3 Graphical Presentation of Data

3.1 Introduction

Graphical displays of data can be very useful in showing the main features of a data set. The appropriate form of graph depends on the nature of the variables being displayed and what aspects are to be shown. However it should be kept in mind that the object is to provide a clear and truthful representation of the data, not to distort and not to impress with unnecessary "fancy" features.

3.2 Qualitative or discrete data: Bar charts

The most useful way to display qualitative or discrete data is usually with a bar chart. In constructing a bar chart, the data are first summarised by counting the frequency with which each distinct value of the variable occurred. The lengths of the bars in the bar chart are then proportional to these frequencies of occurrence. The widths of the bars should be equal to avoid giving a false impression, and a small gap is drawn between each bar to indicate the separate "classes" of data. Figure 1 shows a bar chart for the cell viability data set.

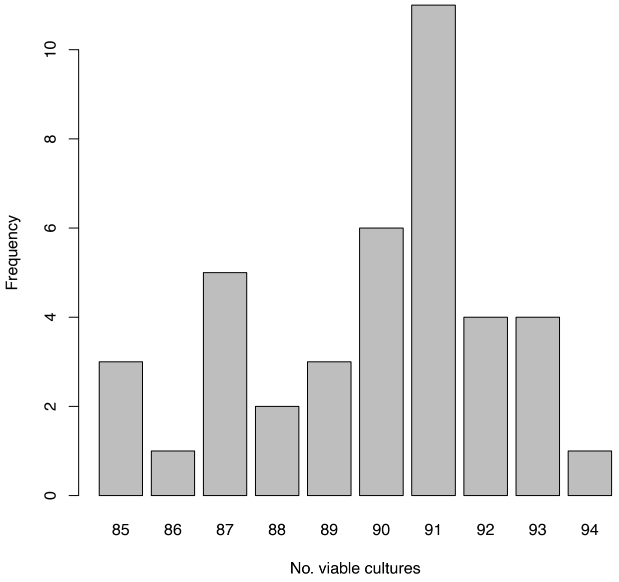

Figure 1: Bar chart of the number of viable cell cultures out of 100 after 14 days.

To do this in R we use the following commands:

```
url = "http://www.mas.ncl.ac.uk/~nseg4/teaching/MAS8380/viability.txt"
viability = read.table(url, header=TRUE)
barplot(table(viability$Number), xlab="No. viable cultures",
    ylab = "Frequency")
```

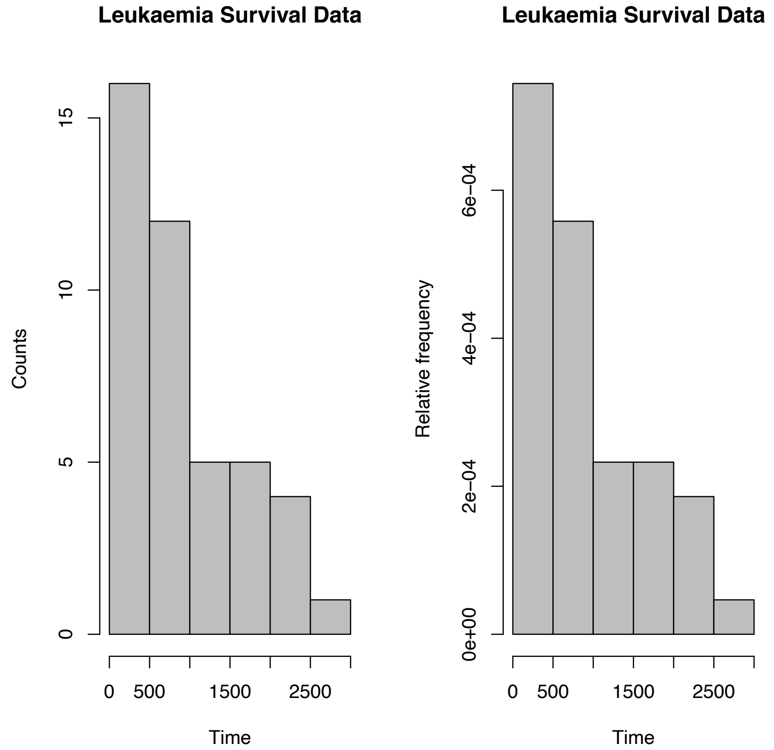

Figure 2: Histogram of the survival time of the leukaemia patients.

3.3 Histograms

Histograms are used to represent the distribution of a sample of values of a continuous variable. The range of values of the variable is divided into intervals, known as bins or classes, and the frequencies in classes are represented by columns. As the variable is continuous, there are no gaps between neighbouring columns, unlike a bar chart. Note also that, strictly speaking, it is the *area* of the column which is proportional to the frequency, not the height. The reason for this is that columns need not be of the same width. However, computer software tends to use columns of the same width. This default can be overridden in R if you really want to. Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation.

The y-axis on a histogram can either show the number of counts in each class (also called the absolute frequency) or the relative frequency of each class. When dealing with relative frequency, we can easily work out the height using this formula:

$$\text{Height} = \frac{\text{number of points in bin}}{n \times \text{ bin-width}}$$

When the y-axis is labelled with relative frequencies, the area under the histogram is always one. If dealing with absolute frequencies, the formula is the same as above, except n is removed from the denominator.

Figure 2 shows a histogram of the survival time of the leukaemia patients.

To generate Figure 2 in R we use the following commands:

```
url = "http://www.mas.ncl.ac.uk/~nseg4/teaching/MAS8380/survival.txt"
survival = read.table(url, header=TRUE)
par(mfrow=c(1,2)) # Creates a multi-panelled plotting window with 1 row
                  # and 2 columns.
hist(survival$Time, col="grey", main="Leukaemia Survival Data",
     freq=TRUE, xlab="Time", ylab="Counts")
hist(survival$Time, col="grey", main="Leukaemia Survival Data",
     freq=FALSE, xlab="Time", ylab="Relative frequency")
par(mfrow=c(1,1)) # Ensures next plot is produced in a single-panelled
                  # plotting window, unless there is another call to
                  # par(mfrow) before it is created.
```

3.4 Stem and leaf plots

Stem and leaf plots are used in a similar way to histograms to represent continuous data, but each bar “contains” actual observations. This is best illustrated with an example (for wind speed in the New York air quality data):

```
The decimal point is at the |

 1 | 7
 2 | 38
 3 | 4
 4 | 016666
 5 | 111777
 6 | 33333333999999999
 7 | 4444444444
 8 | 0000000000066666666
 9 | 2222222277777777777
10 | 3333333333399999999
11 | 555555555555555
12 | 0000666
13 | 2288888
14 | 33333399999999
15 | 555
16 | 1666
17 |
18 | 4
19 |
20 | 17
```

Values on the left (the “stems”) give the first decimal digit(s), and those on the right (the “leaves”) the subsequent digit for each observation in the data set. In this example the

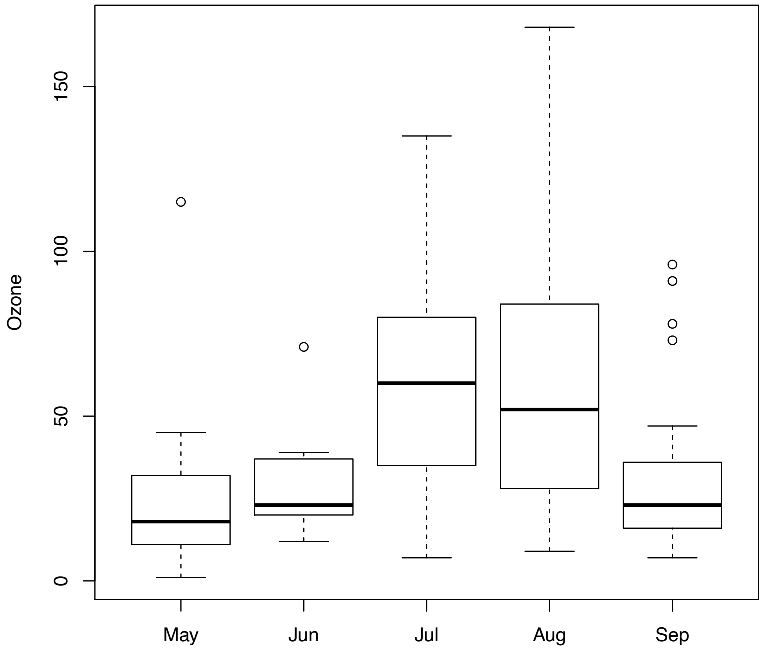

Figure 3: New York air quality data: ozone by month

decimal point is at the | and so the stems represent multiples of 1 and the leaves represent multiples of 0.1, e.g. the first row corresponds to the observation 1.7 and the bottom row corresponds to the observations $20.1, 20.7$.

The plot was generated in R using the following commands:

```
data(airquality)
stem(airquality$Wind)
```

3.5 Boxplots

A boxplot, sometimes called a box-and-whisker plot, is another way to represent continuous data. This kind of plot is particularly useful for comparing two or more groups, by placing the boxplots side-by side. Figure 3 shows a boxplot for the New York air quality data, in this case displaying the ozone data by month.

The central bar in the "box" is approximately the sample *median*. Remember, the sample median is a value M such that half of the observations are greater than M and half are less than M. The top and bottom of the box represent the upper and lower sample *quartiles*. Just as the median represents the 50% point of the data, the lower and upper quartiles represent the 25% and 75% points respectively.

The "whiskers" extend to the maximum and minimum observations, except that this is not always quite true! You will notice that a small number of observations are marked individually beyond the ends of the whiskers. These are “outliers”. That is, they are observations which, according to some rule, are unusually far from the box. These are therefore shown separately and are not used in the calculation of the “median” and “quartiles.” The values used for the box are therefore not exactly the median and quartiles and are given a different name. They are known as “hinges.”

To do this in R we use the following commands:

```
data(airquality)
boxplot(Ozone ~ Month, data = airquality, names =
   c("May", "Jun", "Jul", "Aug", "Sep"), ylab="Ozone")
# Here the formula Ozone ~ Month groups the numeric variable Ozone
# by Month.
```

3.6 Summary

Using the plots described here, we can gain better understanding of the important features of data:

- Is the distribution symmetric or asymmetric?
- Are there any unusual or outlying observations?
- Are the data multi-modal?
- By putting plots side by side with the same scale, we may compare the distributions of different groups.