--- Chapter Three ---

Preparing Biological Data For Statistical Analysis Using R

The first step in any data analysis is to import your data into R, check them for errors, organise them by dividing them into subsets or by joining information from different sources together, and summarise them to help you understand what your data set contains. These basic actions not only allow you to get to know your data better, they can also help you to identify any issues there might be with them. This means the time you spend doing this will be more than repaid by the time saved trying to solve problems you may encounter later on if you do not. Thus, in this chapter, you will learn about the various ways you can import a data set into R, and then how you can check it for errors, divide a data set into subsets, join different data sets together and summarise the information a data set contains.

Before you start the exercises in this chapter, you first need to create a WORKING DIRECTORY folder on your computer and load the necessary data into it. To do this on a computer with a Windows operating system, open Windows Explorer and navigate to the location where you would like to create the folder (such as your C:\ drive or your DOCUMENTS folder). Next, right click anywhere in this location and select NEW> FOLDER. Now call this folder STATS_FOR_BIOLOGISTS_ONE by typing this into the folder name section to replace what it is currently called (which will most likely be NEW FOLDER). To create a WORKNG DIRECTORY folder on a computer running a Mac operating system, open Finder and navigate to the location where you would like to create the folder (such as your DOCUMENTS folder or your DESKTOP). Next, click on FILE> NEW FOLDER, and then type the name STATS_FOR_BIOLOGISTS_ONE before pressing the ENTER key on your keyboard.

Once you have created your WORKING DIRECTORY folder, you are ready to download the data sets you will use for the exercises in this workbook from *www.gisinecology.com/stats-for-biologists-1*. After you have downloaded the compressed folder containing the required data by following the instructions provided on that page, you need to extract all the data files

from it and copy them into the folder called STATS_FOR_BIOLOGISTS_ONE that you have just created.

Next, you need to check that the required data have been extracted to the correct folder. If you are using a computer with a Windows operating system, you can use Windows Explorer to open your newly created WORKING DIRECTORY folder and examine its contents. If all the files from the compressed folder are present in it (there should be a total of 21 of them), you can click on the folder icon at the left hand end of the ADDRESS BAR at the top of the WINDOWS EXPLORER window to reveal its full address. Write this address down as you will need it to set this folder as your WORKING DIRECTORY during the exercises provided in this workbook (see pages 12 and 13 for details of how to modify folder addresses so they will be recognised by R).

If you are using a computer with a Mac operating system, you can use Finder to open your newly created WORKING DIRECTORY folder and examine its contents. If all the required data files are present in it (there should be a total of 21 of them), select this folder in Finder and then press the CMD and I keys on your keyboard at the same time. This will open the GET INFO window where you will find its address (which is also called the pathway). Write this address down somewhere as you will need it to set this folder as your WORKING DIRECTORY during the exercises provided in this workbook (see pages 12 and 13 for details of how to modify folder addresses so they will be recognised by R).

After you have loaded the required data into your WORKING DIRECTORY folder, you can open RGUI or RStudio, depending on which option you wish to use (see Chapter 2 for more details). Once you have opened your preferred R user interface, you need to create a file called CHAPTER_THREE_EXERCISES where you will save the results of your analyses from your R CONSOLE window as you work through this chapter. To do this using RGUI, click on the FILE menu and select SAVE WORKSPACE. To do this in RStudio, click on SESSION and select SAVE WORKSPACE AS. In both cases, save it as a WORKSPACE file with the name CHAPTER_THREE_EXERCISES.RDATA in your WORKING DIRECTORY folder (this will be the one called STATS_FOR_ BIOLOGISTS_ONE that you have just created). If you are using RStudio, you will also want to save the contents of your SCRIPT EDITOR window (where you will enter and edit the R code you will use to carry out specific commands). To do this, click on the FILE

menu and select SAVE AS. Save your file as an R SCRIPT file with the name CHAPTER_THREE_EXERCISES.R in your WORKING DIRECTORY folder. As you work through the exercises in this chapter, remember to regularly save the contents of your R CONSOLE window (which will contain the R objects you have created up to that point) to your WORKSPACE file and, if you are using RStudio, the contents of your SCRIPT EDITOR window to your R SCRIPT file.

Finally, you need to remove any data that are currently held in R's temporary memory. To do this, enter the following command into R (if you wish to copy and paste this command, the required code is directly below the text CODE BLOCK 1 in the document called R_CODE_BASIC_STATS_WORKBOOK.DOC that is included in the compressed folder you just downloaded):

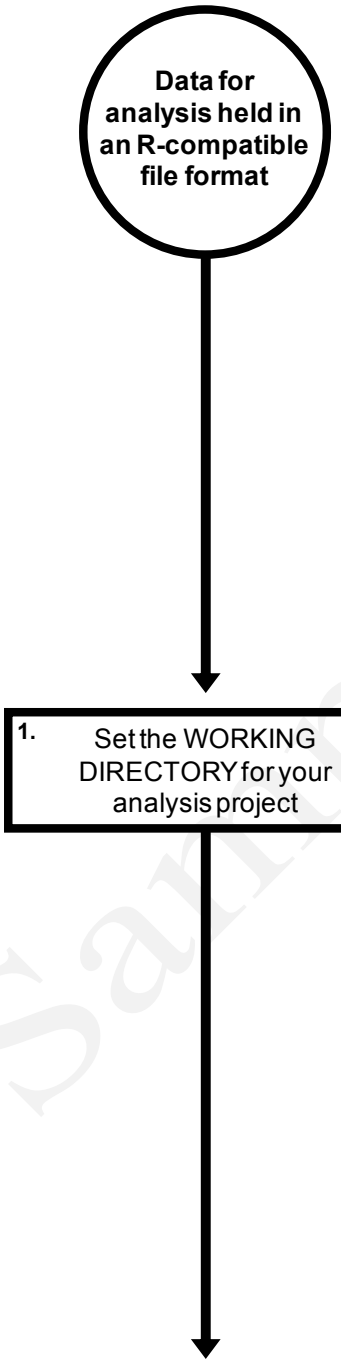
```
rm(list=ls())
```

If you are using RGUI, you can simply type or paste this code after the command prompt at the bottom of the R CONSOLE window (it looks like this: >) and then press the ENTER key on your keyboard to run it. If you are using RStudio, you can type or paste this command into the SCRIPT EDITOR window (the upper left hand window). To run this command, select it and then click on the RUN button at the top of this window. This will run it in the R CONSOLE window (the lower left hand one in the main RStudio user interface). You are now ready to start the exercises in this chapter.

EXERCISE 1.1: HOW TO IMPORT DATA INTO R:

Data can be imported into R from a variety of different file formats. This includes comma separated value (.CSV) files, tab delimited (.TXT) files, data copied to your operating system's clipboard and spreadsheet files. In this exercise, you will learn how import data into R from each of these different file formats, starting with the comma separate value file format. A .CSV file is a file where the data held in different columns are separated by a comma (,) or, when a computer's operating system is set to use commas as the decimal separators, by a semicolon (;). It is one of the most common file formats used to transfer biological data from one program to another, including importing data from spreadsheet software (such as Microsoft Excel or OpenOffice Calc) into R.

The comma separated value file you will import into R in the first part of this exercise is called `blue_tit_occupancy_data.csv`. It contains data on the occupancy of nest boxes by blue tits, a small hole-nesting bird species, along with information on the location of each box and the land elevation at that location. To import the data from this .CSV file into R, work through the following flow diagram:



For this example, the data set you will import into R is stored in a file called `blue_tit_occupancy_data.csv` that is located in the WORKING DIRECTORY folder you created during the introduction to this chapter.

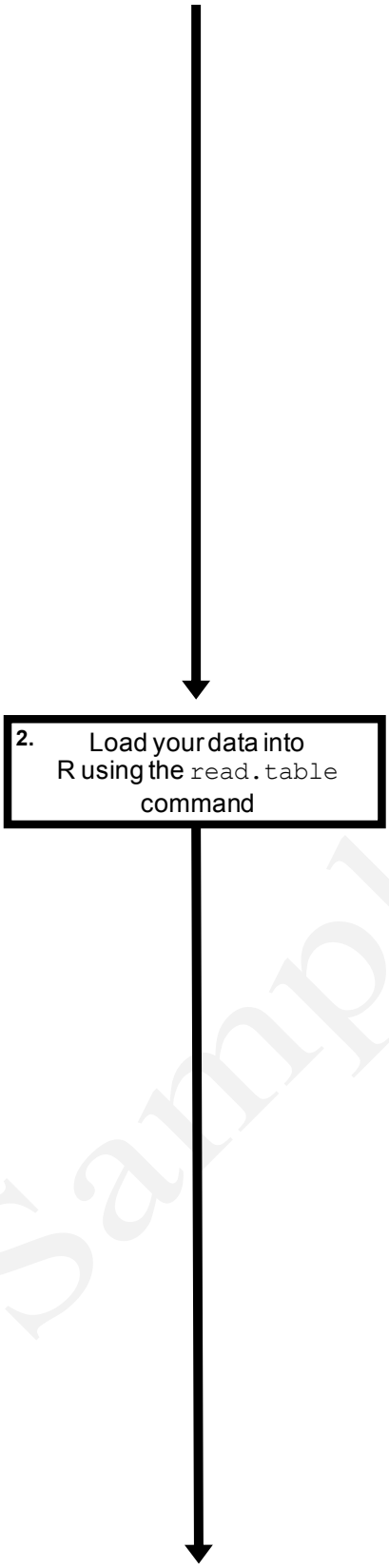
Before you start any analysis in R, you first need to set the WORKING DIRECTORY. To do this, enter the text `setwd("` and then type the address of your WORKING DIRECTORY, using slashes (/) as the folder separators, before entering a second quotation mark followed by a closing bracket, like this `")`. For example, if your WORKING DIRECTORY has the address C:\STATS_FOR_BIOLOGISTS_ONE, your `setwd` command should look like this:

```
setwd("C:/STATS_FOR_BIOLOGISTS_ONE")
```

If you are using RGUI, enter your `setwd` command in the R CONSOLE window (remembering to use the address of your own WORKING DIRECTORY folder in it) and then press the ENTER key on your keyboard. If you are using RStudio, enter your `setwd` command into the SCRIPT EDITOR window. To run it, select it and then click on the RUN button at the top of this window. You will enter all the remaining commands for this exercise in a similar manner, depending on the user interface you are using.

To check that your WORKING DIRECTORY has been set properly, enter the command `getwd()` and carefully check that the address it returns is the same as the one for the STATS_FOR_BIOLOGISTS_ONE folder you created at the start of this chapter.

Before you move on to step 2, make sure that all the data you wish to use in your analysis project are located in this WORKING DIRECTORY folder. In this case, this is a file called `blue_tit_occupancy_data.csv`. **NOTE:** If the data you are going to import into R in step 2 are not located in the WORKING DIRECTORY you set in this step, the import code provided in the next step will not work.

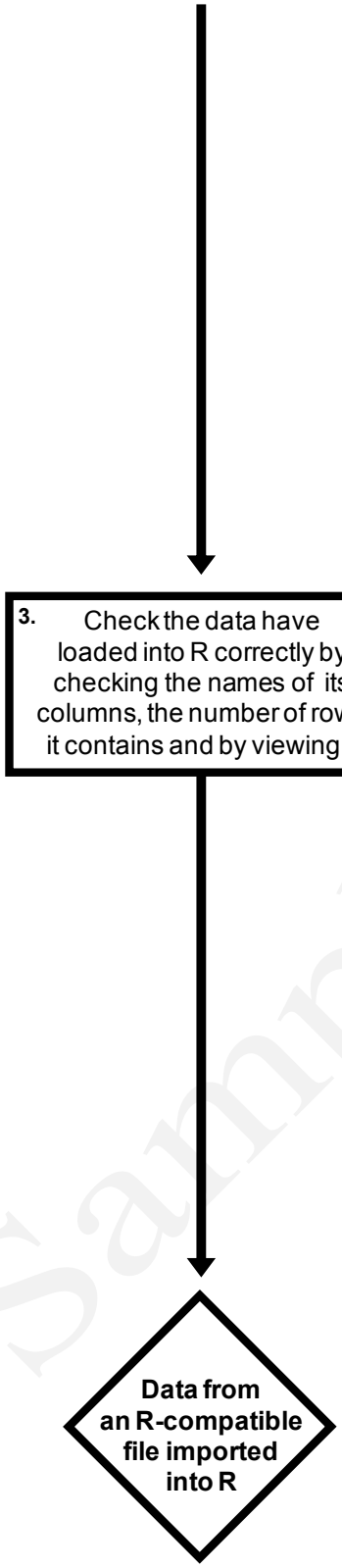


The `read.table` command provides the easiest way to load data held in a .CSV file into R so you can analyse it. To do this for the data being used in this example, enter the following command into R:

```
blue_tit_data <- read.table(file=
"blue_tit_occupancy_data.csv", sep=",",
header=TRUE)
```

This code has to be entered exactly as it is written here or it will not work. If you wish to use the copy-and-paste approach for entering this command, copy the text directly below CODE BLOCK 2 in the document R_CODE_BASIC_STATS_WORKBOOK.DOC and paste it into R.

This command will create a new object in R called `blue_tit_data` which will contain the data from the specified .CSV file. To load a different .CSV file into R, all you need to do is change the file name in the `file` argument to the name of the one you wish to import. However, the specified file must be located in the WORKING DIRECTORY you set in step 1 of this flow diagram. You can also use whatever name you wish for the R object that will be created by this command. To do this, simply replace `blue_tit_data` at the start of the above command with the name you wish to use for it. **NOTE:** If your .CSV data set uses a semi-colon as the decimal separator, you would need to replace the `sep=","` argument with `sep=";"`.



Whenever you import a data set into R, you need to check that it has been loaded correctly. First, you need to check that all the required columns are present in the R object you just created. To do this, enter the following command into R:

```
names(blue_tit_data)
```

This is CODE BLOCK 3 in the document R_CODE_BASIC_STATS_WORKBOOK.DOC. This command will return the names used for each column in the R object called `blue_tit_data` created in step 2. For this example, the names should be: `box_number`, `latitude`, `longitude`, `occupied`, `elevation` and `el_cat`.

Next, you should check the number of rows in your R object to make sure that the entire data set has been successfully imported into R. To do this, you need to specify one of the columns within your newly created R object. For this example, you will use the column called `box_number` in the R object called `blue_tit_data`. To count the number of rows in this column in this R object, enter the following command into R:

```
length(blue_tit_data$box_number)
```

This is CODE BLOCK 4 in the document R_CODE_BASIC_STATS_WORKBOOK.DOC. For the data set being used in this example, the number of rows this command returns should be 198.

Finally, you should view the contents of your newly created R object (called `blue_tit_data` in this example) using the `View` command (**NOTE:** Unlike most commands in R, this command begins with a capital letter). This is done by entering the following code into R:

```
View(blue_tit_data)
```

This is CODE BLOCK 5 in the document R_CODE_BASIC_STATS_WORKBOOK.DOC. This command will open a DATA VIEWER window where you can examine your data set and check that the correct data have been loaded into R.

At the end of the first part of this exercise, the last few lines of your R CONSOLE window should look like this (**NOTE:** Your WORKING DRECTORY folder will have a different address to the one shown here if it has been created in a different location on your computer):

```
> setwd("C:/STATS_FOR_BIOLOGISTS_ONE")
> getwd()
[1] "C:/STATS_FOR_BIOLOGISTS_ONE"
> blue_tit_data <- read.table(file="blue_tit_occupancy_data.csv", sep=",", header=TRUE)
> names(blue_tit_data)
[1] "box_number" "latitude"   "longitude"  "occupied"   "elevation"
[6] "el_cat"
> length(blue_tit_data$box_number)
[1] 198
> View(blue_tit_data)
>
```

While the contents of the DATA VIEWER window should look like this:

	box_number	latitude	longitude	occupied	elevation	el_cat
1	138	56.12644	-4.617821	0	23.27828	20 to 30
2	139	56.12641	-4.618161	0	21.50202	20 to 30
3	141	56.12622	-4.617965	0	20.90909	20 to 30
4	144	56.12598	-4.616986	0	10.00000	0 to 10
5	143	56.12607	-4.616885	0	15.00000	10 to 20
6	142	56.12622	-4.616998	1	20.00000	10 to 20
7	137	56.12642	-4.617008	0	21.58028	20 to 30
8	33	56.13024	-4.614165	0	30.82850	30 to 40
9	30	56.13059	-4.614695	0	31.48324	30 to 40
10	24	56.13129	-4.615251	1	22.87749	20 to 30
11	48	56.12906	-4.614811	1	25.00000	20 to 30
12	13	56.13031	-4.616208	0	42.50000	40 to 50
13	14	56.13052	-4.616450	0	40.00000	30 to 40
14	19	56.13106	-4.616647	0	32.61233	30 to 40
15	301	56.13181	-4.617249	1	32.92912	30 to 40
16	44	56.13033	-4.617238	1	60.00000	50 or more
17	4	56.12952	-4.615516	0	40.00000	30 to 40
18	3	56.12927	-4.615110	1	32.50000	30 to 40
19	300	56.13214	-4.616719	1	25.14756	20 to 30
20	302	56.13208	-4.617021	0	28.33333	20 to 30

While the `read.table` command works for most .CSV files, there may be some occasions where it does not work with a particular one. In such instances, you can modify the code used in step 2 of the above flow diagram to use the more specific `read.csv` command rather than the more generic `read.table` command. The modified version of this command would look like the one provided at the top of the next page (required modifications highlighted in **bold**).