

STATISTICAL TECHNIQUES FOR THE BIOMEDICAL SCIENCES

BMR 617

SPRING 2023

Assignment: Creating an R script to answer a statistical question.

The purpose of this assignment is to get you used to the way you will carry out and submit many of the assignments in this course. Most assignments will involve answering some questions about data. You will analyze the data in R. In order to keep a record of the analysis, the steps in R will be stored in a file known as an “R script”. When you submit your assignment, you will submit (at least) two files: a Word document with the answers to the questions, and the R script that was used to find those answers.

In this introductory assignment, we’ll do some very basic statistics. Since it’s your first time creating an R script and carrying out the analysis this way, you’ll be guided step by step through a solution.

Cholesterol levels of TH and B6 mice fed on a Chow diet.

Six mice from the TALLYHO (“TH”) strain and four mice from the C57/BL6 (“B6”) strain were fed a standard chow diet from weaning until 16 weeks of age. At that time, cholesterol measurements were taken and found to be as follows:

Cholesterol Levels (mg/dl)	
TH	B6
107.98	37.73
97.55	28.53
89.57	53.99
99.23	51.53
102.31	
110.77	

Find the mean and standard deviation of the cholesterol levels for the TH mice.

Find the mean and standard deviation of the cholesterol levels for the B6 mice.

Create a box plot showing the cholesterol levels for each strain.

Submit a Word document with the solution, along with a R script file which generated your results.

Guided solution:

- Open R studio.
- Create a new R script file by choosing “File”, “New File”, and “R Script” from the menu. Note that a new panel opens in the top left of R Studio. This is an editor for your new file.
- In the file, add the following code which will assign the TH cholesterol levels to a variable called `th_chol`:

```
th_chol <- c(107.98, 97.55, 89.57, 99.23, 102.31, 110.77)
```

- If you look in the “Environment” tab, you will notice that there is nothing in the environment. Since we are working in the file editor, the code has not actually been run yet.
- On a new line in the editor, add the following code which will compute the mean of `th_chol`:

```
mean(th_chol)
```

- Add the following code which will compute the standard deviation of `th_chol`:

```
sd(th_chol)
```

- Click on (or move the cursor to) the first line of code, and click the “Run” button at the top of the editor pane. This will run the line of code, and (if everything is correct), you will see `th_chol` appear in the environment. The code that was executed will appear in the console.
- If the cursor is not already on the second line of code, move it there and click “Run” again. This will execute the second line of code. The mean of `th_chol` should be displayed in the console.
- Execute the third line of code. The standard deviation of `th_chol` should be displayed in the console.
- Add the following line of code to assign the B6 cholesterol levels to a variable:

```
b6_chol <- c(37.73, 28.53, 53.99, 51.53)
```

Run this line of code.

- Add lines of code to compute the mean and standard deviation of `b6_chol`. Run these two lines of code.
- To create the box plot, add the following line of code to your script:

```
boxplot(th_chol, b6_chol)
```

- Run this last line of code. It will create a box plot, which will be displayed in the “Plots” tab in the bottom right of R Studio.
 - When you are happy the code is correct and doing what you want, save the file. Click in the editor pane, and from the menu choose “File” and “Save As...”. You should choose or create a folder to save the file in which will make it easy for you to organize your assignments (e.g. you might want to create a “BMR617 Assignments” folder in your “Documents” folder, or any other organization which makes sense to you). The file name should end “.R”, and should be something that will be easy for you to find (e.g. “IntroAssignment.R” or “Assignment1.R” or “2022-01-11_Assignment.R”).
 - Create a Word (or other word processing document) to present your results. Since this is the first assignment, a model solution is included at the end of this assignment.
 - To include the plot, in the “Plots” tab, choose “Export”, and “Copy to clipboard”. In the window that appears, choose “Copy Plot”. Then in Word, you can choose “Edit” and “Paste” (or Control-V) to paste the plot into your results document. We will see better ways to create and save plots later in the course.
-

Discussion:

The entities in R that are followed immediately by parentheses are *functions* (in this case they are built-in R functions):

```
c (...)
mean (...)
sd (...)
boxplot (...)
```

In general, functions take values (specified between the parentheses) and process them, possibly generating another value.

The `c(...)` function means “combine”. It combines values into a single variable. So

```
th_chol <- c(107.98, 97.55, 89.57, 99.23, 102.31, 110.77)
```

combines the six values into a single “vector” of values.

Similarly, `mean(...)` takes a collection of values, and evaluates to the mean of those values, and `sd(...)` evaluates to the standard deviation of those values. The `boxplot(...)` function takes some data and processes it to generate a plot.

By contrast, `th_chol` and `b6_chol` are *variable names* that we have chosen. You can think of these as boxes that contain some kind of values. In R, each variable typically contains a collection of values. You can see all the variables you have defined in R Studio’s “Environment” tab.

The `<-` operator is “assignment”. It takes the value on the right hand side, and assigns it to the variable on the left hand side. So the line

```
th_chol <- c(107.98, 97.55, 89.57, 99.23, 102.31, 110.77)
```

combines the six values into a vector, and puts the result into the variable called `th_chol`.

What do you think the effect of

```
th_mean_chol <- mean(th_chol)
```

would be?

For now, we have just typed the data values directly into R (you can also copy and paste from this document). Fairly soon in the course, we’ll see how to load data easily from files that can be created with Excel. This will eliminate the potential for errors in data entry.

There are advantages and disadvantages to using R. One disadvantage is that it is a steep learning curve. We need to know about the functions we are using, and this takes some practice and experience. As you’ll see soon, if you haven’t already, if you make mistakes the error messages can be difficult to interpret.

On the other hand, the benefits are big. The R Script file we have is a complete record of everything we did to get from the raw data to the summary data (means and standard deviations) and the graphics. This doesn’t guarantee that everything we did is correct, but it makes it possible for anyone (including ourselves) to know exactly what analysis we performed and to evaluate its validity. This makes the analysis completely reproducible and provides a mechanism for a high level of scientific integrity in our data.

Model Solution:

Six TALLYHO ("TH") mice and four C57/CL6 ("B6") mice were fed a standard chow diet from weaning until 16 weeks of age. At that time, cholesterol levels were measured for each mouse (see figure 1). The mean cholesterol level for TH mice was 101.2 mg/dl (standard deviation 7.6 mg/dl). The mean cholesterol level for B6 mice was 42.9 mg/dl (standard deviation 12.0 mg/dl).

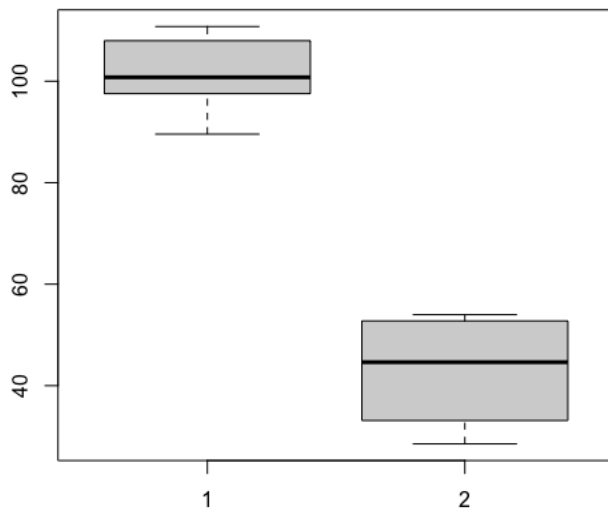


Figure 1. Cholesterol levels for TH (left) and B6 (right) mice fed a Chow diet.