Practical 2: Plotting

Complete this sheet as you work through it. If you run into problems, then ask for help - don't skip sections! Open `Rstudio` and store any files you download or create in a directory called `Practical2`.

1 The data

Downloading the data

In today's practical, we are going to examine the big bang data that was collected in the last lecture.

To load the data from the quiz, type the following:

```
> bb1516 = read.table("http://www.mas.ncl.ac.uk/~nlf8/
  + teaching/mas1802/prac/bb1516.txt",header=TRUE)
```

Do not enter the + sign. I have done this because the code would not fit on a single line in this handout – you should enter the code in a single line!

This downloads the data from the MAS1802 webpage and stores it in the dataframe `bb1516`. The argument `header=TRUE` tells **R** to use the column headings given in the file.

Provided this has all worked smoothly, you should now be able to view the data by typing:

```
> bb1516
```

However, the file might be too big for you to view conveniently on the screen. You can get an idea of what the data looks like by using the `head` function, as was demonstrated in the introductory practical:

```
> head(bb1516)
```

Each row in the dataframe corresponds to a team from last Monday's quiz, with their guesses of the ages of the actors/actresses in each of the columns labelled with a character's name. "Watched" takes the value `TRUE` if most of the team watch The Big Bang Theory; "Gender" takes the values "Male", "Female" or "Both", depending on the make-up of the team. Before you move on, please make sure you familiarise yourself with the dataframe in **R**!

Initial investigations

Since we are going to alter the data set, it's a good idea to create a copy, so assign the dataset to another object – say `d`:

```
> d = bb1516
```

We can inspect the column names using

```
> colnames(d)
```

and we can also look at individual columns, using

```
> d$Name
> d$Shel
```

The dimensions of `d` are found using

```
> dim(d)
```

Complete: Number of columns: _____ Number of rows: _____

2 Scatter plots

2.1 The basics

Let's start with a scatter plot for Group ID against the Sheldon guesses:

```
> plot(d$ID, d$Shel)
```

The default x- and y-axis labels aren't very good, so we use `xlab` and `ylab` arguments, i.e.

```
> plot(d$ID, d$Shel, xlab="Group ID", ylab="Guess")
```

We can highlight the actual age using **abline**, i.e.

```
> abline(h=42)
```

Other commands for **abline** are

```
> abline(h=39, col=1)
> abline(h=40, col=2)
> abline(h=41, col=3,lty=2)
```

Complete: What does the `col` argument do? _____

Complete: What does the `lty` argument do? _____

Complete: What does the `h` argument do? _____

Complete: What happens if you change the `h` argument to `v`? _____

2.2 Separating by Gender

We will now use scatter plots to investigate differences between male and female guesses. First we need to separate the guesses into different categories. The following piece of code extracts the female teams...

```
> x_female = d$ID[d$Gender=="Female"]
> y_female = d$Shel[d$Gender=="Female"]
```

... which we can then plot

```
> plot(x_female, y_female, pch="F")
```

Of course we can add in the `xlab` and `ylab` arguments.

Complete: Create new variables `x_male` and `y_male`. Add these points to your scatter plot, i.e.

```
> points(x_male, y_male, pch="M")
```

Make sure you include all teams in your scatter plot. To extend the x-axis, you may have to explicitly state the x-limits:

```
> ## Change XX & YY to something sensible
> plot(x_female, y_female, pch="F", xlim=c(XX, YY))
```

Complete: What happens if you omit the `pch` argument? _____

Complete: What happens if you change `pch="F"` to `pch=1` or `pch=2`? _____

Complete: Do you think that males guess the age of `Sheldon` better than females?_____

3 Summary Statistics

Use the commands **mean**, **median**, **quantile** and **sd** to calculate summary statistics for the `Sheldon` column. *Hint: look at Chapter 3 in your lecture notes!*

Complete: Mean: _____, Median: _____

Complete: Standard deviation: _____, Q1: _____, Q3: _____.

4 Histograms

We will now investigate the distribution of guesses using histograms. To plot a histogram of guesses for `Sheldon` we use the following code – **use the `xlab` and `ylab` commands to add labels to the axes, and also the `main` command to add a main title to your plots**.

```
> hist(d$Shel)
```

We can also add in vertical lines to show the mean and median of guesses:

```
> abline(v=mean(d$Shel))
> #What does the lty argument do?
> abline(v=median(d$Shel), col=2, lty=2)
```

We can generate two histograms on the same plot using the `par` command, e.g.

```
> par(mfrow=c(1, 2))
> hist(d$Shel, main="Sheldon Guesses")
> abline(v=mean(d$Shel), col=2)
> hist(d$Penny, main="Penny Guesses")
> abline(v=mean(d$Penny), col=2)
```

Complete: How are the graphs different? _____
Hint: Think about summary statistics.

Usually **R** makes a good initial guess when constructing histograms, but we often have to vary the number of bins to achieve the best result. We can do this using the `breaks` argument, i.e.

```
> hist(d$Shel, breaks=5)
> #You can also use the "scott" or the "fd" options as in lectures
> #Typically, these are better than the default.
> hist(d$Shel, breaks="scott")
> hist(d$Shel, breaks="fd")
```

Which rule would you use for the `Shel` data set – i.e. Default, Scott or FD? _____

5 Boxplots

Let's now investigate the error in age guessing:

$$|\text{actual age} - \text{estimated age}|$$

Sheldon's actor's true age is 42, so the absolute error is:

```
> shel_err = abs(42-d$Shel)
```

To do a boxplot of the error, we use the command:

```
> boxplot(shel_err)
```

We can also investigate if people who have watched BB have an advantage when guessing:

```
> #Remember you can use xlab & ylab for labels!
> boxplot(shel_err ~ d$Watched)
> boxplot(shel_err ~ d$Watched, col="bisque") #In colour
```

Do you think watching BB increased guessing precision? _____

The assignment questions are on the next page

Assignment

Before you begin

Points to note:

- Remember to label your axes correctly.
- For this practical, copy your graphs into Word and type up your comments.
- If you are asked to submit histograms, try different bin number rules.
- Don't go overboard with the number of decimal places you quote.
- If you have more than four pages you will lose marks.
- If you have fewer than four pages you will lose marks.
- You can submit your own individual piece of work, or provide a group submission with your friends - up to **five students maximum**. It's up to you - just make sure all names are written clearly on the cover sheet for the assignment.

What I want

The marks for each question are indicated below:

- **Page 1 [25%]:** Choose a Big Bang Theory actor (don't use Sheldon!), and plot a scatter plot of Group ID vs Guess.
  - Add a line to indicate the mean guess.
  - Add two other lines to represent the mean $\pm$ 2 standard deviations. *Hint: use* **sd** *to get the standard deviations, then use* **abline**.
  - Use the `lty` argument to distinguish between the lines.
- **Page 2 [25%]:** Produce a boxplot of the error against Gender for the Big Bang actor you previously chose. Do you think that females are better than males at guessing? What about mixed sex groups?
- **Page 3 [25%]:** Produce histograms of the error for four different characters. Put the histograms on the same plot. Use the **abline** command to indicate your group's guess on each of the plots. Your scales on the four histograms should be the same.
- **Page 4 [25%]:** Guesses from last year's class can be found by loading the "bb1415" dataset in the same way you loaded the "bb1516" dataset.
  - Using some graphs and summary statistics compare this year's guesses to those from last year. Do you think the class in 2014/15 are better at guessing than the this year's class? I don't went every possible plot and summary statistic. Just a brief overview.