Research Methods 2

Week 3: Document 2

Data Description: using graphs

The Box and Whisker plot

A simple but nevertheless very useful way of summarising data graphically is the *boxplot*, or *box and whisker plot*. The data on the heights of five-year-old boys described in 'Data Description: using numerical summaries' is shown in figure 1.

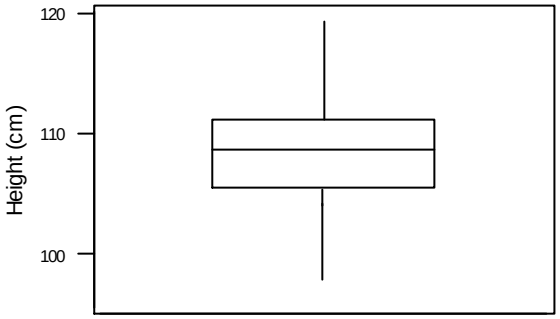

Figure 1

The top and bottom of the 'box' of the box and whisker plot are drawn at the level of the upper and lower quartiles respectively. A line is drawn across the box at the level of the median. This is the usual form of the box and variations from this are very rarely encountered.

The 'whiskers' are lines drawn from the top and bottom of the box to the maximum and minimum value in the sample and this is what is shown in figure 1. This is how the box and whisker plot was originally conceived. However variants on this are quite common. Minitab produces the following boxplot of the height data by default.

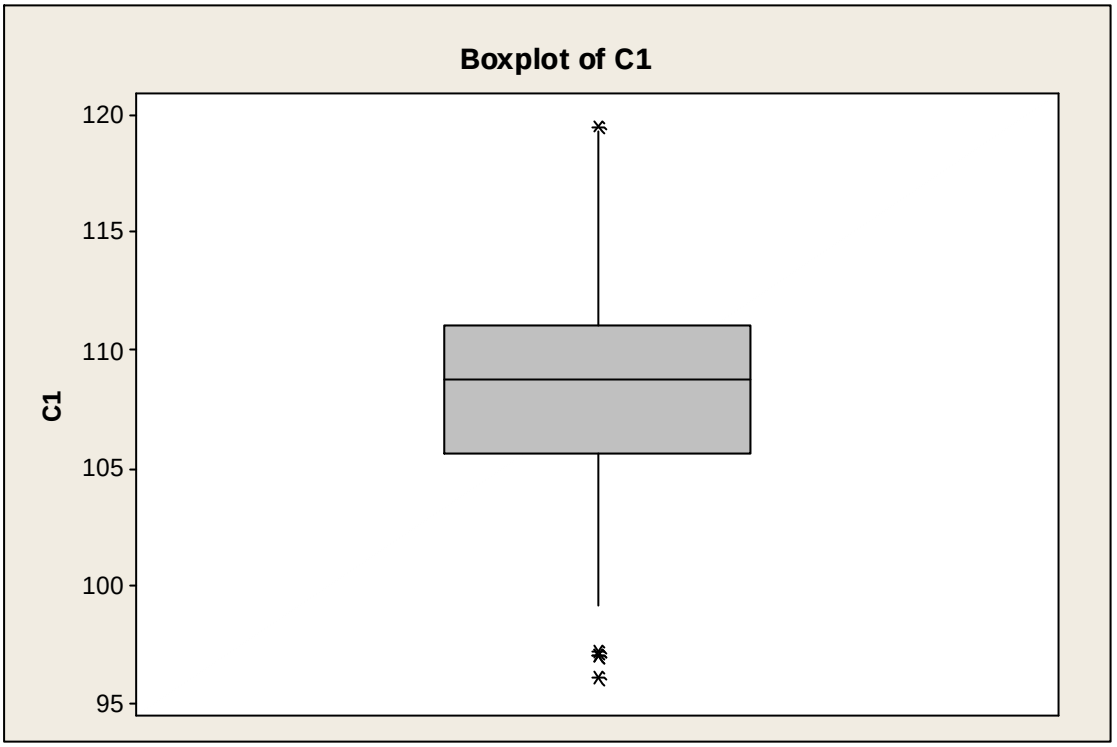

Figure 2

Some points are shown beyond the extent of the whiskers. At first this seems contradictory – how can there be points beyond the minimum and maximum values? Of course there cannot – what has happened is that the whiskers have been defined slightly differently.

The problem with drawing the whiskers out to the maximum and minimum is that if there is a large or small point which is quite out of line from the other values in the sample, this will not be apparent from the conventional box and whisker plot, the extreme point will simply distort the length of the whisker. A slightly more sophisticated approach to the box and whisker plot is to decide on limits as to how widely one might expect the data to be dispersed and to draw the whiskers out as far as the most extreme values *within these limits*. Any points outside these limits are then plotted individually. This is the approach that has been adopted by Minitab in Figure 2. Thus any unusual or outlying values are plotted explicitly and do not simply distort the lengths of the whiskers.

Of course this begs the question of how widely data ought to be dispersed and whether we can prescribe such limits. This is a delicate matter and a careful discussion is beyond the scope of this introduction. It is sufficient to make two points here. First, we are not saying how widely data *ought* to be dispersed, it is simply a way of finding a graphical method that allows us to display data in such a way that unusual points are easily identified. These may be quite genuine observations, in which case we simply have to acknowledge that their value is correct. However, experience shows that in many cases extreme values often arise through some mistake in the process of observation and recording and need to be identified and, if possible, corrected. Second, it turns out that it is sensible to 'expect' data to lie within a few 'boxwidths' of the top and bottom of the 'box' part of the box and whisker plot. This is the way in which the ideas of the paragraph can be implemented. As

mentioned above, the details of this matter are beyond our present scope and you need not be concerned with them.

Histograms

The boxplot, at least in its original conception, is a graphical summary not of the data but of the five number summary of the data.

An alternative approach, which displays the whole of the data, is a *histogram*. The range of the data is divided into intervals of equal width[1] (often called bins) and the number of observations in each interval is counted. The histogram is the plot of bars, one for each bin, with heights proportional to the number of observations in the corresponding bin. The height can be the number in each interval. An alternative is for the height to be given by the number in the bin expressed as a proportion or percentage of the total number of observation in the sample. This gives a slightly more general picture, in that it shows the distribution of the data across its range in a way that is not dependent on sample size.

The analyst must exercise judgment in the choice of number of bins. Figure 3 shows the data on the heights of 99 boys using different numbers of bins. With just two bins there is very little indication of how the heights are distributed. With six bins matters are clearer but a better picture is probably given by the case with 15 bins. When the number of bins increases further there is too little smoothing of the data and the histogram start to look rather jagged.

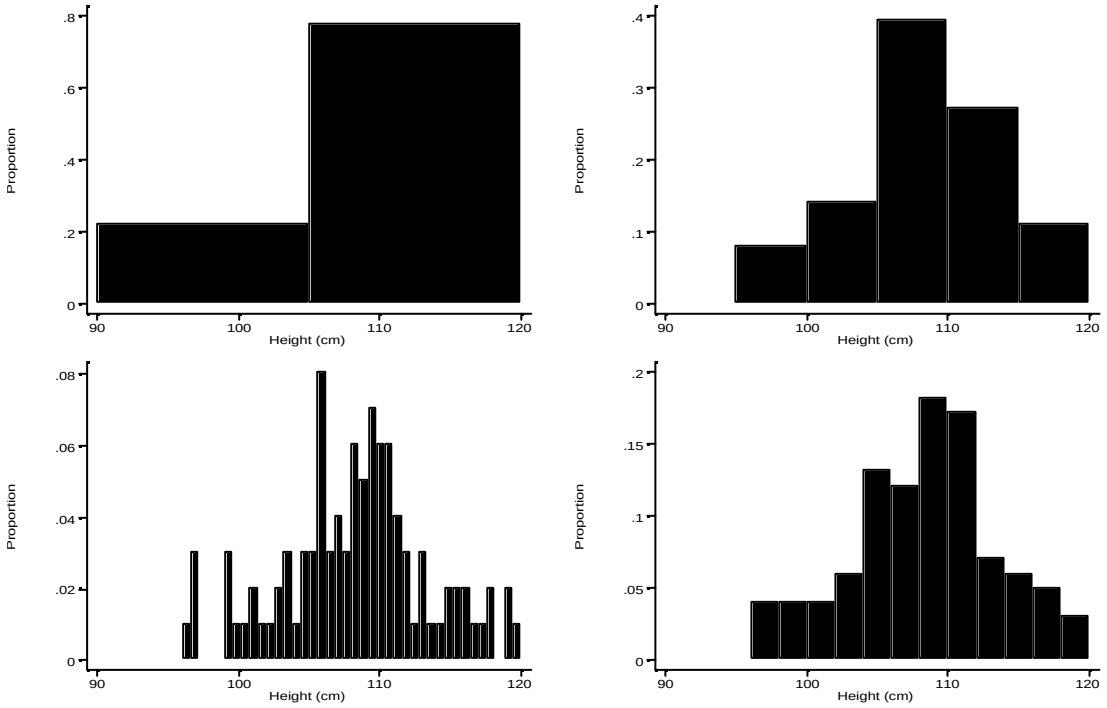

Figure 3: histograms with 2, 6, 15 and 50 bins (clockwise from top left)

[1] The intervals can be of unequal width but this leads to complications and is best avoided

As mentioned above, histograms show the entire sample and allow the full distribution of the sample to be viewed (to within the effects of the chosen bin width). Also, as the size of the sample increases deeper aspects of the nature of the distribution may become apparent. Figure 4 shows histograms for the above sample of 99 heights together with histograms for three larger samples, of sizes 300, 1000 and 10000.

It can be seen that as the sample size gets larger the form of the distribution becomes more regular and seems to approach an idealised form, depicted by the curve which has been superimposed on the last two histograms.

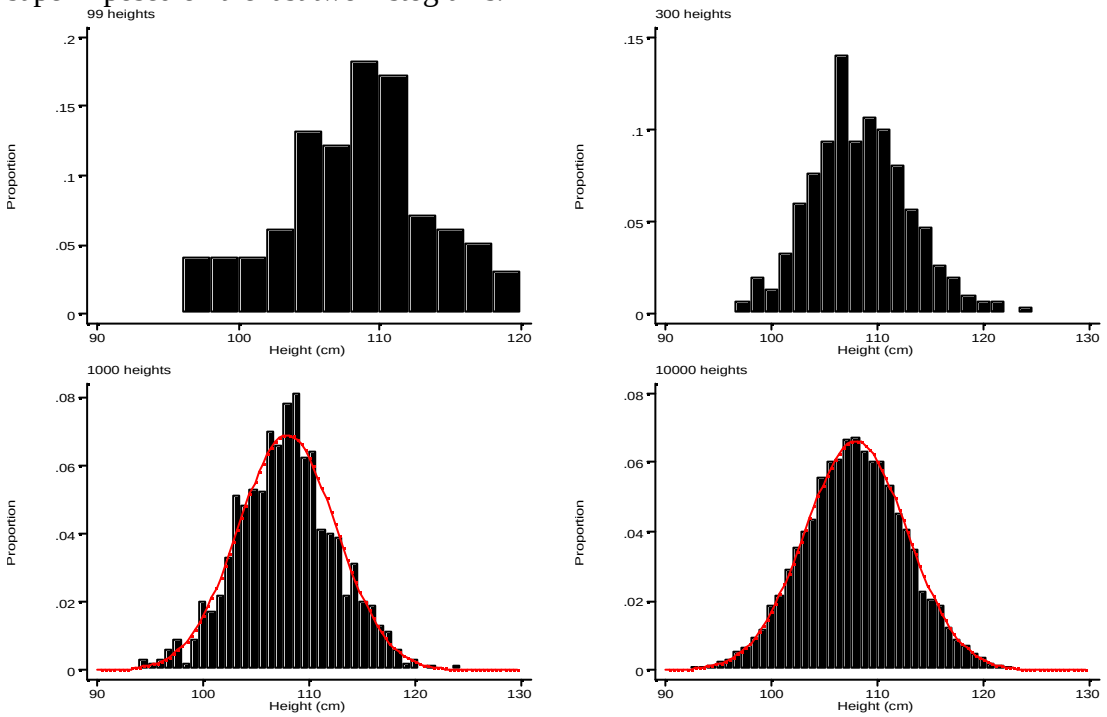

Figure 4: histograms for samples of 99 (top left), 300 (top right), 1000 (bottom left) and 10000 (bottom right) heights

Here the use of histograms has revealed a certain regularity of form, to which it appears the shape of histograms tend as the sample size increases. The form shown in the figure is symmetric and bell-shaped and in this case actually represents the well-known Normal distribution. Not all measurements will tend to distributions with this shape, although many do.

However, before the notion of a Normal distribution can be fully explained, the notion of a *population* must be introduced, and together with it the basic ideas of *inferential statistics*. These will be the matters that concern us next week.