Computer Exercise 1 (R)

If you are new to R, it might be a good idea to skim through chapter 1 and chapter 2 of the R-book before starting the actual exercises. This will not take long, and helps you get into R.

Written commands are entered into the R console, an editor window (of your choice, e.g. note pad, or the built in editor) or into the script window in R studio. Using an editor is advisable as it gives you the chance to save, reload and alter your codes as you work.

1. First, using the File menu in the R console program (in Windows), make sure that the current directory is set to where you have put the data files and other files for the exercise. You can get your current working directory by typing:

```
getwd()
```

and set it by typing:

```
setwd("Your/path/here")
```

You can also set it by going through the menus in R studio. On Macs you can simply set the working directory by dragging the folder you want as a directory to the R symbol in the dock. Ask your teacher if you feel unsure how to go about setting the working directory correctly. Next, read the data in **Metabol.txt** into a so-called data frame in R. Commands for doing it is

R console
```
dat <- read.delim("Metabol.txt")
```

where `dat` is the name of the data frame (you can call it something else if you like) and the symbol combination <- indicates assignment in R. `read.delim` is a shortcut for reading a file that is tab delimited and has headers for the columns. You can check which variables are in the data frame using the command

R console
```
names(dat)
```

which will show you that there are two variables, Bweight and Mrate. You can look at the data in a spreadsheet using the command

R console
```
fix(dat)
```

The variables Bweight and Mrate give the body weights in kg and the resting metabolic rates in kcal/24 hr for 44 women.

To get a look at the data, plot histograms for Bweight and Mrate. To plot a histogram of Bweight, you can use the command

R console
```
hist(dat$Bweight)
```

where `dat$Bweight` is how you specify that you want the variable Bweight from the data frame `dat`.

To get some idea of the shape of a distribution, it can be nice to see a histogram together with an expected normal distribution. For your convenience, someone else has written some R code to accomplish this. To make the code available, give the command

```
source("NormFit.R")
```

assuming that the file NormFit.R is in the current directory. The code contains a function you can call like this

```
norm.fit(dat$Bweight)
```

to get a plot; do this also for Mrate. After having looked at the distributions, use R functions to compute mean, variance, standard deviation, skewness and standard error of skewness for the variables Bweight and Mrate. For the first three, you can e.g. use the commands `mean(Mrate)`, `var(Mrate)`, and `sd(Mrate)`, but there is no R function for skewness. To help you, there is code for this, which you can load by writing

```
source("SkewKurt.R")
```

(there is also code for computing kurtosis in the file). Calling the functions `skew(dat$Bweight)` and `skew.se(dat$Bweight)` gives the estimated skew and its standard error for the Bweight distribution; also check the Mrate variable. What do you note about the distribution of body weights? To read more about skew and kurtosis, look up pp. 350-353 in the R-book.

You might want to clean up your workspace before going on. Use the command `ls()` to see what you have in your work space, and the command `rm()` to clean up, e.g. `rm(dat)` to get rid of dat, `rm(list = ls())` too get rid of everything.

2. Read the data in **Aphids.txt** into a data frame (see previous exercise about how to do this). Each data point is the number of aphids observed on a clover plant (from 0 to 41 on a total of 424 plants; the observations are ranked). Your job is to form an opinion of whether these data follow a Poisson distribution. To get a visual impression, run the code in the file `AphidPoisFit.R` (which your teacher has written) using the command

```
source("AphidPoisFit.R")
```

The code plots a (relative frequency) histogram of the aphid counts, together with a Poisson expected distribution. Does the fit look good? What do you conclude, statistically and biologically about the distribution of aphids on plants? For the interpretation, it might help to look at the mean and variance of the variable. If the variance is greater than the mean the distribution is called clumped and if it is smaller it is called repulsed. What kind of distribution are you dealing with? (Clumped or repulsed?)

3. Read the data in **Chimp.txt**. The variable Glycine gives milligrams of glycine per milligram of creatinine in the urine of 37 chimpanzees. To get a look at the data, plot a histogram for Glycine (you could use the function `norm.fit` (from exercise 1). Then find $\bar{y}, s^2, s$, skewness and standard error of skewness, using the functions skew and skew.se from exercise 1. What do you note about skewness? Further examine the fit to a normal distribution by running the Shapiro-Wilk's test, using the command

R console
`shapiro.test(dat$Glycine)`

What do you conclude?

4. Import the data in **WaterChem.txt**. Plot a histogram for SO4, perhaps using the function `norm.fit` from exercise 1, and then make a so-called normal Q-Q plot to check the distribution for normality (this kind of plot is explained on pp. 346-7 in the R Book). You can use the commands

R console
```
qqnorm(dat$SO4)
qqline(dat$SO4)
```

to get a plot. Then repeat the same procedure for the variable CL. Do also apply the Shapiro-Wilk's test for normality (from exercise 3) to SO4 and CL. What do you conclude? Finally determine the mean, skewness, kurtosis, sample standard deviation and 95% confidence interval for the mean for each of the two variables. To get skewness and kurtosis, use the functions skew() and kurtosis() in the file SkewKurt.R. To get a 95% confidence interval for SO4, an easy way is to give the command

R console
```
t.test(dat$SO4)
```

and the same can be done for CL (don't bother about the actual t-test performed, and its p-value).

5. Using the command

R console
```
x <- rnorm(25, mean=5, sd=1)
```

you can create a sample (with sample size n = 25) from a normal distribution with $\mu = 5$ and $\sigma = 1$. Plot a histogram of the variable x, in order to get a feeling for how normally distributed data can look (why not use the `norm.fit` function from exercise 1). At the same time, test for a fit to normal distribution using the Shapiro-Wilk's test. Write down whether the test rejected the null-hypothesis of a normal distribution. Then find the standard error, using the command

```
sd(x)/sqrt(25)
```

and the 95% confidence interval of the mean of x (see exercise 4). Compare the standard error with the true standard deviation of the sample average (equal to what?). Repeat this (creating a new variable x with the command above) 20 times, and write down how many of the confidence intervals that contain the true mean ($\mu = 5$).

6. A *p*-value represents the proportion that is in one or both tails of a distribution beyond a critical value of a statistic. In this exercise you will learn how to compute critical values from *p*-values, and the other way around, both for normal and *t*-distributions. For a standard normal distribution, compute the two-tailed critical value corresponding to 5% and 1%, and the proportion outside ±1 and ±2, using the functions `pnorm` and `qnorm`. For instance, the command

```
qnorm(0.975)
```

gives the two-tailed 5% critical value (there is 2.5% probability to be in the right tail), and the command

```
2*pnorm(-1)
```

gives the proportion outside ±1. Do the same for a *t*-distribution with 15 degrees of freedom, using commands like

```
qt(0.975, 15)
```

and

```
2*pt(-1, 15)
```

You should check with your teacher if you find it hard to understand what these functions do.

Hints
To go from t to p, do this: 2 * (1 - pt (t))
To go from p to t, do this: qt (1 - p / 2)
The "1 –" is because the distribution is cut off on the right hand side and everything to the left is counted. The "/ 2" or "* 2" is because the cut-off of the distribution is done on one side only. Look at the graphs of the distribution in Lecture 2 for a graphic explanation. Look at sections 7.2 – 7.3.2 on pp. 271 - 281 for more information on the normal distribution and section 8.4.3 on p. 358 for more details on the t-distribution. All in the R Book.