**Biostatistics: A QUICK GUIDE TO THE USE AND CHOICE OF
GRAPHS AND CHARTS**

1. Introduction, and choosing a graph or chart

Graphs and charts provide a powerful way of summarising data and presenting them in
a way that most people find easy to read. They are an important alternative to tables,
especially where a dataset is large, but remember that a graphical presentation of data
is unlikely to allow your reader to access the raw data if they wish, whilst this is possible
with a table. Graphs and charts may alo be useful in assessing data prior to udretaking
data analysis or to fitting a numerical model.

The choice of an appropriate graph or chart depends on what information you are trying
to convey. The three choices at the end of this introduction link to the next three
sections. Finally there is a brief section on plotting graphs and charts using
spreadsheets. Remember that each of the NuMBerS statistical toolkits includes
guidance on the appropriate graphical presentation.

What is your graph going to do?

  Summarise the data - go to Section 2

  Illustrate the average and variability for a single set of observations, or the
  differences between more than one set of observations - go to Section 3

  Illustrate relationships between variables - go to Section 4

2. Graphs and charts that summarise data

There are various ways of presenting data in graphical form, and these typically
dominate the options that are available in spreadsheets (see Section 5). The choice
depends of the type of data (nominal, ordinal, scale-discrete or scale-continuous).

***2.1 Summarising data that are nominal or can be reduced to categories in some
way***

The number of observations within each category can be represented either as a pie-
chart or a bar-chart. In a pie chart, each category is represented as a segment of a disc
(the 'slice' of the 'pie'), whose angle is proportional to the proportion of the overall
observations accounted for by the category (Fig. 1).
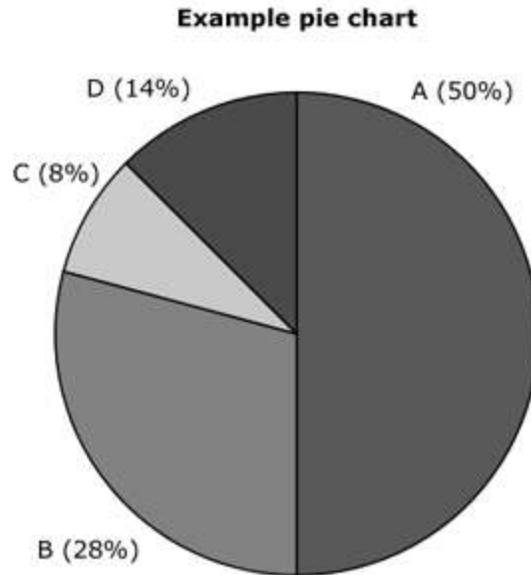
Example pie chart

**Figure 1**. *An example pie chart, where each segment is labelled*

In a bar chart, the magnitude of each category is represented by the length of a bar (usually vertical but occasionally you will meet horizontal bar charts) (Fig. 2). The value for a category may be the number of observations, or the proportion of the total observations in the category.
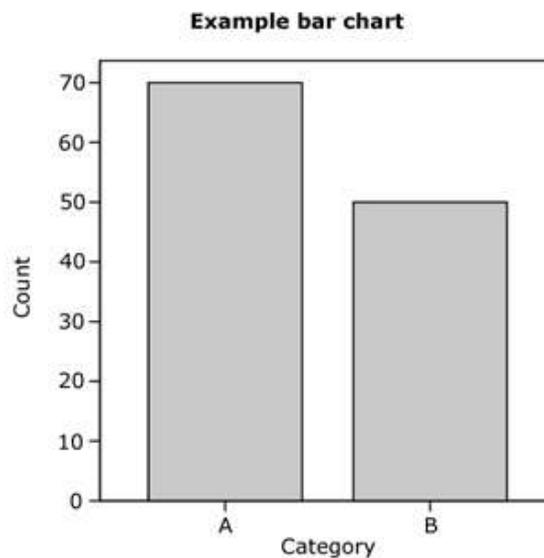
Example bar chart

Figure 2. *An example bar chart, showing the magnitude for each category as a vertical bar*

A pie chart is not an appropriate choice where there is a large number of categories, say more than six and certainly not more than ten. A bar chart may work better than a pie chart where there are more categories, although the advantages may be minimal and it still becomes difficult to label different categories.

Where there is more than one set of observations, each set can be represented by a separate pie- or bar- chart (Fig. 3). In the case of the pie chart, which can only show proportions as the segment angles, different numbers of observations can be represented by making the overall area of the disc proportional to the number of observations in each dataset.
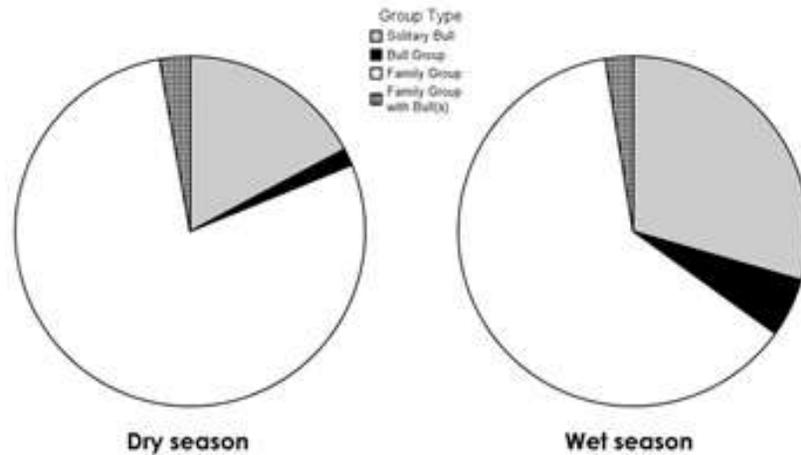
Figure 3. *An example of using two pie charts to compare sets of observations, in this case differences in social groups of elephants in the wet- and dry- seasons*

2.2 Summarising data that are ordinal or scale

Where ordinal or discrete- or continuous- scale data are to be displayed, the most appropriate graphical presentation is a histogram (Fig. 4) where there is a relatively small data range and data are ordinal or discrete, or a continuous distribution curve in the case of continuous scale data with a wide range. A histogram works in the same way as a bar chart, but there is no space between the bars. It is the area of the histogram bar that signifies the value, not its height, and histogram bars may encompass different data ranges.
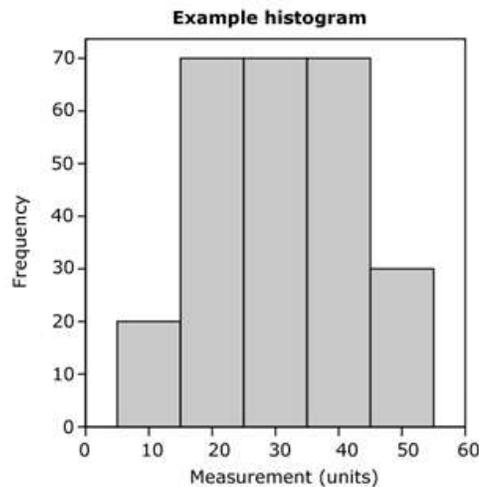

Figure 4. *An example of a histogram used to show a frequency distribution*

3. Illustrating the average and variability within a dataset, or the differences between datasets

There are two common graphical presentations that summarise datasets in terms of an average (for instance the mean) and a measure of variability (for instance the standard deviation). The various measures available are summarised in the two helpsheets on descriptive statistics.

3.1 Average and variability for a single parametric dataset

Parametric data are scale data (not ordinal or nominal) that are at least approximately normally distributed. For such data, the mean is the most appropriate measure of the population average, and variability is typically represented by the 95% confidence interval (CI). These items of information are shown graphically as an error plot (Fig. 5), where the mean is shown as a symbol against an appropriate vertical scale, and the 95% CI is indicated by a vertical line (or 'error bar') passing through the symbol for the mean. In some instances, you may see error plots where the line indicating the value of the 95% CI is terminated with horizontal bars. The caption to the illustration must indicate what measure of variability is used to calculate the length of the error bar.
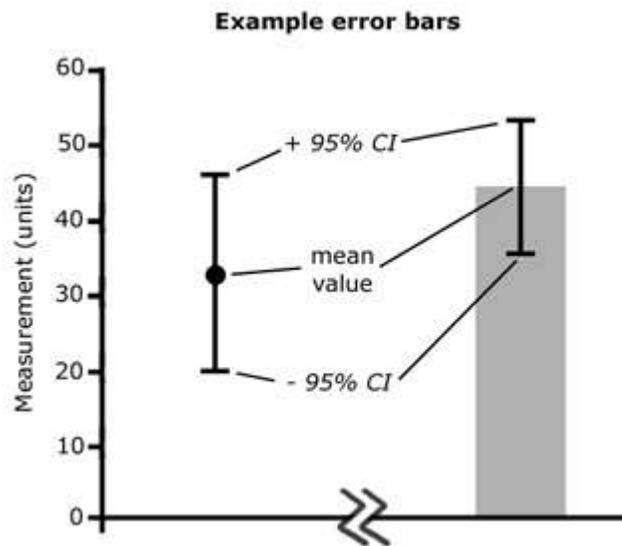

Figure 5. *Examples of two forms of error bars, both indicating the 95% confidence interval about the mean*

3.2 Average and variability for a single non-parametric dataset

Data that are scale or ordinal, but which do not satisfy the criteria for parametric data, will not be described well by the mean as an average, and a symmetrical measure of variability such as confidence interval or standard deviation. A box plot (Fig. 6) is more complicated that an error plot, and displays the median value against an appropriate vertical scale, and the interquartile range as a vertical bar (the 'box'). Typically, the overall data range is also represented by a vertical line (sometimes referred to as a 'box

and whisker' plot), although extreme values and outliers may be separated out from the overall data range.

Example box plot
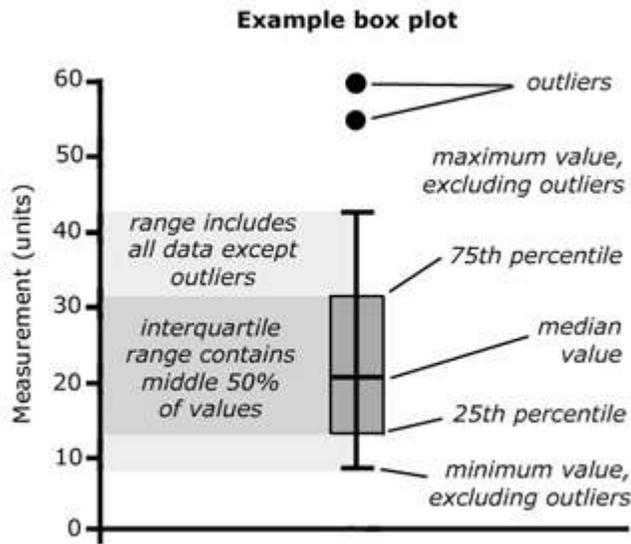
Figure 6. *The main features of a box plot, including outliers or extreme values excluded from the range*

3.3 Comparing average and variability for several datasets

Both error plots and box plots can be used to compare different samples or populations. A chart can include several error plots or box plots, and these allow the user to make an instant comparison between the averages and variabilities of different datasets. The degree of overlap between variabilties is an important initial indicator of the likelihood that differences in means or medians are meaningful (Fig. 7), an assessment that can then be tested more rigorously using the appropriate test.

Overlap in error bars

(a)   (b)   (c)
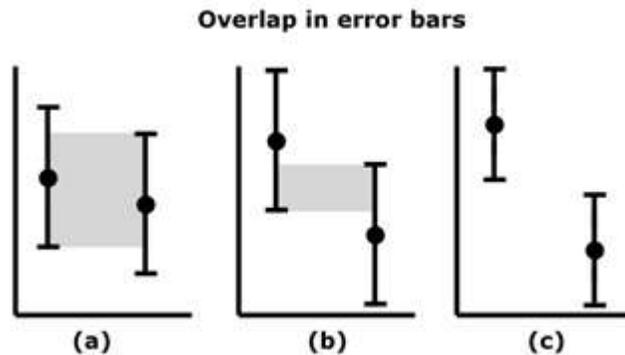
Figure 7. *Different levels of overlap in error bars: (a) large amount of overlap, that includes the mean values, (b) small amount of overlap, that does not include the mean values, (c) no overlap*

If making comparison between several parametric datasets, it is important that the variability is similar in all. If the most variable has a variance measure that is more than ten times the variance of the least variable, statistical comparison of the means is invalid, and a box plot might be a more appropriate way to display the data (Fig. 8).
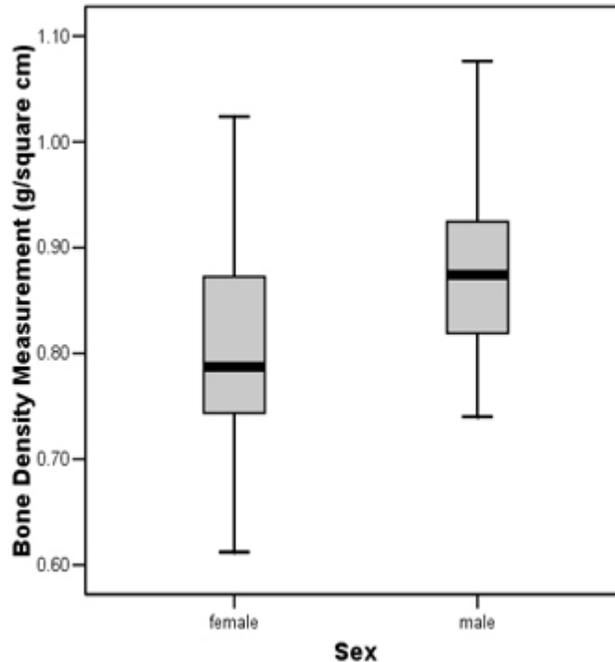
Figure 8. *An example of the use of box plots to compare two samples, in this case illustrating differences in bone density between men and women*

4. Illustrating relationships between variables

Graphical representation is a powerful tool to explore relationships between variables. In the commonest case, two variables are measured for several samples, and the measurements can be plotted on a graph using Cartesian coordinates (horizontal and vertical scales, or an X-Y plot in a spreadsheet).

4.1 Correlation

The simplest way to investigate the relationship between two variables is to use the appropriate test for correlation. The relationship is illustrated graphically using a scatter plot (typically called an 'X-Y plot' in spreadsheets). A high degree of correlation will be shown where data points tend to fall along a straight line, where high and low values of the two variables tend to coincide (positive correlation) or where high values of one variable correspond to low values of the other, and vice versa (negative correlation). A 'shotgun' scatter, or an alignment of points that is close to horizontal or vertical, indicates very low correlation (Fig. 9 below).

Example scatter plots
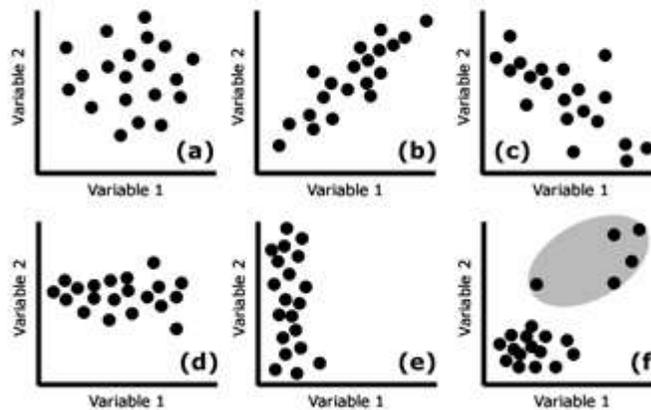
Figure 9. *Illustration of scatter plots with various properties: (a) 'shotgun' scatter, with low correlation, (b) strong positive correlation, (c) strong negative correlation, (d) and (e) low correlation, with very little change in one variable compared with the other, (f) this scatter would generate a spurious high correlation because of the effect of the five points enclosed by the shaded area*

4.2 Linear regression

If there is likely to be a causative relationship between two variables, such that the value of one variable determines most or all of the value of the second, a linear regression can be carried out. Again, a scatter plot is used to explore the relationship. The independent variable (commonly denoted x) is shown on the horizontal axis of the plot, whilst the dependent variable (denoted y) is shown on the vertical axis. If the calculated regression is significant, the fitted line should be added to the scatter plot (Fig. 10).

$$y = 8.56 x + 1.73$$
$$F_{1,19} = 12.63$$
$$P < 0.05$$
$$R^2 = 0.872$$
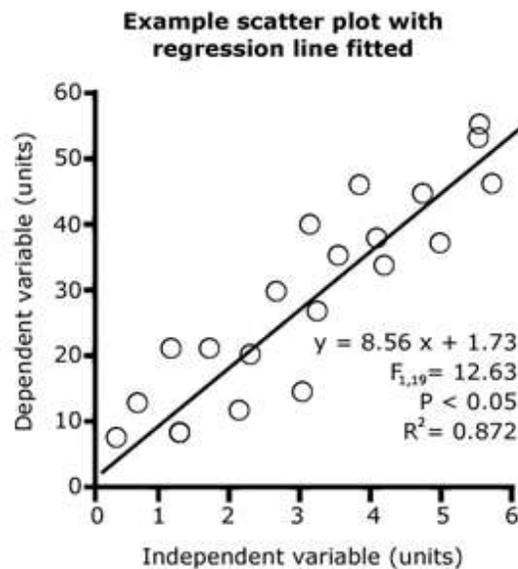
Figure 10. *Scatter plot with fitted regression line*

4.3 More than two variables

Obviously, it is only possible to represent the relationship between two variables in two-dimensional space. It is possible to group scatter plots into a matrix to represent all possible pairs of bivariate relationships within a set of observations of several variables. Alternatively, there are multivariate analyses that represent the overall variability in a dataset with many variables as a series of new summary variables, and the results can then be displayed in two dimensions.

In the case of three variables, an isometric three-dimensional plot can be used. Data points are often shown attached to the 'floor' of the plot (so-called 'ball on stick' plots) to provide a better spatial reference for the two dimensional representation of three-dimensional space. The three-dimensional equivalent of a regression line is a surface.

5. Spreadsheets

Spreadsheet packages such as Microsoft Excel are quick and convenient ways to produce most of the graphs and charts described here. Because spreadsheets were designed originally for business applications, the graph and chart options are biased towards certain types of display. In some cases, it may take some ingenuity to make the spreadsheet produce exactly the type of plot that you want it to.