Introduction to R

Berti E, Mata JC

8 January 2019

What is R and why we use it?

R is a programming language that is compiled (translated to machine readable code) on the run, as you type command into the terminal. This has several advantanges, for instance R is *easier* to learn compared to other compiled languages (C, C++, etc…), and code written in R can be easily and *quickly* modified. However, as in everything else, starting learning a programming language is always a challenging and often frustrating activity. So, why do we do it? The quick answer is because we can perform a huge amount of calculations in short time and in an automatic way. For example, let's say you want to know the mean of one million measures. If you want to evaluate it by hand, and assuming you are very fast at using a pocket calculator, it will take you a bit more than one week to have the answer. By using a common personal computer it will take you around 0.001 seconds. Pretty convinient. The reason why we use specifically R is that it was developed especially for statistical analysis, and has one of the largest collection of packages and libraries that can answer your questions. Loosely speaking, this means that we may not completely understand how a neural network machine learning algorithm works, but, with the needed precautions, still use it to our advantage. But remember: "*with great power comes great responsibility*".

First program: exploring a dataset and first calculations

Import a dataset.csv

The first task you may want to perform is to import a dataset that you want to analyze. First, let's navigate to the folder where the dataset is located with the command **setwd()**, which *set* the *w*orking *d*irectory to the directory specified by *dir*.

```r
setwd(dir = '/home/GIT/BEHAVIOURAL-BIOLOGY-2019/Rintro')
```

NB: R was not developed by a Dane with the scope to be used only by Danes. Danish characters may or may not work, but we strongly recommend not to use them. In general, don't use any character not included in standard ASCII tables. Always think that you are writing also for others, who may want to use your code. If you write with characters not available in most of international keyboards, you are already limiting the applicability of your code.

Now we are ready to import the dataset that we want to analyze. In this case, we want to import the dataset of the body masses of all species belonging to the *Canidae* and *Felidae* families, which is called *body-mass-comparison.csv*. We can import it using the function **read.csv()**, which takes two arguments, the name of the file to import (*file*), and a Boolean variable (*header*, either 0 or 1, *T*rue or *F*alse) that specifies if the file has a header (line with columns' names) or not.

```r
dataset <- read.csv(file = 'body-mass-comparison.csv', header = T)
```

Explore the dataset

When we type *dataset* and press Enter, the content of the **variable** named *dataset* is displayed on the screen. To have a look only at the first six lines of it we can use the command **head()**.

```
head(dataset)
```

```
##              Binomial  Family  Mass.g
## 1    Acinonyx_jubatus Felidae 46700.0
## 2 Atelocynus_microtis Canidae  7750.0
## 3       Canis_adustus Canidae 10249.9
## 4        Canis_aureus Canidae 10345.2
## 5         Canis_dirus Canidae 64000.0
## 6       Canis_latrans Canidae 13406.3
```

This dataset consits of three columns, the first with the name of the species (*Binomial*), the second with the family of the species (*Family*), and the third with the body mass of the species, in grams (*Mass.g*). We can select a single column of the dataset by using the *$* symbol.

```
head(dataset$Mass.g)
```

```
## [1] 46700.0  7750.0 10249.9 10345.2 64000.0 13406.3
```

Subsetting and first calculations

Sometimes, we need only a subset of a given dataset. For example, here we may want to display only the species belonging to the *Canidae* family. In R, a simple way to do this is to use the function **subset()**, which takes two argument, the dataset that we want to extract information from (x), and a condition based on which we want to subset it (in this case Family should be equal to 'Canidae').

```
only_canidae <- subset(x = dataset, Family == 'Canidae')
head(only_canidae)
```

```
##              Binomial  Family  Mass.g
## 2 Atelocynus_microtis Canidae  7750.0
## 3       Canis_adustus Canidae 10249.9
## 4        Canis_aureus Canidae 10345.2
## 5         Canis_dirus Canidae 64000.0
## 6       Canis_latrans Canidae 13406.3
## 7         Canis_lupus Canidae 32183.3
```

NB: the correct syntax in the *subset* function is a double '==' symbol, and not '='. This is because '=' is an assignement operator, while '==' is a relational operator. For example, if we want to assign the value 5 to variable x, the correct way to do it is by typing $x = 5$. If then we want to change it to 6, we would type $x = 6$, and not $x == 6$, since this will perform a comparison between x and 6. Because we assigned 5 to x, the comparison will return the value *FALSE*, meaning that x is not 6. In general, '=' stands for '… is equal to …', whereas '==' stands for 'is … equal to …?'.

Exercise 1

Write a code that calculate the average body mass for the *Canidae* family using the function **mean()**.

```
only_canidae <- subset(dataset, Family == 'Canidae')
avg_can <- mean(only_canidae$Mass.g)
avg_can
```

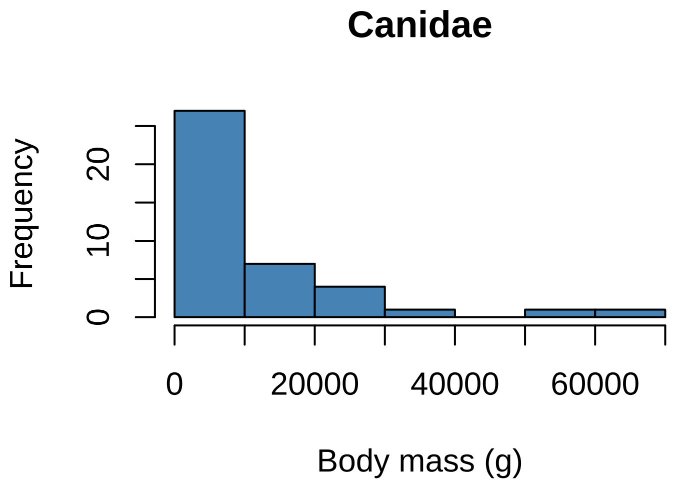
```
## [1] 10973.62
```

Now, this was quick, wasn't it? Let's continue by calculating the median of the body mass for *Canidae* with the function **median()**, and finally display its histogram using the function **hist()**.

```r
median_can <- median(only_canidae$Mass.g)
median_can
```

```
## [1] 5350
```

```r
hist(only_canidae$Mass.g, col = 'steelblue', main = 'Canidae', xlab = 'Body mass (g)')
```

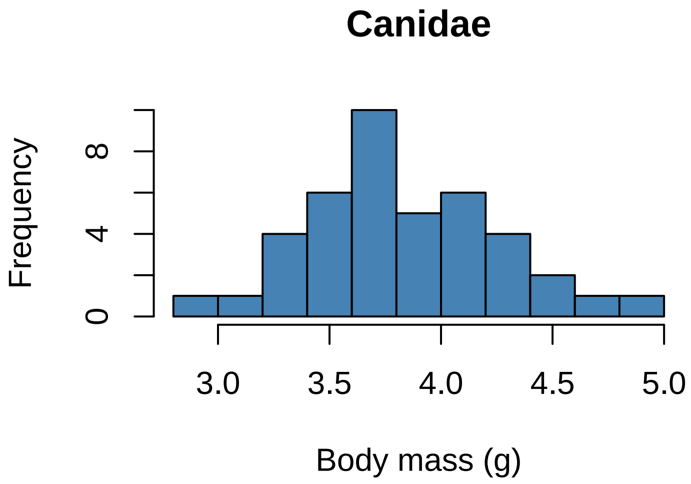


Questions:

- How would you interpret the average and median for this family?
- Would you say that body mass is normally distributed in this case? If not, any ideas?

Hint: there are a lot of small species and few very big. When you have this kind of "unbalanced" distributions try to plot in a log-scale using the **log10()** function.

```r
hist(log10(only_canidae$Mass.g),
     col = 'steelblue', main = 'Canidae', xlab = 'Body mass (g)')
```



- Is it now normally distributed?

Exercise 2 - optional

Evaluate the mean and median body mass for species belonging to the *Felidae* family, and plot the body mass distribution. What are the differences between *Felidae* and *Canidae*?

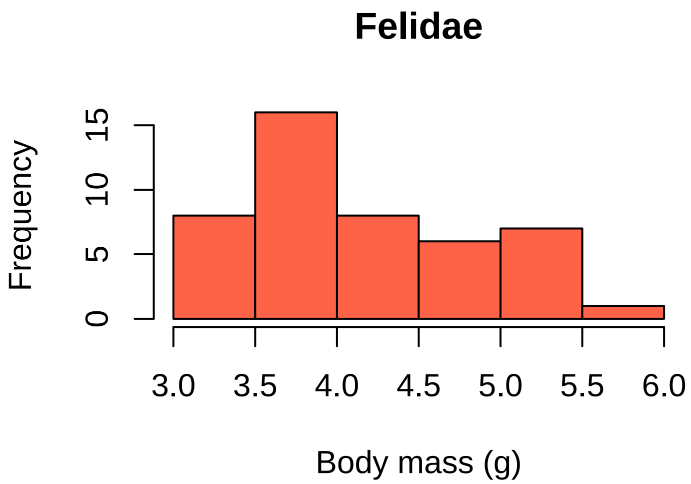
```
only_felidae <- subset(dataset, Family == 'Felidae')
avg_fel <- mean(only_felidae$Mass.g)
avg_fel
```

```
## [1] 52902.56
```

```
median_fel <- median(only_felidae$Mass.g)
median_fel
```

```
## [1] 9386.6
```

```
hist(log10(only_felidae$Mass.g), col = 'tomato', main = 'Felidae', xlab = 'Body mass (g)')
```


Felidae
Frequency
Body mass (g)