Cognition, Language and Communication 2013.
Computerlab 2, Assignment 3

Analysing data with R

The goals of today's computer lab are (i) to learn how to analyse data coming from experiments similar to the one you participated in last week and (ii) to introduce you to R, a programming language that is very suitable for statistical analysis.

1. The data

In the website of the course (http://www.illc.uva.nl/LaCo/clas/clc13/assignments) you will find the data resulting from the experiment you participated in the last computer lab. Download both files.

These files are in csv format. They can be read as a spreadsheet or as text files. There is a file for each condition (2 minutes and 10 minutes exposure). Each row in the file corresponds to the responses of one participant. Columns that have as title some sequences in capital letters correspond to questions in the experiments about that particular sequences. At the right of each of these columns there is an extra column that indicates whether the former column corresponds to a word (w), a partword (pw), or a foil (f).

As you may remember, in the experiment you listened to a nonsense stream of speech syllables, and then you were asked if you had heard some sequences of syllables. We call “words” those sequences that you actually heard in the stream and that have a high transitional probability (TP) between their syllables. “Partwords” are sequences that appeared in the stream but with have low TP between their syllables. “Foils” are sequences that include syllables that have not appeared in the stream.

2. Using R

You can use the operation system that you prefer for this session. R is probably already installed in the computers in the lab. You can also download Rstudio (http://www.rstudio.com/ide/download/desktop), which provides a comfortable user interface.

On the website of the course you will find a tutorial on R. You can find there basic instructions and a reference for the statistical tests and the plots we will use.

3. Analysing the data with R

First load the data into R:

```
table2min <-  read.csv (file.choose(), sep=”,”)
table10min <- read.csv (file.choose(), sep=”,”)
```

This will put the results of calling the function read.csv() into tables called table2min and table10min. Check the contents of these tables by typing table2min and table10min in the console.

To begin with, you can compute the mean of each column and store them in new vectors means.10m.w, means.10m.pw, means.2m.w, means.2m.pw. To do so, you can make use of the built-in functions c(), which creates a vector, mean(), which computes the mean, and the construct $LABEL that will give you the contents of the column labeled LABEL in your data table. Type out the following command and make sure that you understand what all the components mean.

```
means.10m.w <- c(mean(table10min$PU), mean(table10min$NA.), mean(table10min$KA),
mean(table10min$TANA), mean(table10min$LIKA), mean(table10min$PULIKI),
mean(table10min$TADOSU), mean(table10min$BENAKA), mean(table10min$SUPUDOKI),
mean(table10min$LIKATADO), mean(table10min$PUNAKIBE))
```

(Note: One of the columns of our data is called “NA”. Unfortunately, R reserves this as a key word that stands for “No Answer”. To avoid confusion, R will automatically rename the column as “NA.” (it adds a final dot). So when retrieving this particular column, you will have to use “table2min$NA.” .)

Question 1: Complete the commands missing to compute the means for both conditions and for different types of sequences.

Now that you have all these data, you can visualize it with two plots (one for each condition) with the means for each type of sequence:

```
plot (means.2m.w, type="p", col="blue")
par(new=T)
plot (means.2m.pw, type="p", col="red", axes=F)
par(new=T)
plot (means.2m.f, type="p", col="green", axes=F)

...
```

You can save the plots by using the “export” button in RStudio, or using the command `savePlot()`.

Question 2: Complete the commands mising to generate a plot for the 10 minutes condition. Hand in the plot and the commands.

Now create plots with the mean of all sequences of each type:

```
totalmean.2min.w <- mean(means.2m.w)
...


boxplot (totalmean.2min.f, totalmean.2min.pw, totalmean.2min.w)
...
```

Question 3: Hand in the plot and the missing commands. Comment on the differences between both plots. What is the effect of increasing the exposure time?

Now compute a ratio of correct guesses (words) and incorrect guesses (partwords and foils), and another with partwords and foils. Since foils are sequences that have never appeared in the stream, this measure should reflect response biases of the participants.

```
ratio2m <- totalmean.2min.w / (totalmean.2min.pw + totalmean.2min.f)
ratio10m <- totalmean.10min.w / (totalmean.10min.pw + totalmean.10min.f)
```

```
ratio2mpwf <- totalmean.2min.pw / totalmean.2min.f
ratio10mpwf <- totalmean.10min.pw / totalmean.10min.f

boxplot(ratio2m, ratio2mpwf, ratio10m, ratio10mpwf)
```

Question 4: Hand in the missing commands and the plot. Comment on the plot.

Now we will compute a paired t-test on the responses to words in each condition:

```
t.test(means.10m.w, means.2m.w, paired=TRUE)
```

```
data:  means.10m.w and means.2m.w
t = 1.1405, df = 10, p-value = 0.2806
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2588657  0.8017950
sample estimates:
mean of the differences
              0.2714646
```

In this case the p-value is larger than 0.05; hence, there is no significant evidence that the two distributions differ.

Question 5: Apply the same analysis for partwords. Comment on the results.

Question 6: Use the Shapiro-Wilk test (as explained in the tutorial) to test if the means used in the previous exercise follow a normal distribution. Hand in the commands and explain the result.

Homework:

Hand in a document with answers to questions 1 to 6.

IMPORTANT! Remember that on Monday (30/09) you have to hand in on paper a thesis to be discussed in class.