HW 3. Binomial test and t-test

March 8, 2021

This homework is based on the homeworks given to students of the same course in 2019 taught by Olga Lyashevskaya, George Moroz, Alla Tambovtseva and Ilya Schurov.

Your name:

1. Position of verbs in verses

Description of dataset

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 *вина бы* 'I would like to get) wine' (which is rhymed with *жабы* '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.

Research question

We are interested in the following question: can we decide that in verses written in 1820s, verbs are used in the rhyming position more often or less often than expected for verbs in general?

1.1 General expectations

To calculate the probability to come across a verb in written Russian texts (general expectations), use the frequency dictionary of the Russian National Corpus (Lyashevskaya, Sharoff 2009).

1.1.1 Read the file First of all, we have to read the RNC frequency dictionary data.

We will use some functions from `tidyverse` library, so do not forget to put line `library(tidyverse)` in the beginning of your code

Note that the file is tab-separated, so we have to use `read_tsv` instead of `reac_csv` function in order to read it.

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).

Read frequency dictionary file and save the result as variable `freq_rnc`.

```
# YOUR CODE HERE
```

1.1.2 Investigate dataset Use `head` command to show the beginning of the table stored in variable `freq_rnc`.

```
# YOUR CODE HERE
```

Note that verbs are coded as `'v'` in the `PoS` field, and their frequency is shown in the `Freq(ipm)` field (relative frequency, number of items per million words in the corpus).

1.1.3. Find the probability Calculate the probability to see verbs dividing the sum of their frequency by the sum of frequency of all words in the dictionary. Note that due to presence of brackets in the name of column `Freq(itm)`, to access this column, you have to put its name into backticks, i.e.:

```
all_freqs <- freq_rnc$‘Freq(ipm)‘ # access column Freq(ipm)
```

```
## Error in eval(expr, envir, enclos): object ’freq_rnc’ not found
```

```
all_freqs[1:10] # show first 10 elements
```

```
## Error in eval(expr, envir, enclos): object 'all_freqs' not found
```

```
# 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 1820s, verbs are used in the rhyming position 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: *Write your answer here*

1.3 Analyse data

1.3.1. Read the dataset Read the dataset "The last words in verses" and put it variable `lvw`. Filter out the relevant observations from 1820s, 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
```

1.3.2 Do statistical testing

Test stated hypothesis and find p-value. Use `binom.test` that will calculate p-value for binomial test (note that it uses two-sided alternative by default). 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. Would you reject null hypothesis? Answer the initial question: can we decide that in verses written in 1820s, verbs are used in the rhyming posision more often or less often than expected?

Your answer: *Write your answer here*

2. Two-sample Student's t-test

Setting

Let's consider some synthetic data. Assume that variables `sample1` and `sample2` consists of length of sentences, randomly chosen from a large corpora of texts on two languages.

```
sample1 <- c(1, 5, 4, 3, 5)
sample2 <- c(4, 30, 20, 5, 10, 50)
```

We are interested in the following question: is it correct that these two languages has the same average length of sentence?

2.0 Sample means

Find sample means.

```
# YOUR CODE HERE
```

2.1 What intuition says?

We see that the there is a difference in sample means. Is it enough to say that there is a difference in the corresponding languages (i.e. populations)?

Your answer: *Write your answer here*

2.2 Statistical testing

Now let's state our hypothesis.

Null hypothesis: these two samples are drawn from populations with the same mean, i.e. from languages with the same average length of sentence. Difference between sample means should be attributed to sampling error, i.e. randomness in the sampling process (selection of random sentences from the large corpora).

Denote average length of sentence in the first language by μ_1 and in the second language by μ_2. (We don't know these values!) Then mathematical statement of null hypothesis is:

$$H_0 \colon \mu_1 = \mu_2$$

Alternative: samples are drawn from populations with different means. (I.e. languages has difference average length of sentence.) Difference between sample means cannot be attributed to sampling error. We didn't have any ideas which population has larger mean before looking at the data.

Mathematically speaking:

$$H_0 \colon \mu_1 \neq \mu_2.$$

Should we reject null hypothesis in favor of alternative? Let's find p-value to answer this question.

```
t.test(sample1, sample2)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -2.2245, df = 5.1061, p-value = 0.07558
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -34.875664   2.408998
## sample estimates:
## mean of x mean of y
##   3.60000  19.83333
```

Function `t.test` can perform one-sample or two-sample t-test. If we put two vectors in it, it will perform two-sample t-test we need here. There is an option `alternative` that controls alternative hypothesis, by default it's `''two.sided''` (i.e. $\mu_1 \neq \mu_2$). We see the p-value returned by this function.

2.3 Your interpretation Interpret the p-value you see above. Would you reject null hypothesis in favor of alternative? What can you say about your initial research question?

Your answer: *Write your answer here*

2.4 Let's play with data How the result of t-test (i.e. p-value) changes when you change the data? Try to increase or decrease difference between sample values, increase or decrease sample size, increase or decrease sample variances. Provide your experiments, their results and your explanations.

Your answer: *Write your answer here*

Alternatives

What changes if we know before looking at the data that we are interested only in the alternative when mean of the first population is smaller than the mean of the second population? (For example, we do drug trial and values in our samples are disease duration, i.e. the smaller the better. First sample is from the treatment group who actually received the drug and the second sample is from the control group who received placebo; Then we are interested in the case when the drug is effective, i.e. decrease disease duration, but not interested in the case when drug is ineffective, meaning both that it does not affect the disease duration or increase it.)

```
t.test(sample1, sample2, alternative = ’less’)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -2.2245, df = 5.1061, p-value = 0.03779
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.595627
## sample estimates:
## mean of x mean of y
##   3.60000  19.83333
```

2.4 Interpret the result Would you reject null hypothesis in favor of alternative? What is your decision about the drug in this drug test?

Your answer: *Write your answer here*

2.5 Opposite alternative

2.5.1 p-value What happens if we are interested in the opposite alternative? (For example, we do drug trial like previously, but now values in the samples are life expectancy, the larger the better.) Use `t.test` with correct `alternative` argument (look in the documentation which is correct) and find p-value.

```
# YOUR CODE HERE
```

2.5.2 Interpretation Explain, why p-value is so large? Would you reject null hypothesis in favor of alternative in this case? How can you answer your initial question (about drug trial) in this case?

Your answer: *Write your answer here*

Data frames

Two samples can be written in one dataframe using additional variable that will denote the number of group. Let us create such dataframe:

```
dat <- bind_rows(
  data.frame(value=sample1, sample_id=1),
  data.frame(value=sample2, sample_id=2)
)
```

```
## Error in bind_rows(data.frame(value = sample1, sample_id = 1), data.frame(value = sample
```

```
dat
```

```
## Error in eval(expr, envir, enclos): object 'dat' not found
```

You can now perform `t.test` on this dataset in the following way:

```
t.test(value ~ sample_id, data=dat)
```

```
## Error in eval(m$data, parent.frame()): object 'dat' not found
```

3. 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 (for example, 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).

3.0 Data loading

Load data (link), save it to variable `dur_word_freq` and print the beginning of this dataset with `head`. (Note that this is csv-file, not tab-separated, so you have to use function `read_csv`.)

```
# 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.

3.1 Word duration and segment duration

Draw histograms for word duration and segment duration values.

```
# YOUR CODE HERE
```

3.2 Segment duration in slow and fast condition

Run the following code:

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

```
## Error in eval(m$data, parent.frame()): object 'dur_word_freq' not found
```

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: *Write your answer here*

3.3. 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 the graph) *Write your answer here*

3.4 Student's t-test

Now using Student's t-test we want to decide whether the difference between

(a) word duration 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.

3.5.1 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 *Write your answer here*

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

```
# YOUR CODE HERE
```

3.5.3 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 *Write your answer here*