Analysing Data in R

Assignment 3 (Final assignment)

The assignment should be done in an R script. Please create a new R script and complete the assignment inside it, adding comments that explain what you are doing. Then upload the script to the Kampus platform. Use tidyverse verbs and syntax AND/OR base R to complete the task. Use additional packages for statistical analysis if needed. Use the pipe in your code.

Choose one of the following **problems** and provide a full analysis.

In each case you can **choose a simple analysis** (enough to get a 5) **or a more complex one** for additional points (to get a 6).

The instructions only contain the general idea of what you have to do, the rest is up to you, including the way in which you want to analyse and visualize the data. In your analysis, make sure you explore the data visually, check the descriptive statistics and check all the assumptions before you proceed with the statistical analysis.

Problem 1.

In this study, the researchers wanted to use two types of Spanish trisyllabic words: those that have their main stress on the penultimate syllable (PU) and those whose main stress falls on the antepenultimate syllable (APU). The words were manipulated in such a way that the stress was either pronounced correctly (s – standard condition) or on a different syllable (d – deviant condition).

To give an example, the word *semana* 'week' is a PU word in which the primary stress falls on the second-to-last, i.e. penultimate syllable: *seMAna*. In the deviant condition, it is pronounced *SEmana*. The word *pájaro* 'bird', on the other hand, is an APU word pronounced *PÁjaro* in the standard condition and *paJAro* in the deviant condition.

In order to use these words in a correctness judgement task in which the participant hears a word and has to decide whether the word was pronounced correctly or not, the researchers have to first make sure that the words they use in each of the conditions are comparable. To simulate incorrect stress on the penultimate syllable, it has to be produced as if it were a penultimate stress word. Said otherwise, the stress has to be **marked** properly. The JA in *paJAro* has to be like the MA in *seMAna* in terms of key stress properties. To simulate incorrect stress on the antepenult, the SE of *SEmana* has to be the same as the *PÁ* in *PÁjaro*. Only then can we say that when a participant of the study fails to detect incorrect stress, it is because of something else than the physical properties of these syllables. If these parallel syllables are not the same, we could attribute participant behaviour to the fact that the stress was not correctly marked on the deviant, e.g. the syllable was too short or not loud enough.

The stress properties marking stress in Spanish are: stressed syllable **duration**, **F0** and **intensity** (loudness). Hence, there are 3 dependent variables.

Your goal is to explore **one of the two datasets** (Problem1_1 and Problem1_2, choose whichever) and:

- provide descriptive statistics
- visualize group differences (or similarities)
- formulate the hypotheses
- choose the right tests and check all assumptions
- provide a statistical analysis and decide whether you can reject the null hypothesis (Are these words suitable for the experiment?)
- write a short report on the results and visualize them using ggplot with embellishments

In the datasets you have one **grouping variable**: **Word_type** (APUs – antepenult standards like *PÁjaro,* Pud – penult deviants like *SEmana,* APUd – antepenult deviants like *paJAro,* Pus – penult standards like *seMAna).* Hint: we want APUd to be comparable to PUs and PUd to be comparable to APUs.

The other variables are numeric: **SS_F0** (stressed syllable F0), **SS_int** (stressed syllable intensity), **SS_dur** (stressed syllable duration).

Choose one of the databases for your analysis. Each of them compares different words.

*** Additional points:** If you want to gain additional points and a higher grade, combine the two databases and transform the data to get a 2x2 design and provide statistics. In this case you would have stress type (penult, antepenult) and correctness (standard, deviant) as separate factors.

What is the result? Is there an interaction? Document everything you're doing as described in the main task above. Visualize the results properly using an effect/interaction plot. Provide stats and post-hoc analyses.

Important! If you want to do the additional points task, you don't have to do the analysis described above. Your research questions will be different. Explore data, provide main descriptive stats, perform, visualize and report the analysis.

Problem 2.

In this study, the researchers were interested in looking at an electrophysiological response to auditory stimuli in an EEG experiment. The design was simple: participants were hearing the word *FÁbula* 'fable'/'story' pronounced in the standard condition (with antepenult stress, on the first syllable) most of the time, and they sometimes heard a deviant *fabuLÁ* with stress on the last syllable. The deviant stress was produced by manipulating the standard word. The F0 was lowered in the first syllable and increased in the last syllable. In this way, the hearer hears the stress on the last syllable. This experiment type is called a passive oddball paradigm because the participant is passive (does not have to respond, there is no task) and they hear an 'odd one out' from time to time. This, if detected by the brain, should produce an early negative component (MMN – mismatch negativity), which is a negative deflection in the EEG signal in response to the deviant compared to the standard (which is the baseline condition in this case).

In the dataset (Problem2_1), we have the following columns:

- **erpset** equivalent to subject
- **chlabel2** – electrode pool (3 levels: centrally placed electrodes, parietally placed electrodes, frontally placed electrodes)
- **binlabel** – our key fixed factor with standard and deviant stimulus as levels
- **value** – our dependent variable, the voltage in a given condition

In general, we are interested in whether the mean voltage (coded as '**value**') is significantly different between standards and deviants. For this simple analysis, you have to **leave out parietal electrodes**. Exclude them from the analysis. The second question is whether there is a difference in mean voltage between the electrode pools (central vs. frontal). In this second case, ignore the binlabel and also exclude parietal electrodes. Note that the MMN component is a negative one so negative numbers that you see in the data are normal.

For this data:

- explore the differences between the standards and deviants, and then electrode pools graphically
- provide descriptive statistics
- formulate the hypotheses
- choose the right tests and check all assumptions
- provide the analysis and comment on the results – Can we reject the null hypothesis?
- write a short report on the results and visualize them using ggplot with embellishments

*** Additional points:** to get additional points and a higher grade, take the other dataset: Problem2_2, in which you have repeated measures. If you can do a repeated measures Anova, that's wonderful and you can proceed. If not (which is what I would expect since we did not cover this), **convert this database** into one similar to Problem2_1 by averaging over electrodes in a pool (per condition and participant). Your code will tell me how you did it.

After that, you can perform a 2x3 design analysis in which you look at the difference between standards and deviants and whether this depends on the electrode pool. Add an interaction. Here, we expect the MMN component to be located in the front and center of the scalp. So we would not expect MMN in the parietal electrodes. Check this and visualize it properly using an effect/interaction plot. Provide stats and post-hoc analyses.

Important! If you want to do the additional points task, you don't have to do the analysis described above. Your research questions will be different. Explore data, provide main descriptive stats, perform, visualize and report the analysis.

General note: Your short report is just a short paragraph of text with numbers (2-3 sentences).