3. Types of graph

Having explained the principles underlying perception, we can apply these to the various types of graph.

3.1 PIE GRAPH (UNIVARIATE)

Pie graphs, pie charts or pie diagrams have no right to exist in science: the job they do can always be done much better in other ways. They are generally used for data with one numeric and one categorical variable, and display only a few data but take up a lot of space. Moreover, they represent the information as angles, which is low on the scale of decoding accuracy (section 2.3.3). Even worse are 'mock 3D' pies (Fig. 8A), which add insult (distortion) to injury (inaccuracy); they violate the stated rule that the number of data dimensions in a graph should not exceed the number of dimensions in the source data.

Generally speaking, pie-graph data are much better presented in a small table or as horizontal bar graphs (Fig. 8B). Note that many pie-graph designers admit the limitations of pies by adding numeric values and / or percentages to the individual pie segments, thus creating clutter. Pie graphs also often require a detailed key, which more often than not creates extra confusion: colours or shadings are often too similar to clearly identify the segment to which they belong. Generally, a key 'starts at 12 o'clock' and subsequent categories are then listed in clockwise order… but not many readers know that!

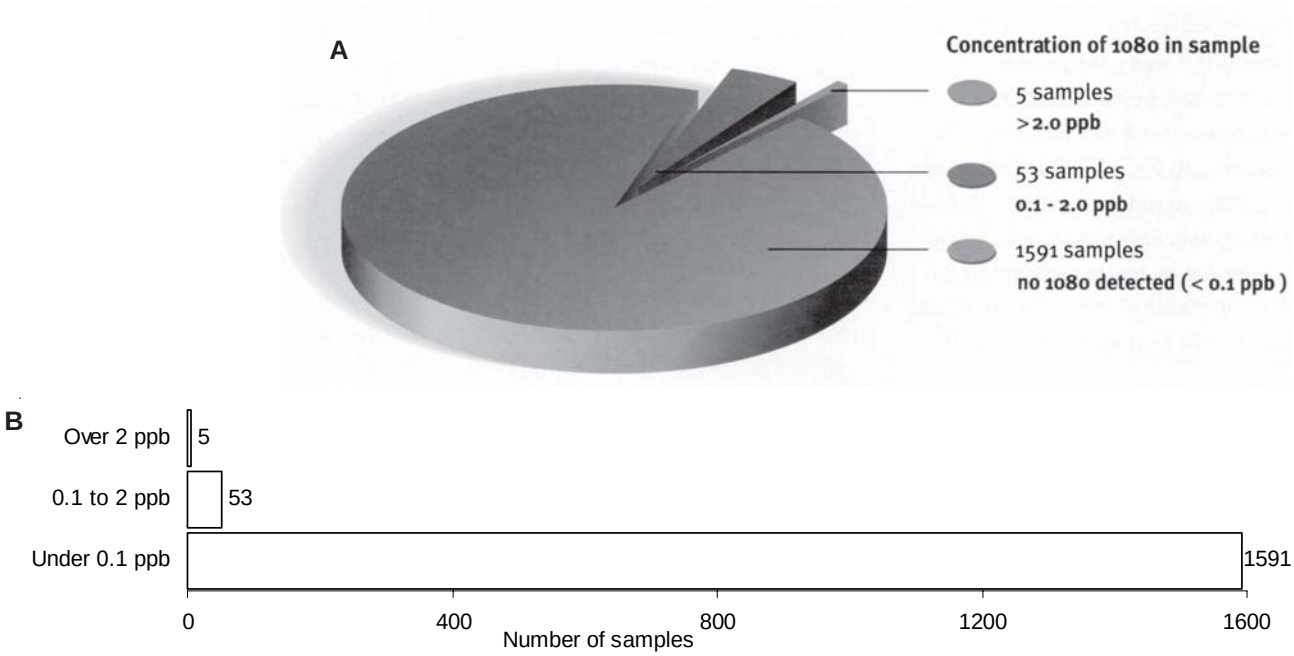


Figure 8. A classic example of a space-wasting pie graph (A), which still requires a table to explain its values. In B, the data from A, particularly the relative sizes of the samples, are much more accurately represented by horizontal bars.

Original caption to A: Results of water monitoring after aerial 1080 operations (1991–2003).

Bigwood & Spore (2003) agree that 'despite their mass popularity, pie charts do not communicate well' (but these authors 'offer some advice on designing and presenting them' in order to 'use them as effectively as possible').

You sometimes see linked pie graphs, where there are several in a row. Instead, if you have three categories that each add to 100%, scored at a number of different sites or samples, consider using a triangular diagram, sometimes called a 'ternary plot'. An example is given in Fig. 9.

In most instances it may be best to represent data from linked pies as a series of column graphs where each column adds up to 100% (Fig. 10). The columns represent the data as length, not angle, and you can run your eye across the values for each category more easily than if they are in pies. Column graphs (bar charts) are discussed in much more detail below.

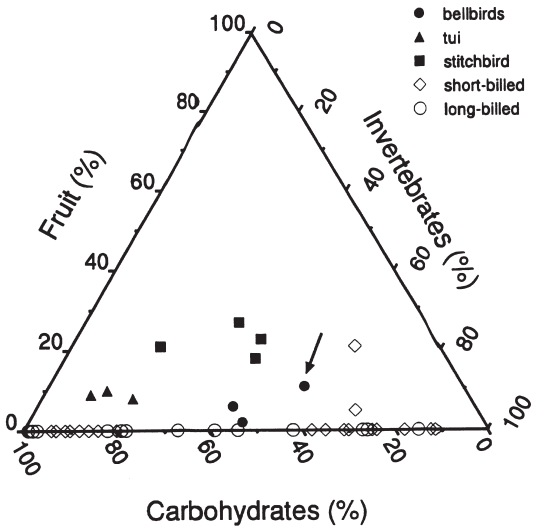
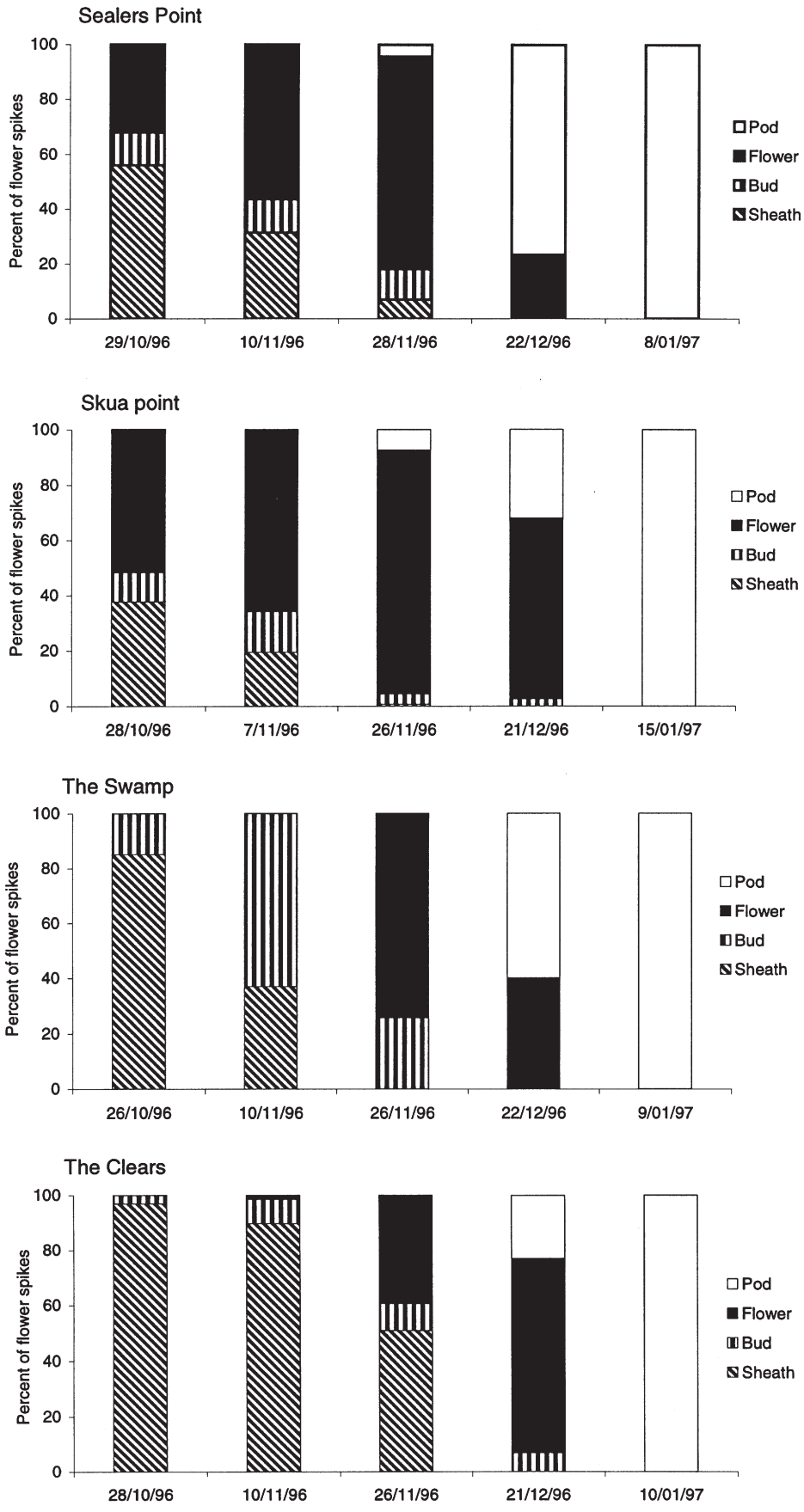


Figure 9. A ternary (triangular) graph, useful for three variables that sum to 100%. These graphs can be difficult to interpret on first encounter. It is easily grasped that the three corners represent 100% of one of the variables and 0% of the other two. In contrast, it is much less obvious that the point dead centre does not represent 50, 50, 50 for a sum of 150%. The reason it actually represents 33, 33, 33 to sum to 100% is that the gridlines run on different angles for the three axes. The left axis (in this case Fruit) gridlines run horizontally; the right axis (Invertebrates) gridlines slope downwards to the left, parallel to the Fruit axis line; and the lower axis (Carbohydrates) gridlines slope upwards to the left, parallel with the Invertebrates axis line. It helps to indicate this if (a) the axis tick mark labels are angled, as here; and (b) the graphs use long angled tick marks (in this case, they are angled, but perhaps too short).

Original caption: Annual mean diet composition of different New Zealand (solid symbols) and Australian (open symbols) Meliphagidae species. Each point on the graph represents the annual mean diet for a species from a single study or site, comprised of the annual mean percentages of the three major Meliphagidae food groups: invertebrates, fruit, and carbohydrates (nectar, honeydew, lerp and manna). Australian species are classified as long-billed or short-billed to distinguish between the two main feeding guilds in the Australian Meliphagidae. The Craigieburn bellbird data are marked with an arrow.

Figure 10. Example of a good 4-by-5 grid of split bars. However, the fills used in the bars run some risk of Moiré effects, see section 4.7.4. Also, the duplication of vertical labels and keys is unnecessary, and the *y*-axis label should read 'Percentage of flower spikes'.

Original caption: Flax flowering at selected plots on Rangatira Island.



3.2 VERTICAL AND HORIZONTAL BAR CHARTS / DOT GRAPHS

Bar graphs can be very clear, but they are overused and there are often better alternatives. Bar graphs tend to have a fairly low information density. They are easy to create using computer software packages such as Microsoft Excel; however, Excel tends to produce graphs that are not readily publishable to a high standard. Appendix 1 describes how such default graphs can be modified to meet science journal publication needs. More than 50% of graphs in DOC Science publications of 2002/03 were bar graphs, mostly produced in Excel—hence our concern with improving them (Appendix 1).

3.2.1 Notes on terminology

What most people call a bar chart has vertical bars, in distinct categories usually separated by white space. Microsoft products, however, call this type a 'column chart'. They are the most commonly seen graph type in all sorts of publications. The vertical arrangement often forces labels on the *x*-axis to be squashed or turned (up to 90°), which makes the axis hard to read, and looks ugly.

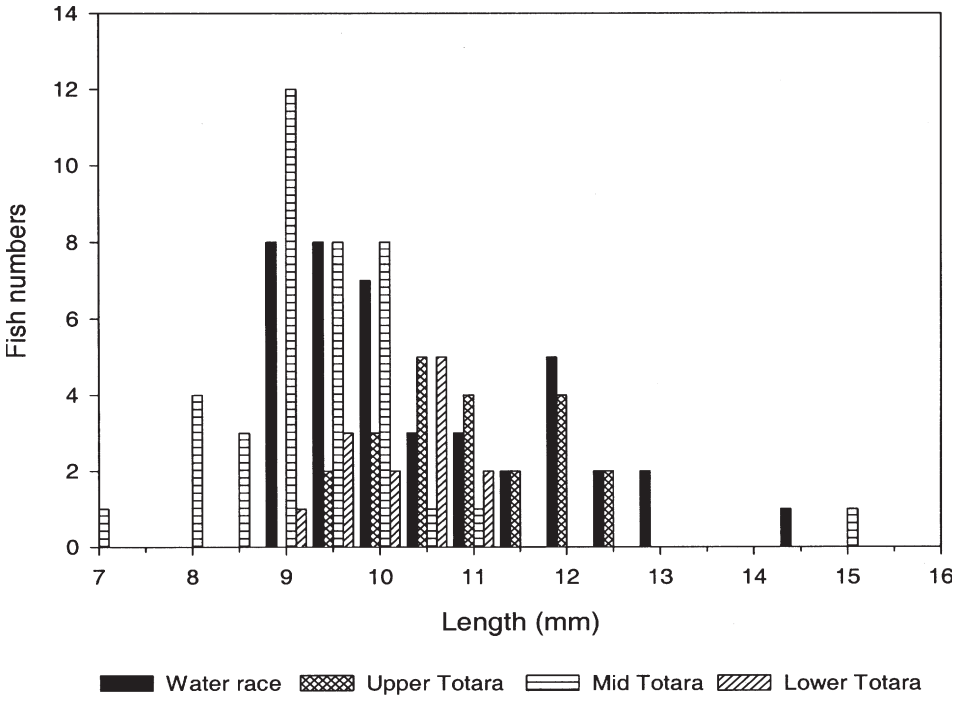
Horizontal bar graphs (bar charts in Microsoft lingo) are especially suitable for wordy categories, avoiding the need for vertical text labels or abbreviations. In this work, we will add the words 'horizontal' and 'vertical' to 'bar graph' where required to avoid confusion. The terms 'graph' and 'chart' appear to be used interchangeably.

Related to the vertical bar graphs are histograms, which display continuous variables with columns touching each other: more about these in section 3.3.

3.2.2 Vertical bar graph

A vertical bar graph displays one numeric variable, on the *y*-axis, against a categorical variable on the *x*-axis (site, species name, etc.). Such bars have a very low information density, and they implicitly present information as the length of the bar. This puts them low on the scale of decoding accuracy, and requires that you include zero on the *y*-axis. For bigger values, this can compromise resolution, and where the *y*-axis has a log scale, this is impossible—which poses a conundrum for good graph design. The information density is slightly higher if you add error bars (Fig. 1), use stacked bars (Fig. 10) or multiple bars (Fig. 11). When full dates do not fit on the *x*-axis, it may be best to abbreviate to the sequence of first letter of the months (i.e. JFMAMJJASOND) or just the day number, and show month and year in the caption.

Figure 11. Multiple vertical bars are not a very good way of presenting data accurately. It is difficult to gain a view of the distribution for each location because the bars are intermingled. Also, in this case, the *x*-axis should show the subdivisions of the length used for the counts. The presentation of an apparently continuous length variable creates distortion and does not clearly reveal that lengths were measured in intervals of 0.5 mm. Fig. 12A shows a more effective example (where the *x*-axis represents categories instead of a continuous variable), but even so better alternatives are available (Figs 12B & 13B).



Original caption: Length frequency distribution of larval galaxiids collected from four sites in Totara Creek on 5 December 1998.

3.2.3 Stacked and multiple bar graphs

Stacked bars (several values one above the other making a single column per category on the *x*-axis) are the general form of the column graph we recommend to replace linked pies (see Fig. 10). Multiple-bar graphs (several variables plotted as adjacent columns next to each category on the *x*-axis) can become hard to read (see Figs 11–13). The bars pile up together and discrimination is difficult, especially in black-and-white representation, where you must use stripes or stipples and a key to identify the various bars. Colour can make a multiple-bar graph easier to discriminate, but when the graph is photocopied in black and white it will be hard or impossible to interpret.

How to improve such graphs? If the *x*-axis is actually a continuous variable (e.g. length in mm or time in years) rather than a categorical one, then draw a standard *x*–*y* graph instead (see section 3.5). The use of different symbols and / or lines allows more than one series to be displayed readily. If you have a complex multiple-bar graph, data may be better represented as a table, where readers can run their eye down each column easily, or as multiple panels (multipanels) in the graph, often with identical axes, depending on the context (see section 3.7 and Fig. 12).

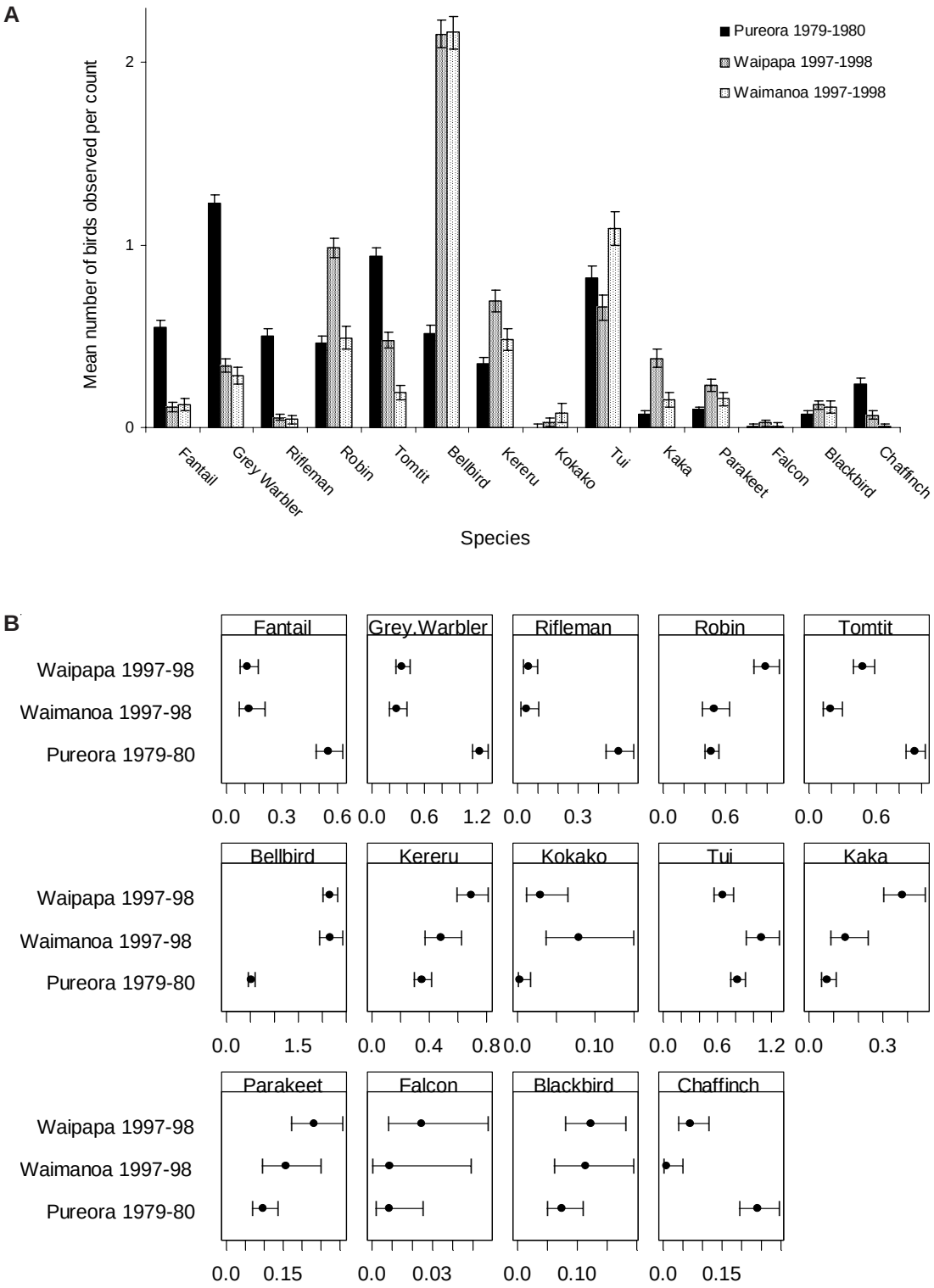


Figure 12. Comparison of a bar chart (A) with a dot chart finally designed for publication (B). The values represent average counts of birds in five-minute observation periods, with 95% confidence intervals. The use of varying scales for different panels is noted in the caption in the original.

Original caption to B: Winter (May and June) mean bird conspicuousness in two studies in Pureora Forest Park, with confidence intervals based on the assumption that the counts have a Poisson distribution. Note that the panels for different birds have varying scales.

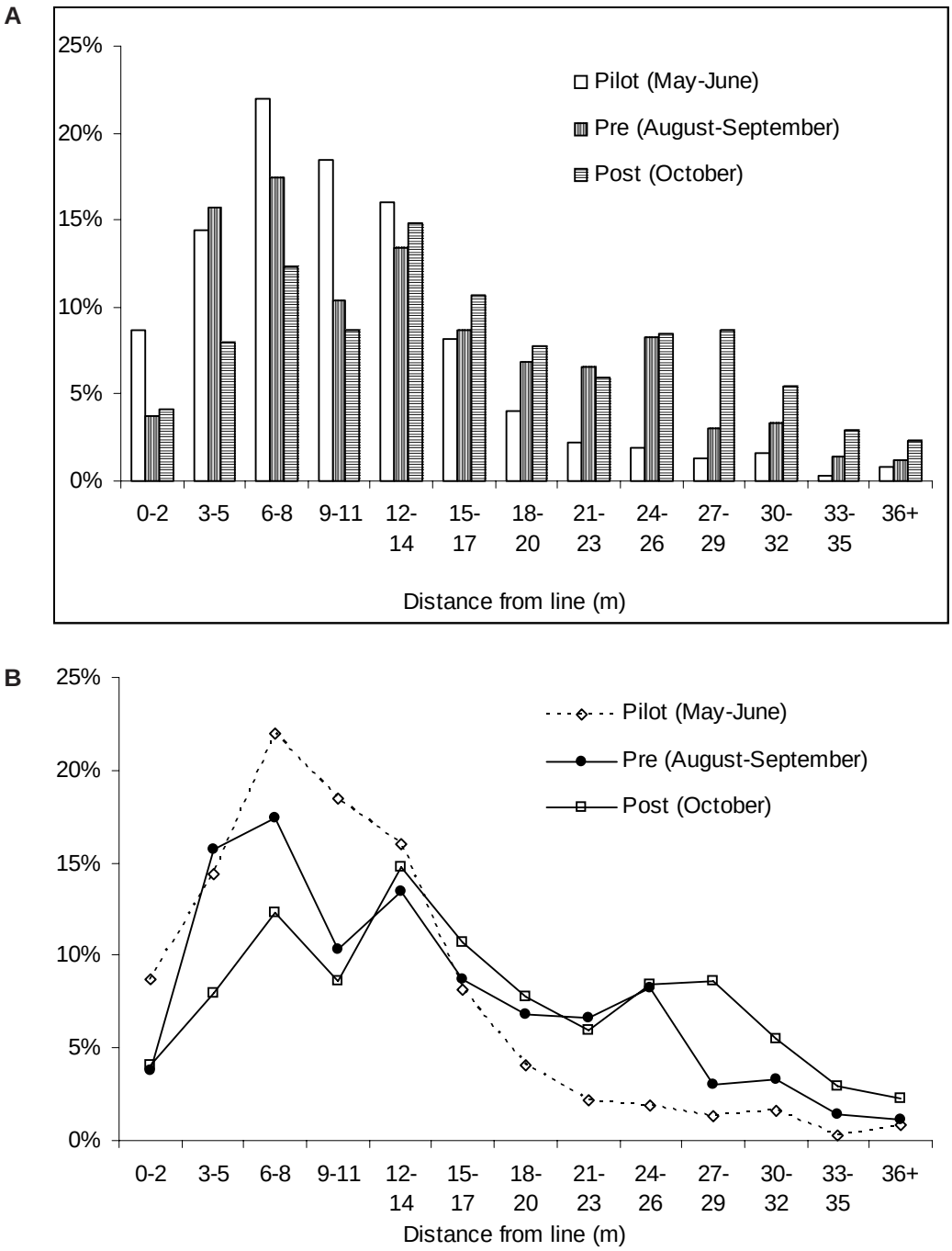


Figure 13. A bar chart (histogram) (A) and a relative frequency polygon (B—as published), based on the same data. Comparing several groups in one histogram destroys the continuity of the *x*-axis. The frequency polygon uses lines joining points to represent a distribution: it can show a modest number of related distributions clearly on one chart.

Original caption to B: Percentage of distance sampling observations in 3-metre distance classes, for three phases of the study: pilot (May-June 2001, *n* = 368), pre-treatment (August-September 2001, *n* = 439), and post-treatment (October 2001, *n* = 425).

3.2.4 Horizontal bar graph

When category labels are too long to reproduce in horizontal type on the *x*-axis of a vertical bar graph, it is better to use a horizontal chart rather than print oblique or vertical type. This graph shape is particularly well suited to categorical data with long names: results of questionnaires, etc. Figure 8B is an example.

3.2.5 Dot chart

A dot chart (dot plot) is a special type of horizontal bar chart, developed by Cleveland (Cleveland 1993). It uses a minimum of ink to optimum effect (which, according to Tufte (1983), indicates good design). The other strength of this design is that by using a dot it is clearly indicating the value by the position of the dot relative to the *y*-axis scale, rather than by the length of the bar, as in a normal bar chart. It may be better in technical works (Fig. 12B), although some authors and readers appear to have difficulty in letting go of the more familiar bars (Fig. 12A).

Dot charts feature:

- Horizontal arrangement (with plenty of room for long labels); usually categorical data.
- A dot marking the data point, not a bar.
- Optionally, light dots on left only (if zero baseline) or, more usually, all the way across to link the dot to its label.

3.3 HISTOGRAM AND FREQUENCY POLYGON

A histogram always has two numeric axes, but the *x*-axis is always a continuous variable, divided into an arbitrary number of categories—usually to show distributions. When drawn for a single variable, the bars of continuous variables by convention touch each other (see Appendix 1); bars for true categorical variables are better presented with spaces between them. Histograms have rather few, fairly specialised uses. They are fine for showing distributions within a large dataset. However, comparing several groups destroys the continuity of the *x*-axis (Fig. 13A), and there is some loss of information compared with showing the scatter, or a cumulative frequency curve, both of which can show the entire dataset.

A frequency polygon (Fig. 13B) is like a histogram, but uses lines joining points to represent a distribution, instead of bars. Its big advantage is that it can show a modest number of related distributions clearly on one chart, using different symbols and / or lines. It has also been shown to be technically superior (Scott 1992).

Histograms and frequency polygons can be based on numbers, or on relative frequencies (relative frequency = the frequency at each point in each category divided by the total for that category), depending on which is more useful.

According to Cleveland (1994), box-and-whisker plots and quantile plots are often better alternatives for assessing distributions. We discuss box-and-whisker plots in section 3.4, but readers are referred to Cleveland (1994: 136) for more on quantile plots.

3.4 BOX-AND-WHISKER PLOT (BOX PLOT)

The box-and-whisker plot (or box plot) is an excellent exploratory graph for summarising the distribution of one continuous variable, possibly broken up into several categories. It is very useful for picking up key aspects of the distribution of samples of modest to very large size.

The most common text-based summary of data involves either just the mean, or the mean and standard deviation, i.e. only a one- or two-number summary. While the mean and standard deviation are very good at summarising data with a normal distribution, most real datasets are not so well behaved.

By contrast, a simple box plot is based around a five-number summary of the data: these are derived by taking all the data and putting the values in order. The derived values are:

- The median (midpoint value in the data, i.e. 50th percentile)
- The upper and lower quartiles (the points midway between the median and the extreme values, i.e. 25th and 75th percentiles)
- The minimum and the maximum

Box plots may also add the positions of potential outliers.

The median and quartiles are used because they are robust: they will not be affected much, if at all, by some odd values in the data. In contrast, the mean, and especially the standard deviation, are very sensitive to the addition of a single extreme value to the data. A box plot example is shown in Fig. 14.

A box plot will show very clearly where the odd extreme values are, and also skewness—where values are systematically further from the middle in one direction than in the opposite direction. The box plot in Box 1, section 2.3.3, illustrates the decoding accuracy of various kinds of data presentation; it shows very clearly the winner: 'position on a common scale' was rated the best for decoding the value of numbers. Not only was the middle value highest, but it was also recorded as the best at every session, and the average ranking varied

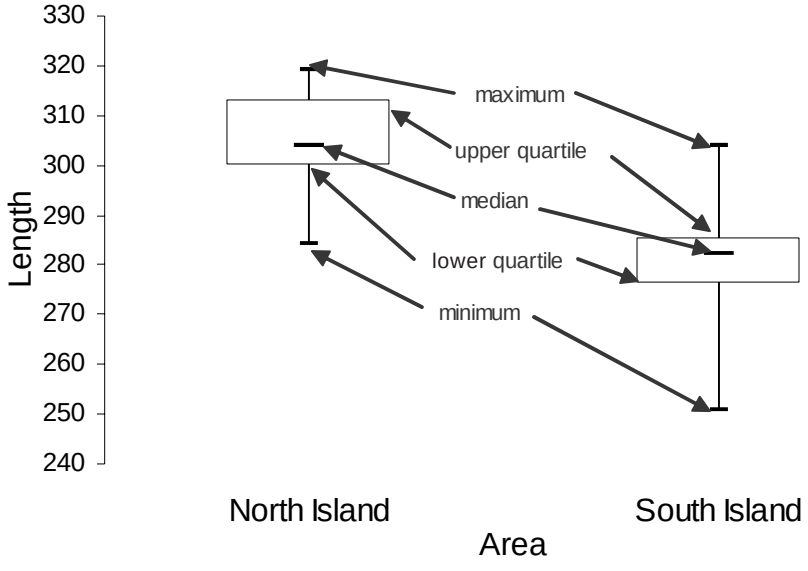


Figure 14. Example of a vertical box plot showing the distribution of Hector's dolphin data for North Island and South Island populations and the various box plot parts.

Original caption: Distributions of five measurements ... for the North and South Island populations, demonstrating the clear morphological separation between them...D, condylobasal length; ... Scale axes ... are in millimetres.

less than any of the alternatives. In contrast, some of the other methods were much more spread, and 'position on identical non-aligned scales' appeared to be skewed—with a median of 4, but many more values well above 4, and few much below.

Unfortunately, Excel does not provide facilities for creating a box plot as a standard type of graph, but there is a file developed at DOC that allows creation of simple box plots for up to 20 groups. The file can be requested from the third author (IW, DOC; email: iwestbrooke@doc.govt.nz).

Small datasets (say fewer than about 10 data points in each category), and some larger ones, may be better plotted as the individual values directly. An example is shown in Fig. 15.

Box plots do not work as well with integer data (e.g. counts) as they do with continuous variables (e.g. length); for integer data, for example, the 25th and 50th percentiles may both be on the same value, which messes up the box plot. This is illustrated in the evaluation data of the 2003 Graphs workshops, which applies the DOC spreadsheet for table format (Fig. 16A) and box plot (Fig. 16B). A simple, Excel-generated dot plot is provided for comparison (Fig. 16C).

More sophisticated box plots are available in statistical packages such as SPSS and S-PLUS. The key difference is that they go beyond the simple box plot by establishing 'fences' (usually 1.5 times the interquartile range—the range between the upper and lower quartiles) beyond the upper and lower quartiles. The whisker at each end stops at the extreme values of the data if within the fence, as in the simple box plot. However, if there are extreme values (possible outliers) outside these fences they are shown individually, with the whisker stopping at the closest data value within the fence. These more complex box plots are even more useful for exploratory data analysis. Because different implementations of box plots display different parts of the distribution with their lines and whiskers, it is always helpful to define these in the caption, e.g. 'The box plot indicates the median, interquartile range, maximum and minimum'.

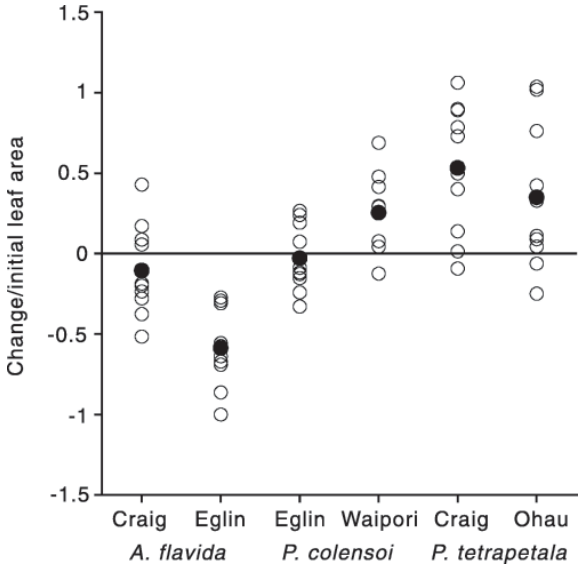


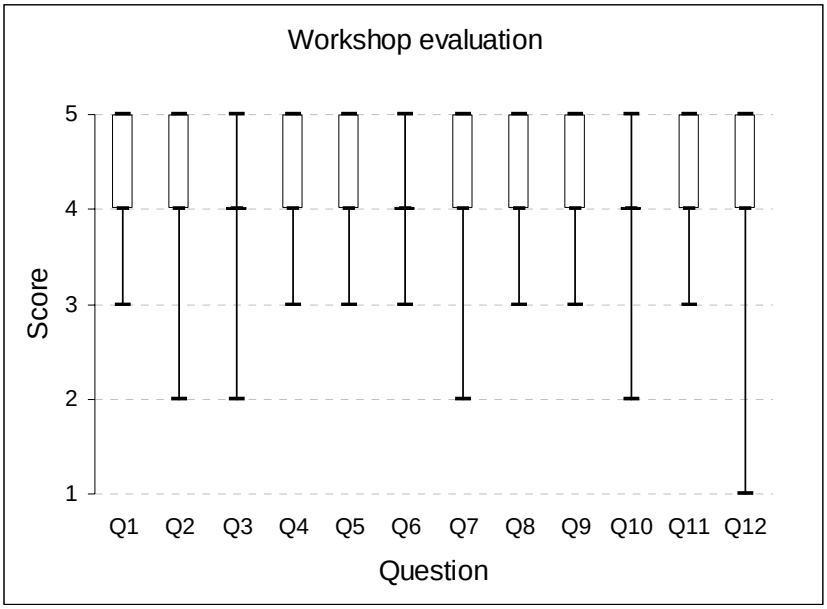
Figure 15. A dot plot showing the data for six categories that are tested statistically elsewhere with one-way ANOVA. Frequently this might be shown as a bar graph with six bars representing the means, and perhaps error bars. However, such bar graphs have a low information density, representing only 12 numbers (6 means and 6 SEMs / CIs). A somewhat more informative version uses boxplots (see Fig. 14). In the plot shown, the same space is used to display the number of data points and their full distribution, along with the means for each group.

Original caption: Overall annual leaf flux (net change in leaf area divided by the initial leaf area) between February 1997 and February 1998 on mapped branches in six populations of New Zealand mistletoes. (○), values for each plant; (●), population means. For full site names see Fig. 1.

A

Statistic	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
maximum:	5	5	5	5	5	5	5	5	5	5	5	5
upper quartile:	5	5	4	5	5	4	5	5	5	4	5	5
median:	4	4	4	4	4	4	4	4	4	4	4	4
lower quartile:	4	4	4	4	4	4	4	4	4	4	4	4
minimum:	3	2	2	3	3	3	2	3	3	2	3	1
number of obs:	143	143	134	142	143	143	115	143	143	142	141	143
mean	4.3	4.3	4.0	4.2	4.4	4.1	4.2	4.4	4.4	4.1	4.3	4.2
standard deviation	0.6	0.6	0.6	0.5	0.5	0.6	0.6	0.5	0.5	0.6	0.5	0.7

B



C

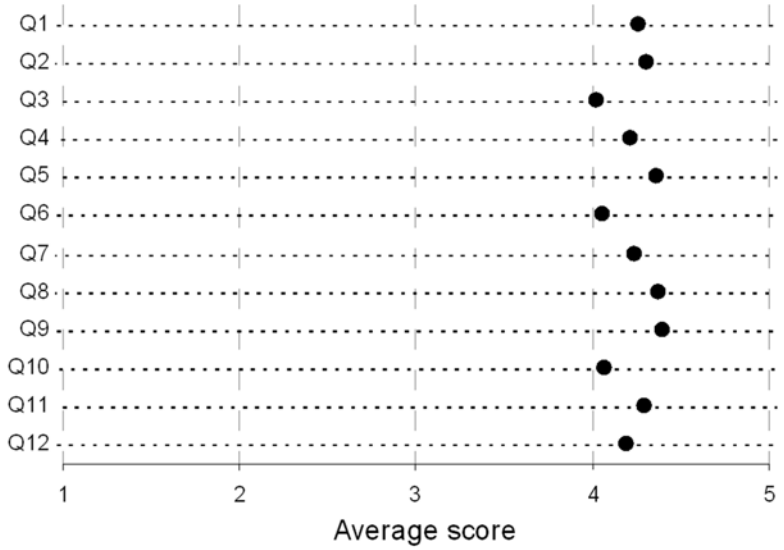


Figure 16. Participants' responses to graph workshop evaluation questionnaires. Scores: 1 Disagree strongly; 2 Disagree; 3 Neutral; 4 Agree; 5 Agree strongly. Figure 16A shows the results in table format, B shows a simple box plot, and C a dotplot of the average. The box plot does not work very well here with only a few response categories.

3.5 *x*–*y* (BIVARIATE, LINE OR SCATTER) PLOT

Bivariate graphs are the bread and butter of scientific graphing. They make excellent illustrations, and you really cannot go wrong using more of them. *x*–*y* graphs display two numeric variables. We can recognise two slightly artificial subtypes: time series, where the *x*-axis is time (more than 75% of graphs in newspapers were like this in the late 1970s (Tufte 1983)), and relational, where neither axis is time (42% of graphs in the journal 'Science' 1978–1980 were of this form (Cleveland 1984)).

There are various types: line graphs (lines only), line-plus-symbol (Fig. 17), or scatterplots (symbols only, Fig. 18), which can apply different symbols for several different variables.

You can include error bars on points; this can be done one way (vertically, as shown in Fig. 17, or horizontally), or both ways (vertically and horizontally), as appropriate.

In a scatterplot, extra text labels to the data points may increase clutter and should generally be avoided. Sometimes you can use a text label as the data point (e.g. using capital letters A, B, C, etc. to mark locations and also identify sites—which gives labelling without increasing clutter: Fig. 18). Avoid letters overlapping.

You can plot a scatter with a fitted line, e.g. a regression line as in Fig. 2. Never show the regression only! It takes no extra space to put the data on and the scatter gives a lot of information about the data. Indeed, the data may well show that even though the r^2 value is close to 1, the interpretation may be suspect (Fig. 19).

A step function graph is a variant of the *x*–*y* graph, where the *y* value is constant over intervals then changes suddenly to a new value (e.g. the price of the daily newspaper over time), so the graph looks like series of irregular (square-edged)

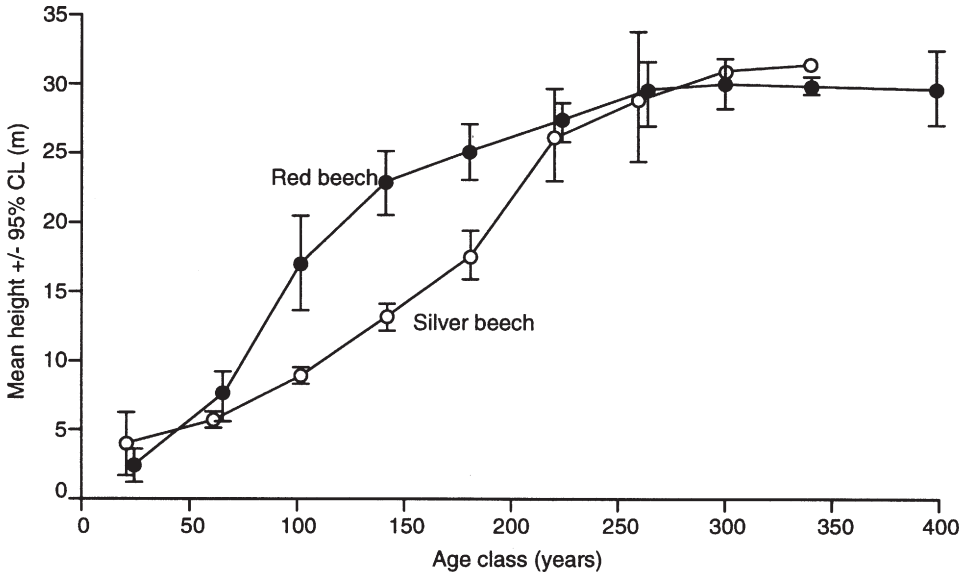
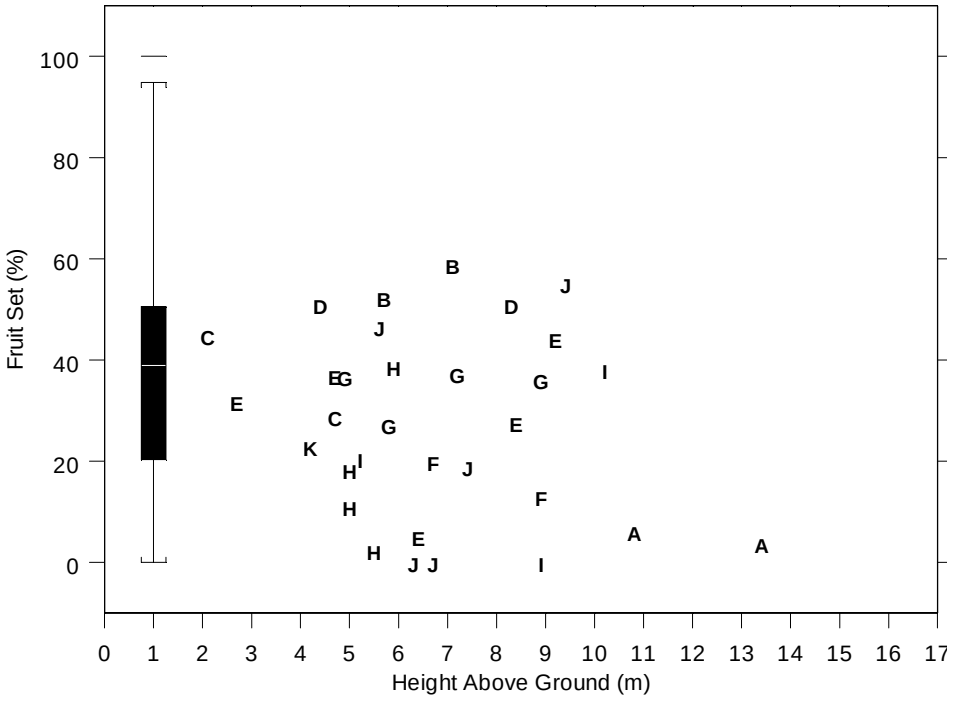


Figure 17. Good example of a clear *x*–*y* plot with suitable symbols, categories, and error bars with explanation (95% confidence interval).

Original caption: Average height growth of red and silver beech trees of different age classes in a stand in the Maruia Valley (After Stewart & Rose 1990).

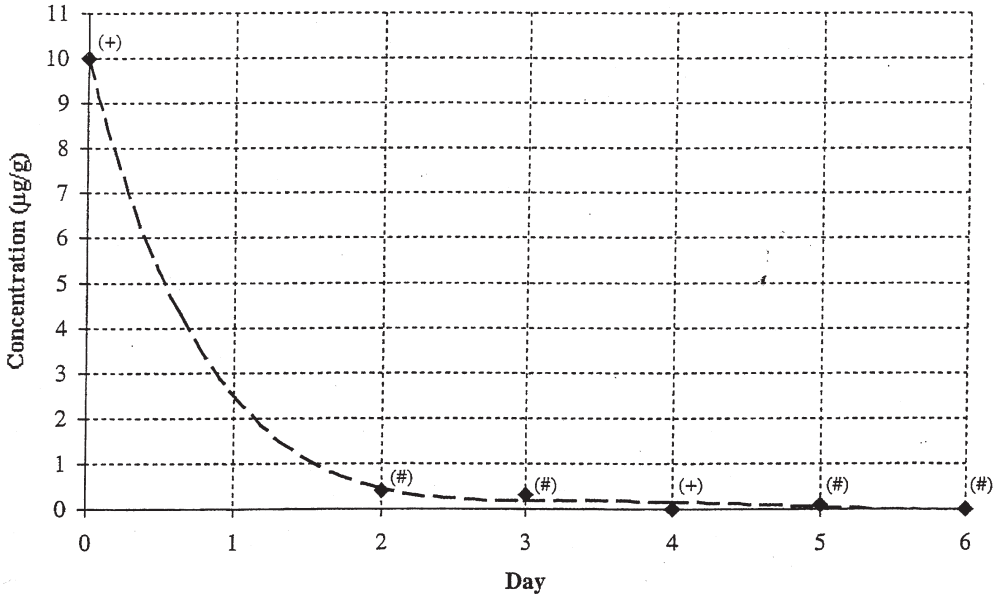
Figure 18. An example of text labels used as data points; the graph also uses a box plot on the left. The text labels serve to locate each point (mistletoe plant) both for height and fruit set rate; they also allow the reader to identify mistletoe plants that share a single host tree. It is not easy to decode the latter, but in this case the authors thought it not especially important to do so, as the overall message is that there is no effect either of height or of individual host tree. If it was important to easily link mistletoes on a single host, the points for mistletoes on the same tree could be joined by lines, but this would make it harder to see the overall picture (here of no relationship between height and fruit set).



Original caption: Fruit set in *P. tetrapetala* at Craigieburn Forest Park in the 1997/1998 flowering season. The box plot shows the range of fruit set values obtained from tagged plants used for our normal pollination treatments (all located within 4 m of the ground) while letters mark the 32 plants located up vertical transects accessed by climbing ropes. Shared letters indicate plants that are located on the same vertical transect.

Figure 19. Not only does the curve interpolate and extrapolate well beyond acceptable boundaries, e.g. curve between the first and second datapoints), it also incorrectly combines data from different sources according to the accompanying text. At best, the points could have been connected by two separate lines: one from (0,10) to (4,0) and another to connect the remainder, just above the *x*-axis.

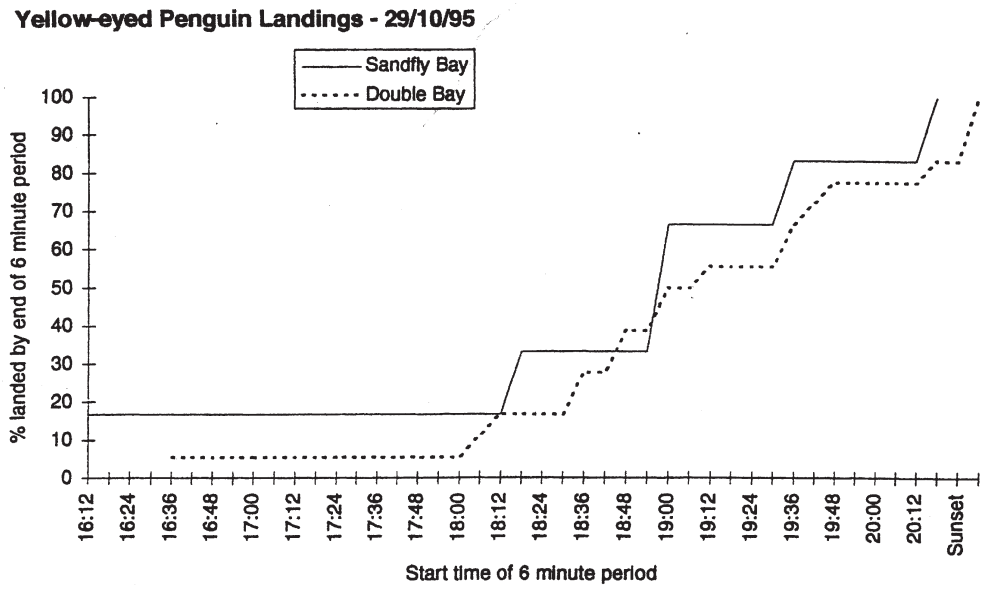
Original caption: Estimated excretion curve for brodifacoum in Orthopteran species. Based on data from this study (#) and Booth, Eason & Spurr 2001 (+).



steps. The step function graph is often used for cumulative proportion below a certain value in a sample, or for representing the estimated proportion surviving over time (Fig. 20).

You can plot several categories or classes on the same *x*–*y* graph, using symbols to separate them, as in Figs 17 and 18. The main concern is symbol / line separation; if this becomes a problem, you may have to present multiple small graphs rather than one large one (see section 3.7).

A



B

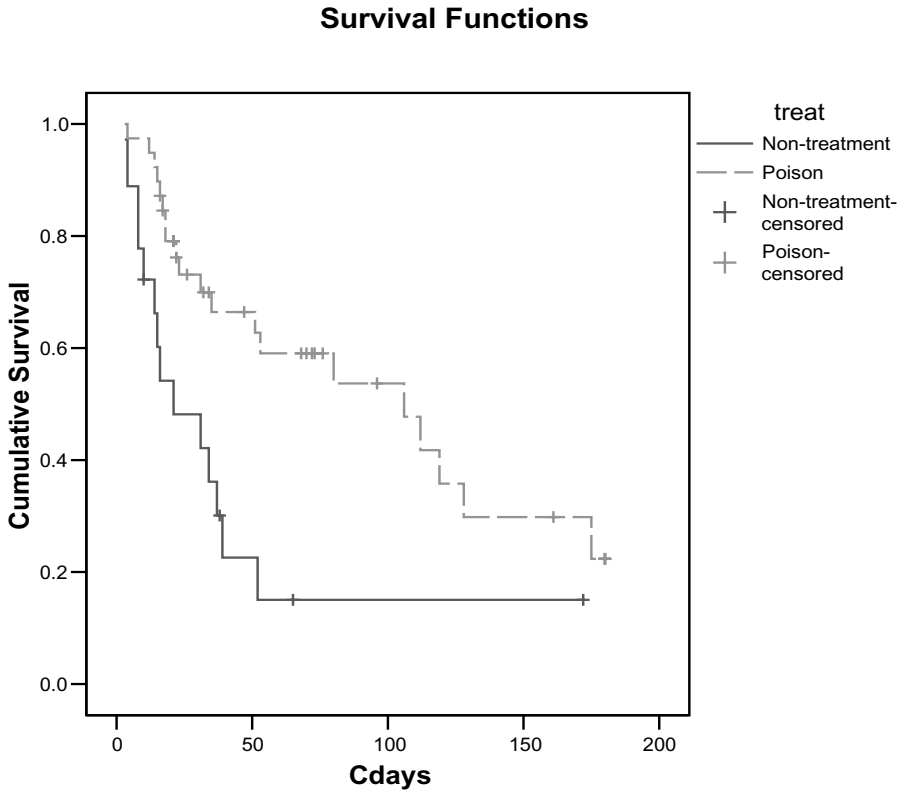


Figure 20. Step function graphs. A, with angled steps. Conventionally the steps drop vertically at each point to the level of the next point, as shown in B (survival rates estimated for two groups of kiwi chicks with and without pest treatment).

Original caption for A: Graph[s] showing comparison of daily yellow-eyed penguin landing times between Sandfly Bay and Double Bay.

Continue to next file: docts32b.pdf