Homework 4

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

Deadline: 22 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.

For this research people were asked to read texts with target German verbs aloud and then the duration of their speech was recorded. Participants had to speak in different conditions, slow and fast. In other words, they were asked to speak slowly/fast or the setting for speaking slowly/fast was created implicitly (so speakers did not understand that).

In this homework you are suggested to compare word duration and text segment duration for fast an slow speaking conditions. On the one hand, it is logical to suppose even without testing that duration in fast speaking condition should be shorter. On the other hand, before doing a more substantial research it might be helpful to check whether this intuitive suggestion holds, i.e. to make sure that the conditions of the experiment were thoroughly maintained (researchers did not swap conditions and recorded results correctly).

Variables of interest:

- `LogDurationA` - log-transformed word duration (i.e. logarithms of word duration).
- `LogDurationW` - log-transformed segment duration.
- `Cond` - condition (slow, fast).

For more information on this data set see the section at the end of this file.

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.

Is it 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 using a Student's t-test we want to decide whether the 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

is statistically significant. 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 (both for cases (a) and (b) above).

2.1.2 Application of test

Apply `t.test` to check the hypothesis (both for cases (a) and (b) above).

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:

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

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.)

2.2.2 Function `MeanCI`

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

(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.)

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 variables as in 2.2.1 and extract the confidence intervals 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 intervals for the same variables as in 2.2.1. Are they wider or narrower than 95% CI's?

Hint: use `conf.level` option.

2.3 ANOVA

We will consider only observations obtained in `slow` conditions in this section. Let us look at three groups with Exponent *-en*, *-n*, *-t*, respectively.

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

2.3.1 Hypothesis

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).

2.3.2. `aov()`

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

2.3.3 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?

2.3.4 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?

How can you explain a contradiction (if any) of this the outcome with outcome of 2.3.3?

2.3.5 ANOVA: Word durations

Repeat 2.3.1, 2.3.2 and 2.3.3 for word durations instead of segment durations.

2.3.6 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 2.3.3 and 2.3.5 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.)

Analyse 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.

```
# TukeyHSD(aov(....), "Exponent")
```

2.3.7 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 2.3.4. How can you explain the difference between these p-values?

2.3.8 Reporting the results

Write down you general conclusions for 2.3.1-2.3.7. 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.

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.

More information on the data set

The rest of the variables:

- `Lemma`.
- `Participant` - participant ID.
- `Exponent` - inflectional exponent of verbs: `-t`, `-en`, `-n`. By default: `-t`.
- `Frequency` - log-transformed frequency of verbs in the corpus.

Seventeen native speakers of German (9 female, mean age: 26, sd: 3), undergraduate students at the University of Tubingen, with no known language impairments, took part in the experiment.

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.