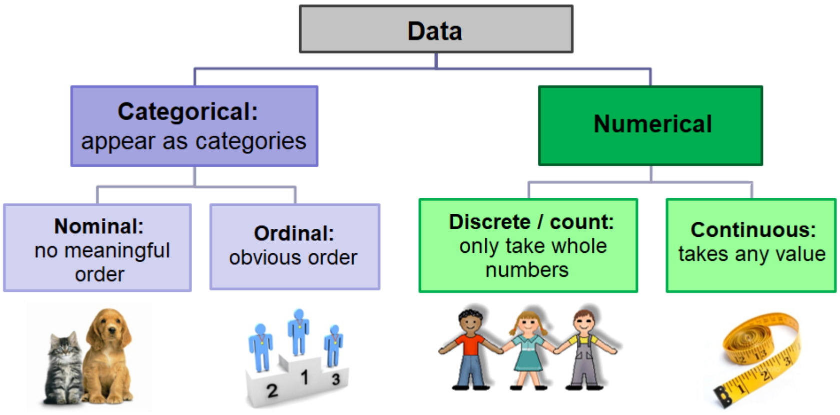
Types of data


Data
Categorical: appear as categories
Numerical
Nominal: no meaningful order
Ordinal: obvious order
Discrete / count: only take whole numbers
Continuous: takes any value
2
1
3

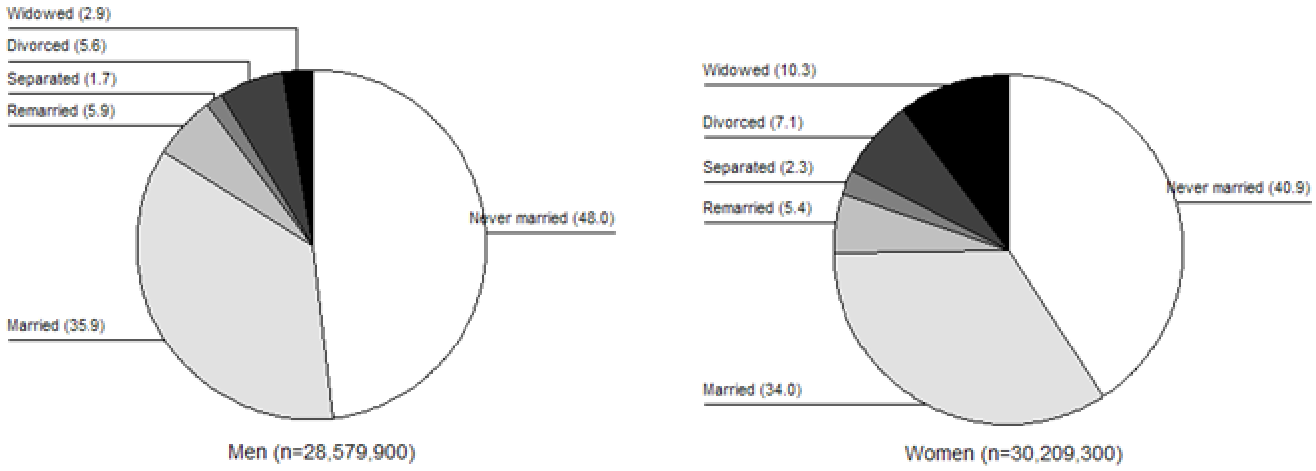
In order to appropriately display data, it is important to understand the different types of data there are as this will determine the best method of displaying them. Briefly, data are either **categorical** or **quantitative**. Data are described as **categorical** when they can be categorised into distinct groups, such as ethnic group or disease severity. Categorical data can be divided into either **nominal** or **ordinal**. Nominal data have no natural ordering and examples include eye colour, marital status and area of residence. **Binary** data is a special subcategory of nominal data, where there are only two possible values, for example (male/female, yes/no, treated/not treated). Ordinal data occurs when there can be said to be a natural ordering of the data values, such as better/same/worse, grades of breast cancer, social class or quality scores.

Quantitative data can be either discrete or continuous. **Discrete** data are also known as count data and occur when the data can only take whole numbers, such as the number of visits to a GP in a year or the number of children in a family. **Continuous** data are data that can measured and they can take any value on the scale on which they are measured; they are limited only by the scale of measurement and examples include height, weight, blood pressure, area or volume.

Basic charts for categorical data

Categorical data may be displayed using either a **pie chart** or a **bar chart**. Figure 1 shows a pie chart of the distribution of marital status by sex for UK adults at the 2001 census. Each segment of the pie chart represents the proportion of the UK population who are in that category. It is clear from this figure that differences between the sexes exist with respect to marital status; nearly half of all men have never

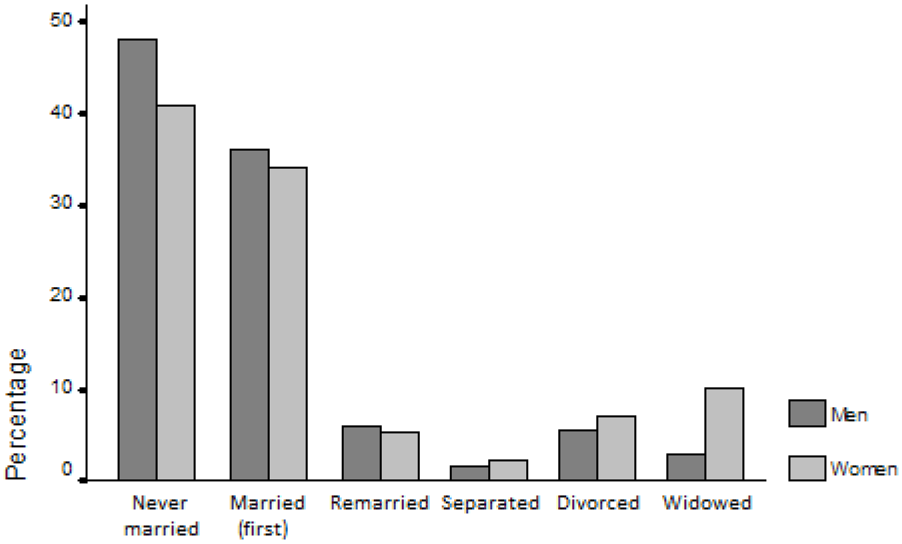
Figure 1: Pie chart of marital status by gender for UK population, 2001 census



married, whilst this proportion was smaller for women. Interestingly the proportion of women who were widowed was about three times that for men.

Figure 2 displays the same data in a bar chart. The different marital status categories are displayed along the horizontal axis whilst on the vertical axis is percentage. Each bar represents the percentage of the total population in that category. For example, examining Figure 2, it can be seen that the percentage of men who are married is about 48%, whilst the percentage of women is closer to 40%. Generally pie charts are to be avoided as they can be difficult to interpret particularly when the number of categories becomes greater than 5. In addition, unless the percentages in the individual categories are displayed (as here) it can be much more difficult to estimate them from a pie chart than from a bar chart. The relative proportions falling in the different categories is much clearer in Figure 2 than in Figure 1. For both chart types it is important to include the number of observations on which it is based, particularly when comparing more than one chart. And finally, neither of these charts should be displayed as 3-D as these are especially difficult to read and interpret.

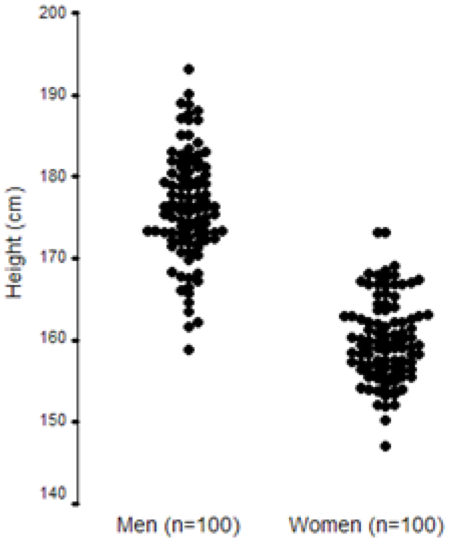
Figure 2: Clustered barchart of marital status by gender for UK population, 2001 census



Basic charts for quantitative data

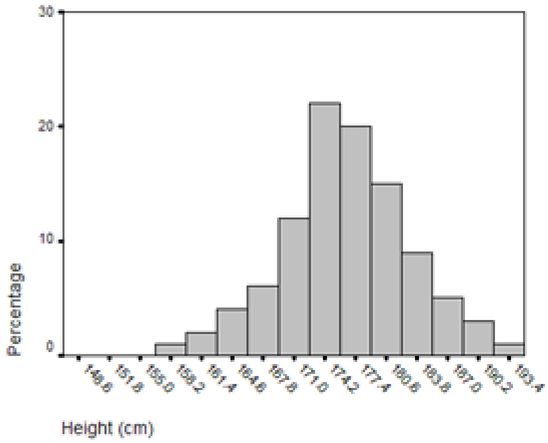
There are several charts that can be used for quantitative data. **Dot plots** are one of the simplest ways of displaying all the data. Figure 3 shows dot plots of the heights for a random sample of 100 couples. Each dot represents the value for an individual and is plotted along a vertical axis, which in this case, represents height in metres. Data for several groups can be plotted alongside each other for comparison; for example, data for the 100 randomly sampled couples are plotted separately by sex in Figure 3 and the differences in height between men and women can be clearly seen.

Figure 3: Dots plots of height for 100 men and 100 women



A common method for displaying continuous data is a **histogram**. In order to construct a histogram the data range is divided into several non-overlapping equally sized categories and the number of observations falling into each category counted. The categories are then displayed on the horizontal axis and the frequencies displayed on the vertical axis, as in Figure 4. Occasionally the percentages in each category are displayed on the y-axis rather than the frequencies. If this is done, it is important to include the total number of observations that the percentages are based. The choice of number of categories is important as too few categories and much important information is lost, too many and any patterns are obscured by too much detail. Usually between 5 and 15 categories will be enough to gain an idea of the distribution of the data.

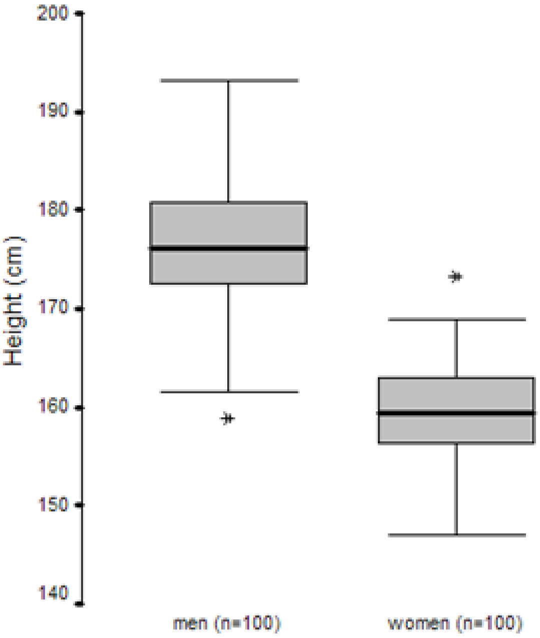
Figure 4: Histogram of heights for 100 men



A useful feature of a histogram is that it is possible to assess the distributional form of the data; in particular whether the data are approximately Normal, or are skewed. The histogram of Normally distributed data will have a classic 'bell' shape, with a peak in the middle and symmetrical tails. The **Normal distribution** (sometimes known as the Gaussian distribution) is one of the fundamental distributions of statistics, and its properties, which underpin many statistical methods, will be discussed in a later tutorial. **Skewed** data are data which are not symmetrical, negatively skewed data have a long left-hand tail at lower values, with a peak at higher values, whilst conversely positively skewed data have a peak at lower values and a long tail of higher values.

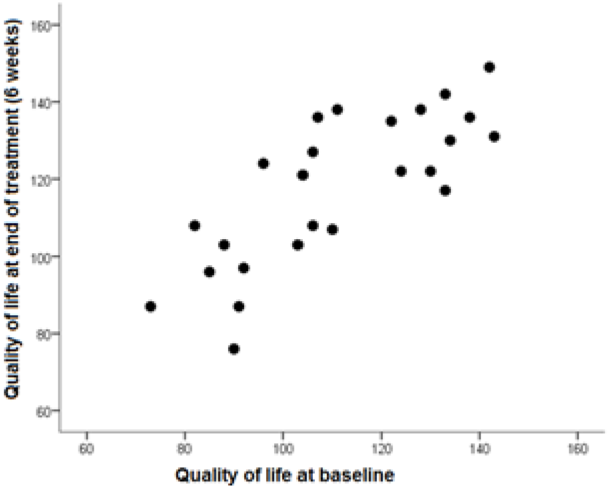
Another extremely useful graph for continuous data is a **box-and-whisker** or **box plot** (Figure 5). Box plots can be particularly useful for comparing the distribution of the data across several groups. The box contains the middle 50% of the data, with lowest 25% of the data lying below it and the highest 25% of the data lying above it. In fact the upper and lower edges represent a particular quantity called the interquartile range. The horizontal line in the middle of the box represents the median value, the value such that half of the observations lie below this value and half lie above it. The whiskers extend to the largest and smallest values excluding the outlying values. The outlying values are those values more than 1.5 box lengths from the upper or lower edges, and are represented as the dots outside the whiskers. Figure 5 shows box plots of the heights of the men and women. As with the dot plots, the gender differences in height are immediately obvious from this plot and this illustrates the main advantage of the box plot over histograms when looking at multiple groups. Differences in the distributions of data between groups are much easier to spot with box plots than with histograms.

Figure 5: Boxplots of height for 100 men and 100 women



The association between two continuous variables can be examined visually by constructing a **scatterplot**. The values of one variable are plotted on the horizontal axis (known as the X-axis) and the values of another are plotted on the vertical axis (Y-axis). If it is known (or suspected) that the value of one variable (independent) influences the value of the other variable (dependent/outcome), it is usual to plot the independent variable on the horizontal axis and the dependent variable on the vertical axis. Although it is not always obvious, it is often clear which variables to place on the X- and Y-axes. Experimentally the X-axis would be something that the experimenter controls while the Y-axis would be the response to the X-axis. Figure 6 shows the scatter plot of quality of life at the start of treatment and at the end of treatment; lower values indicate poorer quality of life. There is a positive association between the two time points: those with poor quality of life at the start had poor quality of life at the end and those with better quality of life had better quality of life at the end.

Figure 6: Quality of life at the start and end of treatment



Good practice recommendations for charts

Good charts have the following four features in common: clarity of message, simplicity of design, clarity of words and integrity of intentions and action(1). A chart should have a title explaining what is displayed and axes should be clearly labelled; if it is not immediately obvious how many individuals the chart is based upon, this should also be stated. Gridlines should be kept to a minimum as they act as a distraction and can interrupt the flow of information. When using charts for presentation purposes care must be taken to ensure that they are not misleading; an excellent exposition of the way in which charts can be used to mislead can be found in Huff(2).

Box 1 : Guidelines for good practice when constructing charts

1. The amount of information should be maximised for the minimum amount of ink
2. Charts should have a title explaining what is being displayed
3. Axes should be clearly labelled
4. Gridlines should be kept to a minimum
5. Avoid 3-D charts as these can be difficult to read
6. The number of observations should be included

Reference List

(1) Bigwood S, Spore M. Presenting Numbers, Tables and Charts. Oxford: Oxford University Press, 2003.

(2) Huff D. How to lie with statistics. London: Penguin Books, 1991.