# Chapter 7: Summarising Your Data - Frequency Tables and Charts.

Questions

1. In providing a discussion of these questions, greater detail will be given to how this information is used in data analysis. *Frequency* is the term given to the number of cases with a particular variable value. For example, if there are ten respondents aged 25 years, the frequency is 10. One difficulty though is that some respondents fail to answer a question, or the data has been incorrectly entered into the database. In the former case, this is referred to as a *missing value* and the data cell is left empty. If data has been entered incorrectly and is obviously wrong, you will need to designate this as a *missing value*. In effect, you will delete the data from the database. Where there are a lot of data categories (e.g. respondent age where the range of values is large), it is often more convenient to display the results as *grouped data* to collapse the data into a smaller number of categories that become easier to interpret or publish. For example, you may decide to create age groups such as: 0-4, 5-9, 10-14, 15-19, 20-24. Each category must be of equal interval size. *Cumulative frequency* is a useful measure when dealing with what may be termed continuous interval data. This is the running total of cases as you move through the categories. For example, respondent age is usually a continuous variable and **Table 7.1** in the chapter shows how the running total has been recorded as *cumulative percent*. It is sometimes useful to plot this on a graph and its shape provides information such as *quartiles* and *interquartile range* (see chapter 8). The shape also provides a visual impression of age distribution. The two graphs below were produced using *cumulative frequency* and *cumulative percentage* values of respondent age.
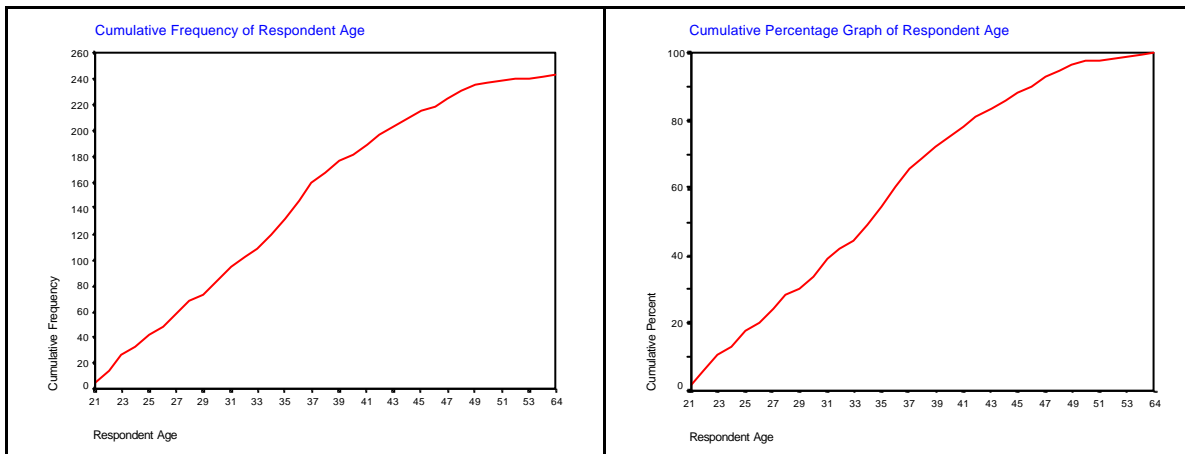

Since the general gradient (or slope) of the graphs is relatively constant up the mid-40's it is possible to state that there is a generally even age distribution of respondents. However, the slope starts to flatten from around ages 50 indicating that older respondents are not as evenly represented amongst the sample. However, that is not the same as saying that the sample is unrepresentative as there may well be fewer older people in the research population. What you would need to do is to compare this cumulative frequency graph of the sample with one for the entire population. If they are similar, then it is fair to say that the sample's age distribution is representative of the population.

2. An *outlier* is a value that is obviously outside the general spread of values. It is either very high or very low compared to the rest of the data. For example, taking the age distribution of the UK population, people aged over 100 years form a very small proportion of the population, but if you found someone aged 110 years, this age would be so high as to lie outside the distribution curve and would be an obvious exception. The effect of *outliers* on the data as a whole can sometimes be profound. For example, in chapter 8 measures of central tendency are described and outliers can have a profound effect by distorting these measures. Several examples of this are discussed in that chapter. It is a paradox, but while an outlier is in many ways more interesting that the remaining cases in your sample (e.g. it is probably more interesting to know how someone has managed to live to be 110 than, say, 90), an outlier may be so different and unusual that their inclusion in the analysis may be inappropriate. The decision to do this must be justified. *Rogue values*, on the other hand, are much easier to deal with. These values are obvious errors and can be designated as a *missing value*. For example, is a respondent's age is recorded as 220, it is probably safe to assume this to be a data entry error. You either correct this by checking the original questionnaire (or data collection form) – assuming this can be identified – or you make it a *missing value*.

3. A *histogram* plots cases for your variable in the form of grouped data. SPSS will do this automatically, but if you use some other form of graphical presentation package, you may need to collapse the data yourself first. The SPSS feature of collapsing the data for you will save you the trouble of having to do it yourself as it generates the appropriate group ranges based on its analysis of the data. There are two histograms below and show slightly different ways of labelling the data.
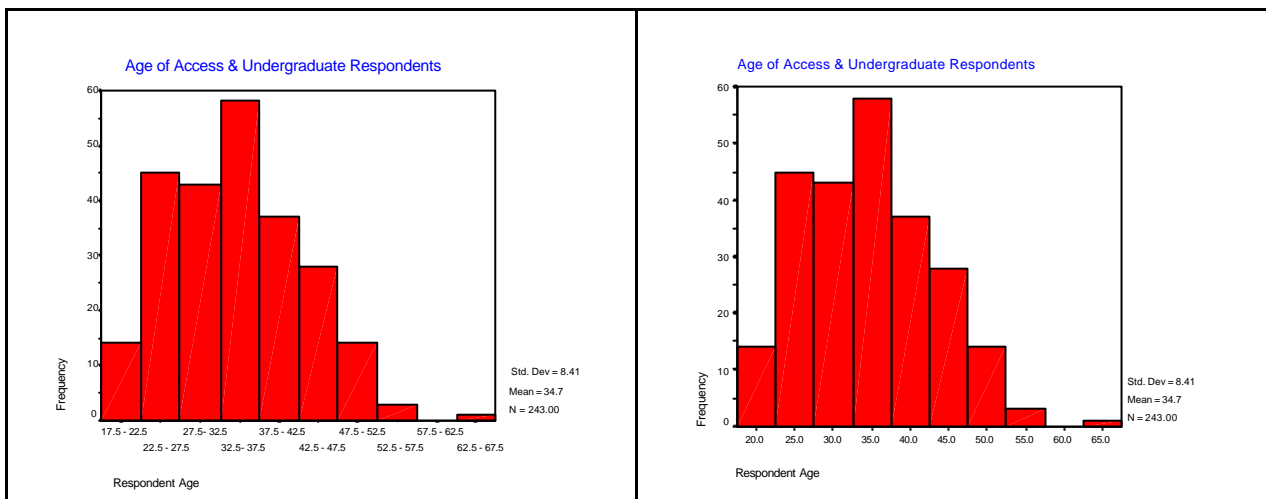
The histogram bars represent the proportion of cases in each category and so there is a range of values within each bar. You can format the horizontal axis to show either the mid-points of each range or the range of values for each bar.

A *boxplot* is designed to show several of the measures described in chapter 8 (i.e. median, 1st and 3rd quartile, interquartile range) and also any outliers. The two whiskers at either end of the plot show the minimum and maximum values. Outliers and extreme values are shown as points above and below the whiskers. SPSS will identify each outlier by labelling it with its case number (i.e. record number in the database). You can identify outliers from this and exclude them from more detailed analysis if that is appropriate. An example of a *boxplot* is shown below.

Boxplot of Mature Students Ages
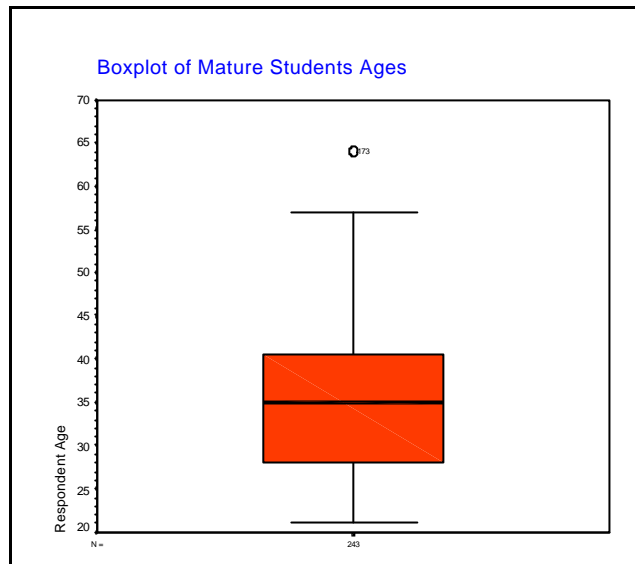
The line across the centre of the shaded box is the ***median*** value, the box itself represents the ***interquartile range*** (see chapter 8). Note the outlier labelled 173 (i.e. record 173 in the database).

Stem and leaf charts show the distribution by displaying every case within grouped data as shown below.

```
stem-and-Leaf Plot of Mature Students

 Frequency     Stem &   Leaf

      4.00        2 .   1111
     22.00        2 .   2222222222333333333333
     17.00        2 .   44444455555555555
     16.00        2 .   6666677777777777
     14.00        2 .   88888888889999
     22.00        3 .   0000000000111111111111
     13.00        3 .   2222222333333
     24.00        3 .   444444444444555555555555
     28.00        3 .   6666666666666677777777777777
     16.00        3 .   8888888999999999
     13.00        4 .   0000001111111
     14.00        4 .   22222222333333
     12.00        4 .   444444555555
     10.00        4 .   6667777777
     10.00        4 .   8888889999
      3.00        5 .   001
      2.00        5 .   23
       .00        5 .
      2.00        5 .   77
      1.00  Extremes    (>=64)

 Stem width:   10
 Each leaf:        1 case(s)
```

Extreme values are also identified in a stem and leaf plot.