# Home work 1

## Homework 1 - to be done as groups

Names:

Group:

For deadlines etc, see absalon.

You have to supply both the answer (whatever it is: numbers, a table, plots or combinations thereof), as well as the R or Linux code you used to make the plots. This should be done using this R markdown template: we want both the R markdown file and a resulting PDF. For PDF output, you may have to install some extra programs - R studio will tell you.

Note that:

1. If the R code gives different results than your results, you will get severe point reductions or even 0 points for the exercise
2. Some questions may request you to use R options we have not covered explicitly in the course: this is part of the challenge
3. While this is a group work, we expect that everyone in the group will have understood the group solution: similar or harder question might show up in the individual homework. So, if something is hard, it means you need to spend more time on it
4. The results should be presented on a level of detail that someone else could replicate the analysis.

For statistical tests, you have to:

1) Motivate the choice of test
2) State exactly what the null hypothesis is (depends on test!)
3) Comment the outcome: do you reject the null hypothesis or not, and what does this mean for the actual question we wanted to answer (interpretation)?

### Question 1

Install the package babynames and look at the data babynames:

```
install.packages("babynames")
```

```
library(babynames)
head(babynames)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

a) List the top 5 female baby names starting with P, regardless of year, as a table.

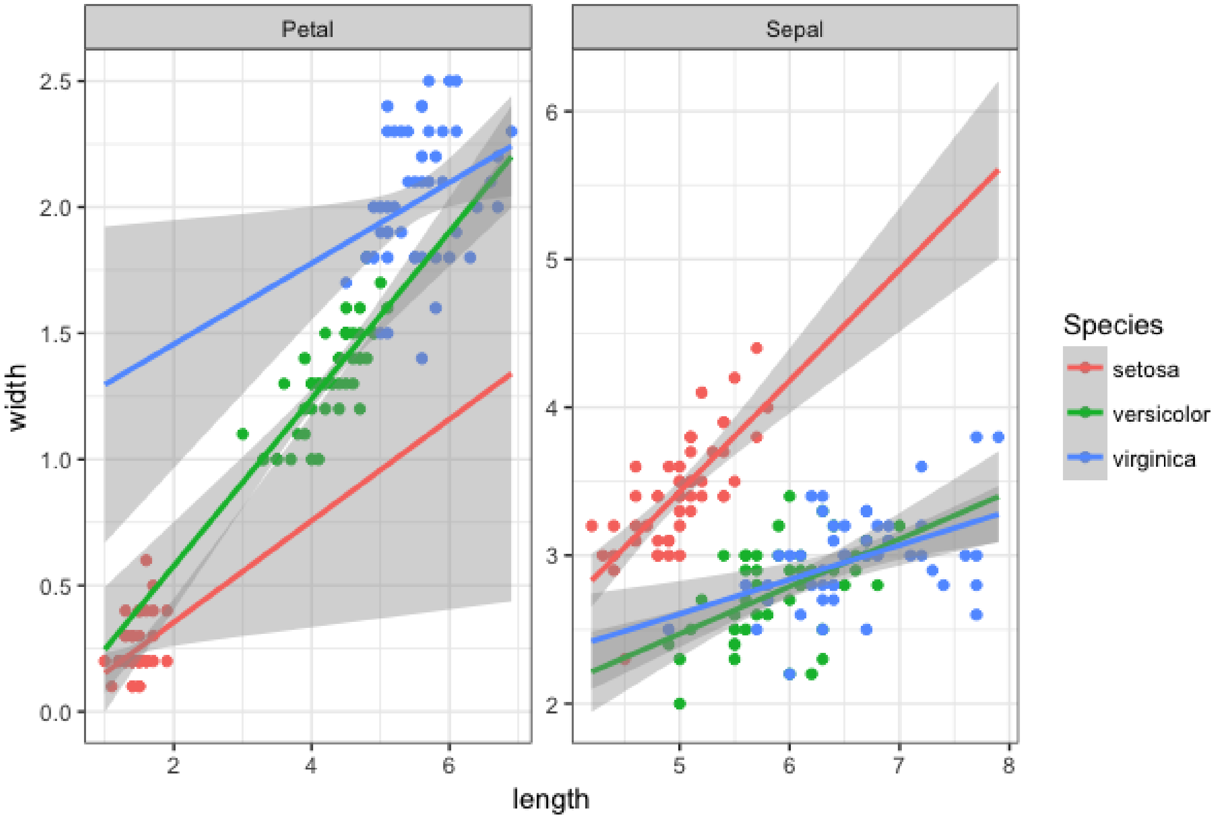
b) Using the results from a, plot their occurrences as a function of year using a line plot. Comment on your results. If you get strange results, explain them and/or improve the plot.

## Question 2

In the same dataset, is the name Arwen significantly more (or less) common in 2004 vs 1990? Is the change significant? What is the likely cause? Do not use hard-coding.

## Question 3

Produce the following plot starting from the flowers dataset. A potentially useful function that you may not have seen: bind_rows(): merges two tibbles by rows so that the joint tibble becomes longer, not wider



## Question 4

We are given a file with binding sites of a certain transcription factor, made with the ChIP-seq technique (you will hear a lot more about the technique later in the course) by a collaborator. In the homework directory, there is a data file 'chip_mm5.txt' from the collaborator, representing binding sites from a Chip-chip experiment, with a column for chromosome, start, end, and score, where score is how 'good' the binding is. Our collaborator has two hypotheses:

1: Binding scores are dependent on chromosome

2: Binding site widths (end-start) are dependent on chromosome

Can you prove/disprove these two hypotheses statistically?