Homework 4

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

Deadline: 17 February, 23:59

Frequent words, their acoustic duration and co-articulation effects

Many studies report shorter acoustic durations, more co-articulation and reduced articulatory targets for frequent words. The study of Fabian Tomaschek et al. (2018) investigates a factor ignored in discussions on the relation between frequency and phonetic detail, namely, that motor skills improve with experience. Twenty-seven German verbs with the vowel [a:] in the stem were used. All verbs were presented in a *sie ...* phrase which is disyllabic in its canonical form. Nine of these verbs were also presented in a phrase eliciting a monosyllabic verb form. Verbs were selected to cover a wide range of relative frequencies according to written and spoken corpus data.

Values:

- `LogDurationA` - log-transformed word duration (i.e. logarithms of word duration).
- `LogDurationW` - log-transformed segment duration.
- `Lemma`.
- `Participant` - participant ID.
- `Cond` - condition (slow, fast).
- `Exponent` - inflectional exponent of verbs: `-t`, `-en`, `-n`. By default: t.
- `Frequency` - log-transformed frequency of verbs in the corpus.

1.0 Data loading

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

For brevity, below we will refer to variables `LogDurationA` and `LogDurationW` as "word duration" and "segment duration" correspondingly despite the fact that they are actually logarithms of the durations.

1.1 Word duration and segment duration

Draw histograms for word duration and segment duration values.

1.2 Word duration and segment duration in slow and fast condition

Group the data by speaking condition (`Cond`) and estimate whether there is a difference in the word duration with the help of boxplot. Do the same for the segment duration.

It is reasonable to expect that both durations are shorter for fast speaking condition than for slow speaking condition? Can the graph you plotted confirm this? What kind of assertions can you make from the graph? E.g. can you assert something like "sample/population mean/median of word duration for fast speaking condition is shorter/longer than in slow speaking condition"?

2.1 Student's t-test

Now we want to check statistical significance of difference between

(a) word durarion in fast condition and word duration in slow condition,

(b) segment duration in fast condition and segment duration in slow condition using a Student's t-test. In other words, we want to check, is it true that these durations differ not only in the samples, but also in the populations.

2.1.1 Hypothesis

First of all, state the null hypothesis and the alternative you consider.

2.1.2 Application of test

Apply `t.test` to check the hypothesis.

2.1.3 Interpretation

Interpret results of the t-test performed. Report p-values obtained. Can you confirm that there is a difference between word duration in fast condition and word duration in slow condition in the population? The same question for the segment duration.

2.2 Confidence intervals

2.2.1 Explicit formula

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

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

Use it to find 95% confidence interval for a population mean of word durations.

2.2.2 Function `MeanCI`

Use function `MeanCI` from package `DescTools` (you have to install and load it first; use `install.packages` to install and `library` to load) to find the same confidence interval.

(The result 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.)

2.2.3 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 variable as in 2.2.1 and extract the confidence interval from the output. Does it coincide with the results of sections 2.2.1 or 2.2.2?

2.2.4 Different confidence level

Use function `MeanCI` to find 99% confidence interval for the same variable as in 2.2.1. Is it wider or narrower than 95% CI?

3.1 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.