Dataset (s) used###

File: "pig_adg" originally from Bernardo TM, Dohoo IR, Donald A. Can J Vet Res. 1990 Apr;54(2):278–284. Growth rates, measures of ascarid burden, and the levels of anteroventral pneumonia and atrophic rhinitis at slaughter were determined for 352 hogs born between March 8 and March 28, 1987 on 15 farms located in Prince Edward Island. Regression analyses were used to determine associations between average daily gain (ADG) and independent variables controlling for sex, farm, and litters nested within farm.

Your tasks###

1. *Some descriptive functions in R*

R is an object orientated programming language. This means that everything in R is an object. Some are simple collections of numbers whilst others are more sophisticated. R contains a large collection of functions which operate on data frames and other objects and which often produce new objects in return.

Table2: Some useful commands to explore your dataset in R.

mean	Mean
median	Median
var	Variance
sd	Standard deviation
max	Maximum value
min	Minimum value
sum	Sum
rank	Rank
summary	Summarize data
tapply	Used to apply a function to each group of components of the first argument defined by the levels of the second component (see below for example)

Set the working directory and open the "pig_adg.csv" file. Call the data frame "Ascar"

Check the structure of the data frame and that the variables are correctly recognised (factors as factors, numerical as numerical etc…) using function *str()*. If you need to transform a numerical variable into a factor, use the function *as.factor()*.

We will explore some of the useful functions offered by the "descr" package. Install it and load it for this session.

A quick exploration of a dataset should be done to obtain the summary statistics of all variables. This can be achieved in a single command *summary().* The function *summary()* is from the base library. From the "descr" package, a similar function can be found: *descr()*.

Try both on your dataset. What do they do?

2. *Describing continuous variables*

The term average refers to any one of several measures of the central tendency of a data set.

Identify a continuous variable and (with the help of table 2) calculate its mean and median.

There are a number of measures of the spread of the data, each of which has different attributes.

 Calculate the range, inter-quartile range, variance and standard deviation of your variable.

A frequency distribution shows the frequencies of occurrence of the observations in a data set

 Use the commands *hist()* and *boxplot()* on your variable. What do they show? If you are unsure, use the help functions in R. An interesting function from the "descr" package is *histkdnc()*. Try it out.

The assumption of Normality is central to many statistical procedures. The easiest approach to establishing Normality is to observe the Q-Q plot in which the horizontal represents the ordered numerical values of the variable, and the vertical axis represents the standardized normal deviates (you don't need to worry today about how these are estimated). If the data are Normally distributed, the points will follow the straight line.

 Follow this example to plot a QQ plot for your variable. FYI: "col" argument set a specific color in your graph. In this example we choose red but many options are available. You can check later on internet.

```
qqnorm(Ascar$adg)
qqline(Ascar$adg, col="red")
```

It is also possible to compare the distribution of a variable between two groups.

 The functions *aggregate()*, *tapply()* and *compmeans ()* (the later from the package "descr") allow you to do just that. Try them out!

```
compmeans(Ascar$adg,Ascar$sex)#"descr" package
aggregate(Ascar$adg, by=list(Ascar$sex,Ascar$lu), FUN="median")
tapply(Ascar$adg, Ascar$lu, FUN="range")
```

3. ***Describing categorical variables***

When a variable is categorical or qualitative, we are interested in the frequency of occurrence of the observations in every class or category of the variable.

 We can display this information in a table in which each class is represented, or in a diagram such as a bar chart or pie diagram. Play with the *table()*, *prop.table()*, *pie()* and *barplot()* functions.

HELP
freq() from the "descr" package offers a very interesting alternative to the above.

Cross tabulation is a statistical process that summarises categorical data to create a contingency table. They provide a basic picture of the interrelation between two variables and can help find interactions between them. The mosaic plot is a graphical representation of a contingency table.

 Create contingency tables using *crosstab()* and *CrossTable()* (from "descr" package).

Other useful resources###

Aviva Petrie & Paul Watson 2013: Statistics for Veterinary and Animal Science, 3rd Edition, Wiley-Blackwell. -> Chapter 2: Descriptive statistics.

R Programming/Descriptive Statistics:
http://en.wikibooks.org/wiki/R_Programming/Descriptive_Statistics

Graphics and Exploratory Data Analysis in R:
http://bio.fsu.edu/miller/docs/Tutorials/Tutorial3_Graphics.pdf