HW 2. Binomial test and t-test

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

Deadline: 24 February, 23:59

Your name:

1. Position of verbs in verses

The dataset "The last words in verses" contains a sample of lines taken from the RNC Corpus of Russian Poetry. Actually, there are two samples comprising the texts written in the 1820s and 1920s. We took only one line per author to keep our observations as independent as possible.

Variables:
- Decade — decade of creation: 1820s, 1920s.
- RhymedNwords — the number of words in the rhyming position (usually one word, but there are two words in cases such as 'I would like to get) wine' (which is rhymed with 'toad', see http://russian-poetry.ru/Poem.php?PoemId=18261)).
- RhymedNsyl — the number of syllables in the rhyming position.
- UPoS — part of speech of the last word. - LineText — a sampled verse.
- Author — the author of the text.

Can we decide that in verses written in 1920s, verbs are used in the rhyming posision more often or less often than expected for verbs in general? To calculate the probabilty to come across a verb in written Russian texts (general expectations), use the frequency dictionary of the Russian National Corpus "https://raw.githubusercontent.com/LingData2019/LingData/master/data/freq_rnc_ranked.csv" (Lyashevskaya, Sharoff 2009).

1.1 General expectations

Read the RNC frequency dictionary data. Verbs are coded as 'v' in the `PoS` field, and their frequency is shown in the `Freq.ipm.` field (relative frequency, # items per million words in the corpus).

Note that the file is tab-separated, so you have to add `sep='\t'` option to `read.csv` function. You probably also need to activate UTF-8 locale using command `Sys.setlocale(locale='UTF-8')` to see cyrillic letters in `Lemma` column (but it is not necessary for this exercise).

```
### YOUR CODE HERE
```

1.1.1

Calculate the probability to see verbs dividing the sum of their frequency by the sum of frequency of all words in the dictionary.

```
### YOUR CODE HERE
```

1.2 State hypothesis

Assume that every time the author decides which word to use as a last word of a verse, they toss a coin. If they get head, they use verb, otherwise they use word of different part of speech. Denote the probability to get head (i.e. use verb) by p. Recall that we are interested in the following question: can we decide that in verses written in 1920s, verbs are used in the rhyming posision more often or less often than expected for verbs in general?

State null hypothesis H_0 and alternative H_1 in terms of p.

Your answer

1.3 Analyse data

Read the dataset "The last words in verses". Filter out the relevant observations from 1920s, calculate the number of verbs observed in the sample, and the sample size.

Hint. You can use function `table` to calculate how many times each value occurs in a vector of factors.

```
### YOUR CODE HERE
```

1.3.1

Test stated hypothesis and find p-value. Use `binom.test` that will calculate p-value for binomial test. You have to put actual number of successes (i.e. *heads*), number of trials (i.e. coin-tossings) and expected probability of success that is used in null hypothesis.

```
### YOUR CODE HERE
```

1.4 Interpret results

Give your interpretation of obtained p-value. Answer the initial question: can we decide that in verses written in 1920s, verbs are used in the rhyming posision more often or less often than expected?

Your answer

2. 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:

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

2.0 Data loading

Load data (link), save it to variable `dur_word_freq` and look at the summary of the loaded data frame.

```
### YOUR CODE HERE
```

For brevity, below we will refer to variables `LogDurationW` and `LogDurationA` 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.

```
### YOUR CODE HERE
```

1.2 Segment duration in slow and fast condition

Uncomment and run the following code:

```
# boxplot(LogDurationA ~ Cond, data=dur_word_freq)
```

The result is so-called *box and whisker plot* for variable `LogDurationA` groupped by variable `Cond`. You can read about meaning of the elements of box plot in Wikipedia article (also, in Russian). What can you say about difference between values of `LogDurationA` that correspond to fast and slow conditions? Is it reasonable to expect that segment 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 segment duration in fast speaking condition is shorter/longer than in slow speaking condition"?

Your answer

1.3. Word duration in slow and fast condition

Repeat 1.2 using word duration instead of segment duration. Run appropriate code and interpret the resulting figure.

```
### YOUR CODE HERE
```

Your answer (interpretation of a graph)

2.1 Student's t-test

Now using 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). Justify your choice of alternative hypothesis.

Your answer

2.1.2 Application of test

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

```
### YOUR CODE HERE
```

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.

Your answer