

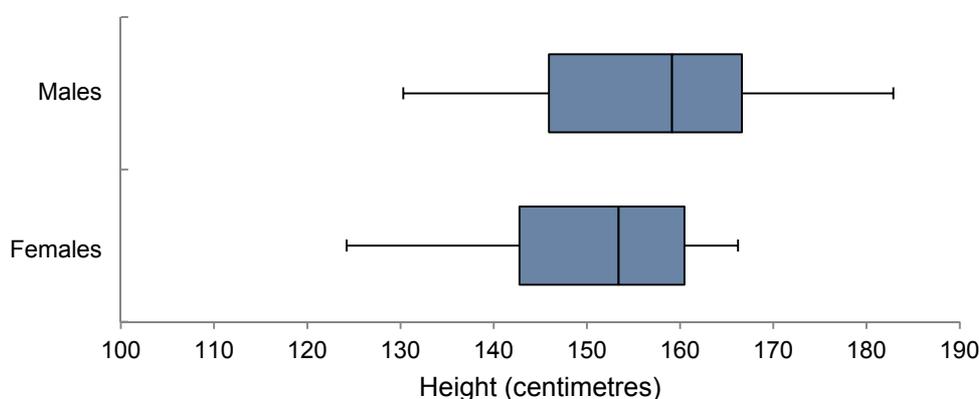
This group of statistics is sometimes called a five number summary. Box whisker plots are named for their design, which includes a box around the lower quartile, median and upper quartile, and two 'whiskers' that extend to the minimum and maximum.

A box whisker plot is especially useful for indicating whether a distribution is skewed and whether there are unusual observations (outliers) in the dataset. If there are outliers in the data, these are represented by dots outside of the whiskers, and the whiskers extend to the minimum and maximum values in the dataset *excluding the outliers*. If there are no outliers, then the whiskers extend to the actual minimum and maximum value.

Box whisker plots are also very useful when large numbers of observations are involved and when two or more datasets are being compared. Box whisker plots are ideal for comparing distributions because the centre, spread and overall range are immediately apparent.

The example below shows box whisker plots for male and female student heights, and enables us to compare the two distributions.

Figure 31: Heights of Grade 8 students, Kenmore High, at 30 June 2014



Base: All students in Grade 8 at Kenmore High School (n=120, 60 per group)

## Pie graphs

Pie graphs show parts or percentages of a whole. While they are colourful and popular, they have limitations:

- It is difficult to tell the difference between estimates of similar size.
- Error bars or confidence limits cannot be shown on a pie graph.
- Legends and labels on pie graphs are hard to align and read.
- The human visual system is more efficient at perceiving and discriminating between lines and line lengths (such as are found in bar graphs), rather than two-dimensional areas and angles (such as are found in pie graphs).
- Pie graphs simply don't work when comparing data.

## Summary

Graphs are a way to present data in visual form. They are a useful tool for displaying many types of data, and one of the easiest ways to see relationships between variables and/or compare numbers. To ensure that graphs are easy to interpret, they need to be presented in a way that enables them to “stand-alone”. They should be clutter free and use appropriate titles, legends, axis titles and footnotes. There are a range of different types of graphs that can be used. Care should be taken to ensure that the type of graph chosen is appropriate for the type of data that is being plotted. An appropriate and properly prepared graph can be a powerful way in which to convey statistical information.

## Further information and advice

The Queensland Government Statistician's Office has extensive experience in a broad range of data presentation techniques and is able to provide specialist consultation and advice on request. For more information please see the QGSO website, [www.qgso.qld.gov.au](http://www.qgso.qld.gov.au), or email [govstat@treasury.qld.gov.au](mailto:govstat@treasury.qld.gov.au).