***Statistics for Humanities Researchers – day 2***

**Read the following instructions with the Day2Exercise1.Rmd file open.**

**The exercises should have 2 outputs:**

* **This document with answers to the questions. Cut and paste any plot you’re asked to produce.**
* **The Rmd file with your code written in and “knitted”**

**Part 1 – Contrastive design – Some conceptual questions**

*Background:* Two researchers are studying the effect of height in elections. They both hypothesize that when there are two candidates, the taller one is more likely to be elected.

1. One researcher, Hidetoshi Hirata, chooses to analyse American gubernatorial elections. For each historical election, he compares the two frontrunners on height and their chances of success.
2. Another researcher, Diana Alekseeva, studies elections in the laboratory. She presents subjects with pairs of fictional candidates of varying height, and asks the subjects who they would vote for in a general election.

*Exercise 1.1)* Describe the advantages and shortcomings of the two designs relative to each other.

*Exercise 1.2)* What are the null and alternative hypotheses in the two studies?

*Background:* One of the researchers in question 1, Dr. Alekseeva, has found that in a direct run-off election, taller candidates are more likely to be elected (p = .034).

Exercise 1.3) How should you interpret the p-value that Dr. Alekseeva has reported (assuming α = .05)? State what it says about the null and about the alternative hypothesis.

**Part 2 – Let’s get to real data – R exercises**

*Background:* Previous research conducted in the USA has shown that people with depression tend to speak differently from comparison groups (see Pennebaker, *The secret life of pronouns*). In particular, they tend to use fewer words, fewer positive words, more negative words and more past tense. A Danish student has decided to replicate these findings in the Danish population. The data you have come from his research.

*Exercise 2.1)* Characterize the sample population in terms of i) diagnosis, ii) gender, and iii) education. Would you consider this population externally valid?

*Exercise 2.3)* Are the American results replicated? Write a short paragraph + plot summarizing the results for each of the 4 indexes: i) total words, ii) positive words, iii) negative words, iv) past tense. At the end discuss possible reasons for the replication or lack of replication.

*OPTIONAL Exercise 2.3)* The dataset contains other linguistic indexes, test whether they are different in depressed and comparison participants.

*OPTIONAL Exercise 2.4)* The great thing about lm() is that you can just add new predictors without changing the structure of the code. Have fun by adding gender and age to the model: lm(DV~IV1+IV2,Data).