Part III Systems Biology

MAN – Extended Exercise

Matt Castle – mdc31@cam.ac.uk

Overview:

This practical consists of an extended exercise that will allow you to practice and utilise all of the different skills that we have introduced over the past few days. It is deliberately designed to replicate a real and realistic piece of work that you might be expected to undertake within your lab. There are two slightly different versions of this exercise:

- The first utilises data that already in a tidy format and merely requires you to work out a way of combining everything using the tools we've already covered.
- The second starts with the data stored in a common, but distinctly non-tidy format, and will require much more effort to work out a way to get the data into a usable format. It will require use of functions that we haven't explicitly introduced and so you will have to use Google to find out what they are and how they work. This is a much more realistic situation, and one that you will encounter in your future work and as such I strongly recommend that you at least have a go at this version.

The versions only differ in the skills you require to get the data into R in the first place and both versions ask you to perform the same analyses on the datasets at the end.

1 Exercise

An experiment investigated the effect of diet on the early growth of chicks. A number of chicks were fed one of four different diets and their weight (in g) was measured at birth and then at every other day until day 20. A final measurement was made on day 21.

1.1 Version 1 – easy

The data are stored in text files in a directory entitled “EE_easy”. There is one text file for each chick. The format of the data in each text file is the same:

- each file contains a tidy table with four columns: ID, Diet, Days, Weight

Create a function that that reads in the data from a single file and stores it in a single R object. The function should:

- **Accept a single argument (a character string of the file location)**

1.2 Version 2 - challenging

The data are stored in text files in a directory entitled “EE_chal”. There is one text file for each chick. The format of the data in each text file is the same:

- The name of the file gives the ID of the chick
- The first line in the file gives the diet (1,2,3 or 4)
- The number of days since birth is given in the first column with the weight of the chick in the second column.

Create a function that reads in the data from a single file and stores it in a single R object. The function should:

- **Accept a single argument (a character string of the file location)**
- **Extract the chick ID from the filename**
- **Get the diet from the first line of the file**
- **Get the days and weight timeseries from the remaining lines in the file.**

Hints:

- You may have to read the file in more than once to get all of the data).
- The following functions may be useful:
  - `read.table() (and the arguments skip and nrows)`
  - `rep(), numeric(), character()`
  - `paste(), strsplit()`

1.3 For both versions

1. **Modify this script to read in all files and store the data in a single data frame object called "my_chicks". The data frame should have four columns entitled "ID", "Diet", "Days", and "Weight"**

   Hints:

   - The function `list.files()` might be useful here

2. **How many chicks died before the end of the experiment? Modify the script to create a separate data frame called "my_chicks_survive" that only includes the data from surviving chicks.**

3. **Modify your script to perform the following manipulations:**
   a. **Calculate the mean increase in weight (end weight minus birth weight) for all of the surviving chicks.**
   b. **Calculate the mean increase in weight for each subset of surviving chicks depending on their diet.**
   c. **Produce box plots of increases in weight for the surviving chicks on each diet (i.e. four box plots side by side)**
   d. **Find the first recorded time that each surviving chick weighed more than 130g and produce a histogram of these times. (Remove from the analysis any chicks that didn't ever make this weight)**
   e. **Produce a single scatter graph plot which shows the weight of each chick against time. Colour the lines differently depending on the diet of each chick.**
   f. **Calculate an average growth curve for each diet (i.e. calculate the average of the weights of each chick at each time point) and produce a single plot showing the four average growth curves.**
   g. **Construct a data frame with 6 columns and write it to a text file.**
      i. **The first column should contain the original times.**
      ii. **The second column should contain the mean weights of all of the chicks on all diets (who survived).**
      iii. **The third through sixth columns should contain average weights of the chick on each diet expressed as a percentage of the mean average weight of all chicks.**