HW 3. An overview

Deadline: May 15, 23:59

Your name:

Part 1: frequent words, their acoustic duration and co-articulation effects (continued)

We continue using dataset from the study of Fabian Tomaschek et al. (2018), see HW2 for full details.

Variables of interest:

- `LogDurationW` - log-transformed word duration (i.e. logarithms of word duration).

- `LogDurationA` - log-transformed segment duration.

- `Cond` - condition (slow, fast).

- `Exponent` - inflectional exponent of verbs: `-t`, `-en`, `-n`. By default: `-t`.

- `Frequency` - log-transformed frequency of verbs in the corpus.

1.1 Confidence intervals

1.1.1 Explicit formula   Recall the formula for 95% confidence interval discussed at the lecture:

$$\text{CI} = \left[\bar{x} - 1.96 \times \frac{\text{sd}(x)}{\sqrt{n}}, \ \bar{x} + 1.96 \times \frac{\text{sd}(x)}{\sqrt{n}}\right].$$

Use it to find 95% confidence intervals for population means of word durations for fast and slow conditions. (You have to obtain two confidence intervals, one for fast condition and another for slow condition.)

```
# YOUR CODE HERE
```

1.1.2 Function `MeanCI`   Use function `MeanCI` from package `DescTools` to find the same confidence intervals.

```
# YOUR CODE HERE
```

(The results will be a little bit different compared to the result of the previous section due to the fact that the formula above is only approximation and `MeanCI` uses a more precise formula. However, for our data the difference is very small.)

1.1.3 Function `t.test`   You can also use function `t.test` for one sample to obtain the confidence interval for a mean. Apply `t.test` to the same variables as in 1.1.1 and extract the confidence intervals from the output. Does it coincide with the results of sections 1.1.1 or 1.1.2?

```
# YOUR CODE HERE
```

1.1.4 Different confidence level   Use function `MeanCI` to find 99% confidence intervals for the same variables as in 1.1.1. Are they wider or narrower than 95% CI's?

```
# YOUR CODE HERE
```

Hint: use `conf.level` option.

1.2 ANOVA

1.2.0 Data preparation   We will consider only observations obtained in `slow` conditions in this section, so you have to filter your dataset appropriately. Filter out only observations with `slow` condition and save resulting dataset as `dat_slow`.

```
# YOUR CODE HERE
```

1.2.1 Hypothesis   Let us look at three groups with Exponent -*en*, -*n*, -*t*, respectively.

We will use a one-way independent ANOVA to see whether there any statistically significant difference between segment durations in these three groups.

First of all, state the null hypothesis and the alternative you consider. (Note that the alternative hypothesis of ANOVA states that at least two groups are different.)

(your answer here)

1.2.2. `aov()`   Use the `summary` of `aov()` to perform an ANOVA.

```
# YOUR CODE HERE
```

1.2.3 Interpreting the ANOVA results   Let us use 3% significance level instead of usual 5% to make a decision. Can you conclude that there is significant difference (i.e. difference in the corresponding population means) between these three groups at 3% significance level?

(your answer here)

1.2.4 Pairwise t-test   Use two-sample t-test to compare segment duration for expontents *-t* and *-en*. Can you conclude, according to the results of t-test, that there is significant difference between these two groups at 3% significance level?

```
# YOUR CODE HERE
```

How can you explain a contradiction (if any) of this outcome with the outcome of 1.2.3?

(your answer here)

1.2.5 ANOVA: Word durations   Repeat 1.2.1, 1.2.2 and 1.2.3 for word durations instead of segment durations.

```
# YOUR CODE HERE
```

1.2.6 Post-hoc testing   Since an ANOVA provides only information on the overall significance of the differences between groups, you can perform an additional test (called post-hoc test in experimental studies) to find out which pairwise differences are significant. Perform the Tukey Honest Significant Difference test for the results of 1.2.3 and 1.2.5 using the `TukeyHSD()` function.

It requires an `aov()` object and a variable that defines the groups. For each pair of groups, the test provides the differences between the means (diff), their confidence intervals (lwr and upr), and the p-value of the difference. If p-value is less than the significance level, then the difference is significant.

As we are using 3% significance level, use 97% confidence interval by passing option `conf.level` to `TukeyHSD`. (By default, it's 95% confidence interval that corresponds to 5% significance level.)

Analyze p-values provided by the Tukey test to see if there is any pair of groups in which the difference is not significant at 3% significance level.

```
# YOUR CODE HERE
# TukeyHSD(aov(....), "Exponent")
```

1.2.7 Comparison with pairwise t-test   Compare p-value for the difference between segment durations for -*t* and -*en* Exponents, given by `TukeyHSD`, with the result of 1.2.4. How can you explain the difference between these p-values?

your answer here

1.2.8 Reporting the results   Write down you general conclusions for 1.2.1-1.2.7. Report the full name of the ANOVA (one-way independent ANOVA), its p-value, the name of post-hoc test (if used) as well as its results: p-value and the difference between the means (with CI) for those pair of groups that you want to report.

your answer here

1.3 Summary in a `dplyr` style

```
require(tidyverse)
```

Group you data by exponent and condition and calculate a mean, a median, and a standard deviation of word duration. You have to use `dplyr` (`tidyverse`) for these operations.

```
# YOUR CODE HERE
```

Part 2. Lexical decision

In this home assignment you are suggested to work with the data set with the results of a psycholinguistic experiment dedicated to lexical decision. In studies of lexical decision participants (also called subjects) are asked to decide whether the word shown on the screen is real or not. In other words, whether a word exists in the language or it is just an artificial word created using grammatical rules.Then the reaction time is measured: how fast a person clicks on the button *word* or *non-word*.

This data set is taken from the library `languageR`, it contains lexical decision latencies elicited from 21 subjects for 79 English concrete nouns, with variables linked to subject or word. Data collected by Jen Hay, University of Canterbury, Christchurch, New Zealand, 2004.

Some variables of interest:

- `Subject`: participant's id;

- `RT`: logarithmically transformed reaction times;

- `NativeLanguage`: a factor with levels `English` and `Other`, distinguishing between native and nonnative speakers of English;

- `Correct`: a factor with levels `correct` and `incorrect` coding whether the word was correctly responded to as a word rather than a nonword;

- `Word`: word shown;

- `Frequency`: logarithmically transformed lemma frequencies as available in the CELEX lexical database;

- `FamilySize`: log-transformed count of a word's morphological family members;

- `SynsetCount`: log-transformed count of synonym sets in WordNet in which the word is listed;

- `Length`: word's length in letters.

The description of all the variables in this data set can be found here.

For brevity, below we will refer to variable `RT` as "reaction time" despite the fact that it is actually the logarithm of time measured in ms.

2.1 Lexical decision: correctness and native language

Imagine that you are suggested to conduct a small research on lexical decision. And before proceeding to more substantial analysis you want to check whether the correctness of decision (`Correct`) depends on the person's native language (`NativeLanguage`).

2.1.0 Data loading   Load data (link) and look at the summary of the loaded data frame.

```
# YOUR CODE HERE
```

2.1.1 Descriptive statistics   How many correct answers were provided by native English speakers? And by speakers of other languages? Answer the same questions, but for incorrect answers. Provide your R code used to answer these questions and answers as well.

```
# YOUR CODE HERE
```

2.1.2 Method   State the method that is applicable to check whether the correctness of decision (`Correct`) depends on the person's native language (`NativeLanguage`). Explain your choice.

(your answer here)

2.1.3 Hypothesis   State the null hypothesis you are going to test. State the alternative hypothesis as well.

(your answer here)

2.1.4 Analysis   Perform the analysis using R. Provide your R code.

```
# YOUR CODE HERE
```

2.1.5 Interpretation   Based on the output obtained, can you conclude that the correctness of decision depends on the person's native language? Explain your answer.

(your answer here)

2.2 Lexical decision: short words compared

Imagine that you have to check whether the average reaction time is different for different short words (less than 5 letters). The question is: is it true that it is harder to recognise some short words than others? If it is true, we can proceed to more sophisticated analysis and think of factors that can affect the reaction time.

2.2.1 Data prepraration   Using `tidyverse` (`dplyr`) choose rows that correspond to correctly named words (column `Correct`) and words consisting of less than 5 letters (column `Length`). Save them to a new dataset, you should use this data set for this task.

```
# YOUR CODE HERE
```

2.2.2 Visualization   Create boxplots of reaction time for different words (column `Word`). Provide your R code. Judging by the graph, report two words with the highest median value of reaction time.

```
# YOUR CODE HERE
```

2.2.3 Method Choose an appropriate statistical method to answer the question stated at the beginning of this task. Perform the analysis and provide your R code.

```
# YOUR CODE HERE
```

2.2.4 Interpretation

2.2.4.1 Hypothesis State the null hypothesis you tested. State the alternative hypothesis as well.

(your answer here)

2.2.4.2 Conclusion Based on the output obtained, can you conclude that the reaction time differs for different words? Explain your answer.

(your answer here)

2.3 Lexical decision: reaction time and word features

Now you are suggested to check how reaction time (`RT`) is related to several word features: word frequency (`Frequency`), word family size (`FamilySize`) and number of synonyms (`SynsetCount`). Here you should use the original data set, not the filtered one from the previous task.

2.3.1 Visualization Plot a scatterplot matrix (table of scatterplots) for the pairs of variables chosen at the previous step using `GGAlly` library (function `ggpairs`). Find all pairwise correlations between these variables (use function `cor`). Which variables are positively associated? And negatively associated? The association between which variables is the strongest? Explain your answer.

```
# YOUR CODE HERE
```

2.3.2 Significance Check whether the correlation between reaction time and word frequency is statistically significant.

2.3.2.1 Hypothesis State the null hypothesis you are going to test. State the alternative hypothesis.

(your answer here)

2.3.2.2 Testing   Test the hypothesis stated above using R. Provide your R code.

```
# YOUR CODE HERE
```

2.3.2.3 Interpretation   Based on the output obtained, can you conclude that the reaction time is associated with word frequency? Explain your answer.

(your answer here)

2.3.2.4 Reporting   Report the correlation coefficient obtained. If it is statistically significant, interpret its value: state the direction of association (positive or negative) and the strength of association (approximately, strong or weak). There are no strict rules how to interpret the strength, for example, see here (p.9).

(your answer here)

2.3.3. Bivariate model

Create a bivariate linear model that will explain how reaction time is affected by word frequency.

2.3.3.1   State which variable is independent and which is dependent in our case.

(your answer here)

2.3.3.2   Run this model in R and report your code.

```
# YOUR CODE HERE
```

2.3.3.3   How does (on average) reaction time change when word frequency increases by one?

(your answer here)

2.3.3.4   Based on the output obtained, can you conclude that the word frequency affects the reaction time? Explain your answer. Can you conclude that higher word frequency leads to lower reaction time? Explain your answer.

(your answer here)

2.4. Lexical decision: multivariate regression

Now you are suggested to check how reaction time (`RT`) is affected by several word features: word frequency (`Frequency`), word family size (`FamilySize`), length of a word in letters (`Length`), word class (`Class`) and person's native language (`NativeLanguage`). Here you should use the original data set.

2.4.1. Using `tidyverse` (`dplyr`), choose rows that correspond to real words (see the column `PrevType`). Save them to a new dataset, you should use this data set for this task.

```
# YOUR CODE HERE
```

2.4.2 Create a multivariate linear model that will explain how reaction time is affected by word features stated above. Run this model in R and report your code.

```
# YOUR CODE HERE
```

2.4.3 Interpret the R output.

2.4.3.1 Write the equation of the model using the R output obtained, with estimates of coefficients substituted into the equation.

(your answer here)

2.4.3.2 All else equal, how does the reaction time change (on average) when the length of a word increases by one?

(your answer here)

2.4.3.3 Interpret the coefficient of `Class` (explain what does it show).

(your answer here)

2.4.3.4 Interpret the coefficient of `NativeLanguage` (explain what does it show).

(your answer here)

2.4.4

Perform some model diagnostics.

2.4.4.1   Report the R^2 of this model.

(your answer here)

2.4.4.2   Plot any graphs that can show whether there are some patterns in the residuals distribution. Can we conclude that residuals of this model are scattered randomly (with no patterns)?

```
# YOUR CODE HERE
```