1.1 A Framework for Statistical Problems

Many statistical problems can be described by a simple framework in which we have:

- a population of objects
- a real-valued variable X associated with each member of the population
- some quantity of interest determined by the overall distribution of X values in the population
- a sample of n members of the population
- a data set $x_1, \ldots, x_n$ of observed values of X for the sampled members.

The problem is to infer the unknown value of the population quantity from the known sample data.

For example, the population might consist of all students graduating from a particular university in a given year, the variable might be the level of debt incurred by each student at the time of graduation, and the quantity of interest might be the average level of debt in the population. Alternatively, the population might consist of all lightbulbs that have been produced by a certain manufacturer, the variable might be the lifetime of each lightbulb, and the quantity of interest might be the proportion of lifetimes that exceed two years.

More generally, we may have sample data from more than one population and we may want to determine whether there is any pattern in the way the quantity of interest varies from population to population. For example, we may have data on the debt at graduation for a sample of students from several universities, and we may want to explore how the average debt in the population of graduating students varies from university to university. Or we may have data on the lifetimes of a sample of lightbulbs from several manufacturers, and we may want to explore how the proportion of lifetimes in the population that exceed two years varies from manufacturer to manufacturer.

Often, we can reasonably assume that the sample has been chosen in such a way that:

- each population member is chosen independently of the other sample members, and
- each population member is equally likely to be included in the sample.

In this case, we say that the sample data values $x_1, \ldots, x_n$ are the observed values of a *simple random sample* of size n from the population.

For simple random samples, the data values are representative of the values in the population as a whole, in the sense that, on average, different values occur in the sample in the same proportion as they occur in the population. Thus we can use the data values from the (possibly small) sample to make inferences about the values in the population as a whole.

1.2 Exploratory Data Analysis (EDA)

Exploratory Data Anaylsis is a descriptive name for a collection of techniques for initial exploration of a data set.

EDA can provide us with a means of checking that the data is compatible with our assumptions that:

– the observations are independent
– the observations all come from a common distribution
– and any assumptions (as we will see later) about the type of distribution

EDA can also provide us with simple direct estimates of some population quantities, in a form which does not depent on the assumption of any particular type of distribution.

Features of EDA which we will use include:

- Initial graphical plots
  – stem-and-leaf plots
  – histograms (or bar charts)
  – time plots
  – boxplots
- Numerical summaries of the centre or location of the data
  – median, mean, trimmed mean
- Numerical summaries of the spread of the data
  – variance and standard deviation, hinges, quartiles and inter-quartile-range
- In graphical plots, the things you should check include:
  – the overall pattern/shape of variation within the data (e.g. symmetric, skew, bi-modal)
  – any unusual features within a pattern or striking deviations from a pattern (outliers)
  – whether any unusual features are just random occurrences or are systematic features
  – any evidence of clustering or granularity (data clumping at certain sequences of values, reflecting measurement scale)
- For more than one variable, begin by examining each variable by itself, then move on to study relationships between variables.
- For data from more than one population, the things you should check include:
  – evidence of variation/pattern within each data set relative to variation between data sets
  – using numerical (e.g. summary statistics) or graphical (e.g. boxplots) summaries

1.3 Initial Graphical Plots

- **Stem-and-Leaf Plot**

1. If necessary, truncate or round the data values so that all the variation is in the last two or three significant digits.
2. Separate each data value into a **stem** (consisting of all the digits except the final rightmost digit), and a **leaf** (the final rightmost digit).
3. Write the stems in a vertical column – smallest at the top – and draw a separator (e.g. a vertical line) to the right of this column.
4. Write each leaf in the row to the right of the corresponding stem, in increasing order out from the stem.
5. Record any strikingly low or high values separately from the main stem, displaying the individual values in a group above the main stem(low values) or below the main stem (high values).

A stem-and-leaf plot of a set of data can be produced in R using the command `stem`.

For example, R contains data on the duration in minutes of eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The data is stored in **R**, under the variable name `eruptions` in the dataframe `faithful`. The command `data(faithful)` includes the dataframe in the current workspace, the command `attach(faithful)` makes available the variables in `faithful`, and the command `stem(eruptions)` then produces the plot.

```
> data(faithful)
> attach(faithful)
> stem(eruptions)
```

- **Histogram**

1. Divide the range of data values into say K intervals (cells or bins) of equal width. If the width is too large, the plot may be too coarse to see the details of any pattern; if it is too small there may be lots of cells with just one or two observations.
2. Count the number (frequency) or the percentage of observations falling into each interval. Be consistent with the allocation of values that equal the end points of intervals.
3. Display the outcome as a plot of joined columns or bars above each interval, with height proportional to the count or percentage for that interval.

A histogram of a set of data can be produced in R using the command `hist`.

For example, having brought up and attached the Old Faithful data as above, you can plot a histogram with the command:

```
> hist(eruptions)
```

- **Time Plot**

  A plot of the data in the order it was obtained (or recorded) may give valuable information when the values represent the successive outcomes of repetitions of a single statistical experiment repeated over time.

  In **R** this be obtained with the command `plot`, e.g.

```
> plot(eruptions)
```

- **Boxplot** and **Five number summary (including Hinges)**

  A boxplot is a very simple graphical summary of a data set, devised by John Tukey. It is based on just five numbers calculated from the data – the median, the upper and lower hinges (similar to quartiles) and the maximum and minimum values. These numbers roughly divide the data into four equally-sized groups.

  The plot consists of a box with the top drawn level with the value of the upper hinge, the bottom drawn level with the value of the lower hinge, and a horizontal line drawn across the middle, level with the median. Vertical lines (sometimes called *whiskers*) are drawn from the top of the box to a point level with the maximum value, and from the bottom of the box to a point level with the minimum value.

  Note that if there are any *outliers* (points at a distance of more than 1.5 times the inter-quartile range (IQR) away from the box), then the whiskers are drawn to the largest data value within $1.5 \times$ IQR from the corresponding hinge and the remaining outlier data points are plotted individually.

  In **R** a boxplot and the corresponding five numbers on which it is based can be obtained with the commands `boxplot` and `fivenum`, e.g.

```
> boxplot(eruptions)
> fivenum(eruptions)
```

  A typical boxplot looks something like the following plot.

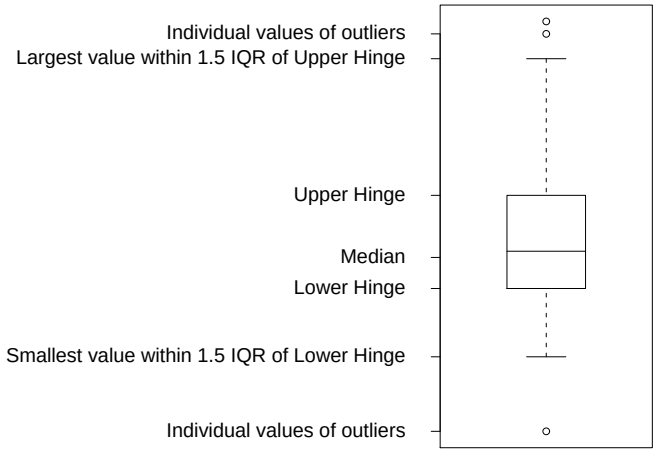

Figure 1:

1.7 Example - Earthquakes data

The `quakes` data set records the time in days between successive serious earthquakes worldwide. An earthquake was included if its magnitude was at least 7.5 on the Richter scale, or if over 1000 people were killed. Recording started on 16 December 1902 and ended on 4 March 1977. There were 63 serious earthquakes and so 62 recorded waiting times. The data values vary in size from `9` to `1901`. Note that you may need to load the folder of data sets for the Statistics 1 unit, but you only need to do this once each session. You can inspect the data sorted in increasing size by typing `sort(quakes)` after the prompt, and produce a stem-and-leaf plot with the command `stem(quakes)` as follows:

```
> load(url("http://www.maths.bris.ac.uk/%7Emapjg/Teach/Stats1/stats1.RData"))

> sort(quakes)
 [1]    9   30   33   36   38   40   40   44   46   76   82   83   92
[14]   99  121  129  139  145  150  157  194  203  209  220  246  263
[27]  280  294  304  319  328  335  365  375  384  402  434  436  454
[40]  460  556  562  567  584  599  638  667  695  710  721  735  736
[53]  759  780  832  840  887  937 1336 1354 1617 1901

> stem(quakes)
The decimal point is 2 digit(s) to the right of the |
   0 | 133444445888902345569
   2 | 01256890234788
   4 | 034566678
   6 | 0470124468
   8 | 3494
  10 |
  12 | 45
  14 |
  16 | 2
  18 | 0
```

Note that R decided to put the decimal point 2 digits to the right of the | (the bar) and use a scale where each stem corresponds to intervals of 200 days – and hence each leaf corresponds to an interval of 10 days. Thus the first line represents (rounded) data values of 10, 30, 30, 40, 40 ,. . . , 90, 100, 120, . . . , 190 in that order. Because of the scale and small number of data values, it can be difficult to tell that the last line, for example, represents the data value 1900 rather than 1800. You can change the scale on which the data is displayed by specifying the `scale`. For example, you might try `stem(quakes, scale=2)` which produces a scale where each stem corresponds to 100 days.

```
> stem(quakes, scale=2)
The decimal point is 2 digit(s) to the right of the |
   0 | 1334444458889
   1 | 02345569
   2 | 0125689
   3 | 0234788
   4 | 03456
   5 | 6678
     ⋮
  18 |
  19 | 0
```

Similarly, you can produce a histogram of the times between earthquakes. The standard histogram is produced using the command `hist` below.

```
> hist(quakes)
```

The plots can be customised in many ways using sub-commands, such as:

- by using `freq=FALSE` to display density (i.e. proportions) rather than frequency counts
- by specifying `breaks` to give a certain number of cells or to give cells of a desired width;
- by adding titles using `main="Plot name - Your ID"`;
- by adding labels to axes using `xlab="Label for X axis"` or similarly for `ylab`.

ALWAYS MAKE SURE YOU HAVE ADDED YOUR OWN ID before printing a plot.

For example we can customise the histogram above as follows:

```
> hist(quakes, breaks=seq(0,2000,100), freq=FALSE,
+ xlab="Time between earthquakes in days",
+ main="Earthquakes - mapjg")
```

which gives a histogram of proportions rather than frequencies; with the breaks between cells forming a sequence starting at 0 days, finishing at 2000 days, and of length 100 days apart; adds a label to the x-axis; and revises the title of the plot.

Note how the prompt in **R** changes from `>` to `+` when a command is continued onto a new line.

The standard histogram is shown on the left below, and the customised version on the right.

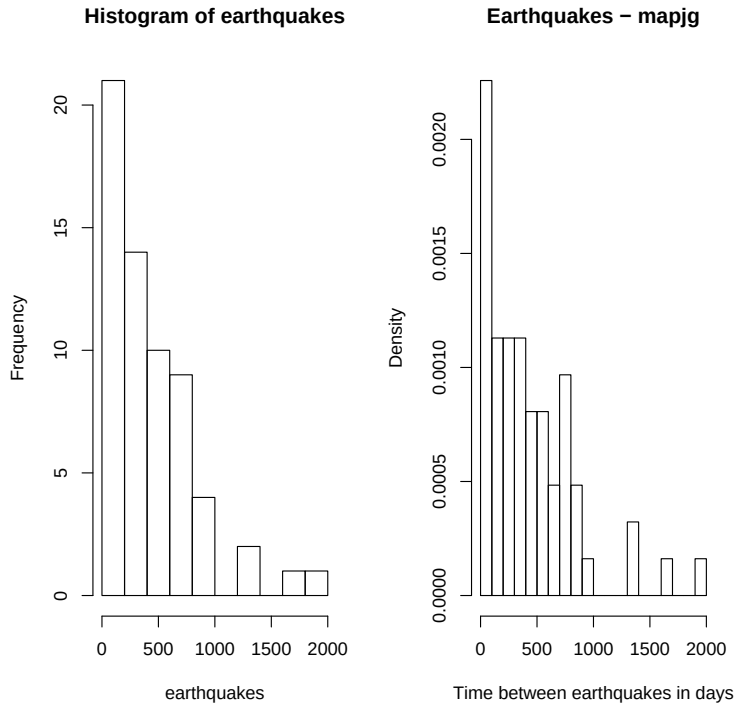

Figure 2:

1.8 Example - Newcomb data

The `newcomb` data set records the results of a series of 66 experiments performed by Newcomb in 1882 to try to determine the speed of light. Newcomb measured the time in seconds that it took for a light signal to pass from his laboratory to a mirror and back, a total distance of about 7400 metres. The data values here record each time in terms of its deviations from a standard time of 0.000024800 seconds - so a data value of 28 indicates a recorded time of 0.000024828 seconds. The commands shown below first display the data, then produce a standard time plot and then a more customised histogram. If the Statistics 1 data sets are not currently loaded, you may need to type
`load(url("http://www.maths.bris.ac.uk/%7Emapjg/Teach/Stats1/stats1.RData"))`

```
> newcomb
[1] 28 26 33 24 34 -44 27 16 40 -2 29 22 24 21 25 30 23 29 31
[20] 19 24 20 36 32 36 28 25 21 28 29 37 25 28 26 30 32 36 26
[39] 30 22 36 23 27 27 28 27 31 27 26 33 26 32 32 24 39 28 24
[58] 25 32 25 29 27 28 29 16 23
> plot(newcomb, main="Plot of Newcomb data - mapjg")
> hist(newcomb, breaks=seq(-45, 45,2.5),
+ main="Histogram of Newcomb data - mapjg")
```

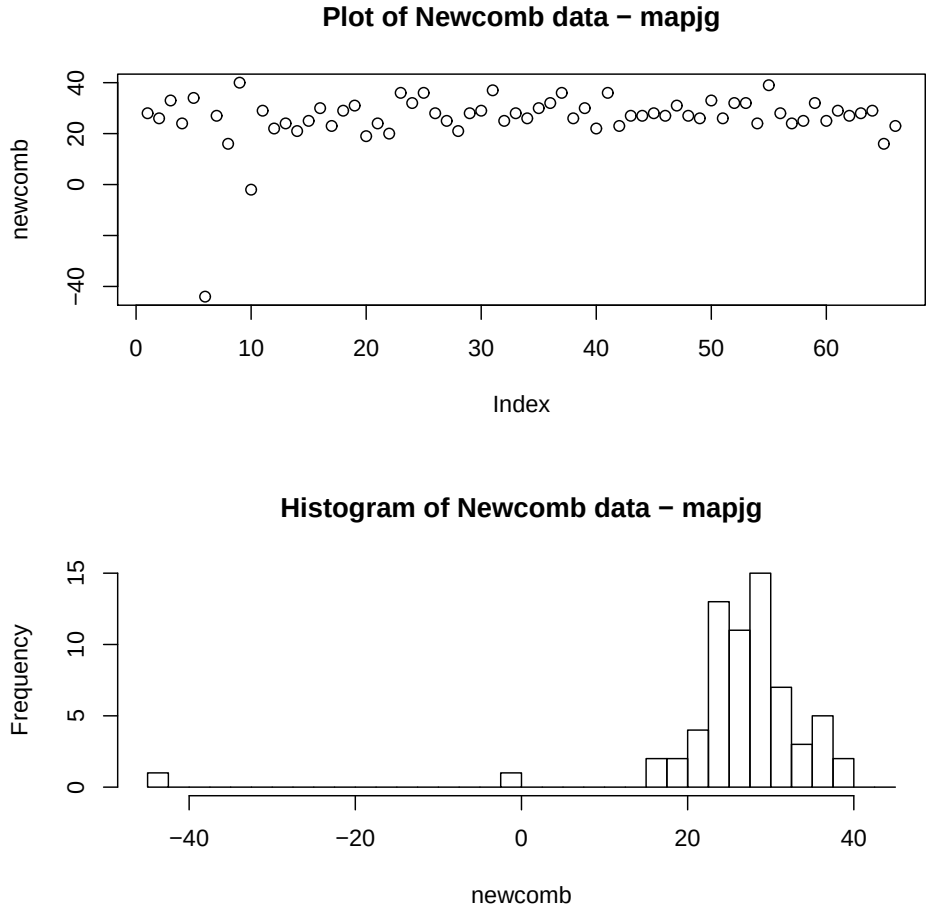

Figure 3: Time plot and histogram of the Newcomb data

We can use R commands `median(), mean(), summary(), var(), sd()` and `IQR()` to calculate a number of helpful numerical summaries.

```
> median(newcomb)
[1] 27
> mean(newcomb)
[1] 26.21212
> mean(newcomb, trim = 0.1)
[1] 27.42593
> mean(newcomb, trim = 0.2)
[1] 27.35
> summary(newcomb)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -44.00   24.00   27.00   26.21   30.75   40.00
> var(newcomb)
[1] 115.462
> sd(newcomb)
[1] 10.74532
> IQR(newcomb)
[1] 6.75
```

We can also produce a graphical summary of the data with the command `boxplot()`, for which the relevant numerical values are given by the command `fivenum()`.

```
> boxplot(newcomb, main="Boxplot of Newcomb data - mapjg")
> fivenum(newcomb)
[1] -44  24  27  31  40
```

The boxplot scale can be distorted by the presence of outliers. We can produce a boxplot of the `newcomb` data set which *excludes* the outliers (the 6th and 10th data values) with the command `boxplot(newcomb[-c(6,10)])`. The standard boxplot is shown on the left below, while a boxplot without outliers is shown on the right.

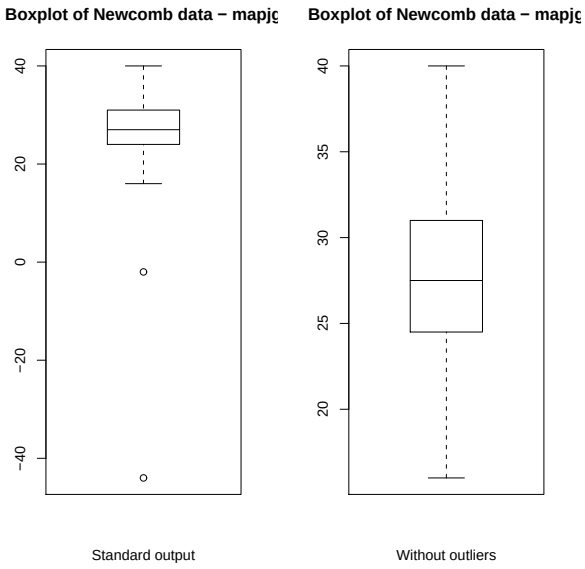

Figure 4: