Homework 2

Linguistic Data: Quantitative Analysis and Visualisation. Linguistic theory group

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

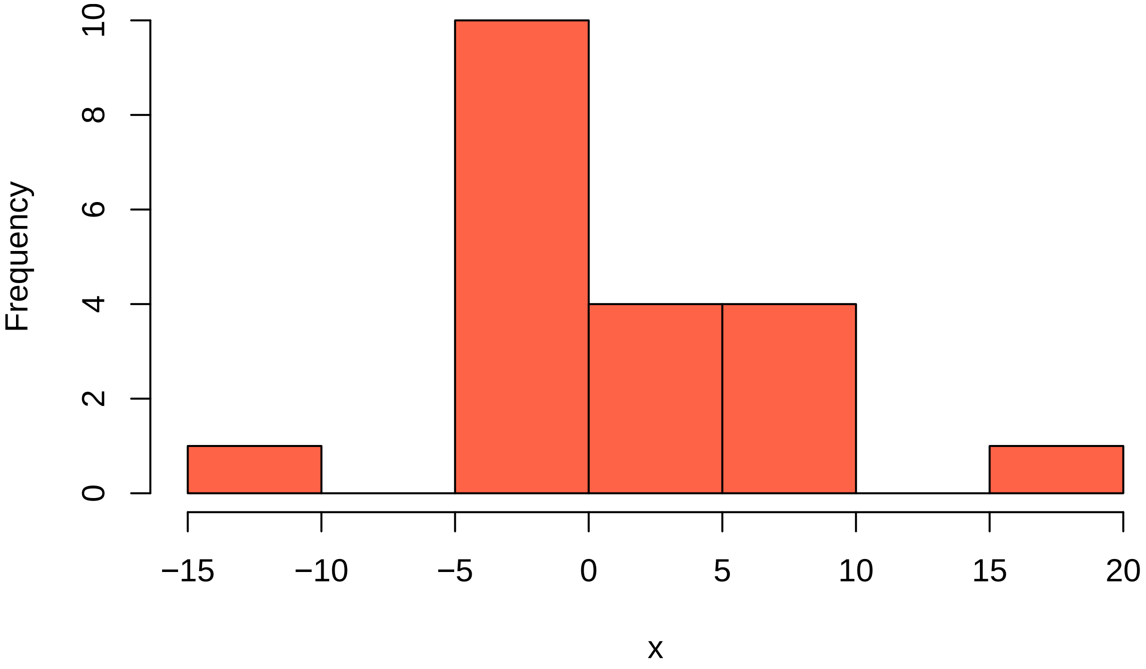
Course webpage: http://math-info.hse.ru/s20/g

Part 1. A preliminary training

Do not use R (RStudio) to solve problems in Part 1.

Problem 1

Look at the following histogram and answer the questions.



1.1

What is the proportion of values in the sample that exceed 5? Explain your answer.

Your answer and explanation here

1.2

Indicate the interval where the median of this sample can lie. Explain your answer.

Your answer and explanation here

1.3

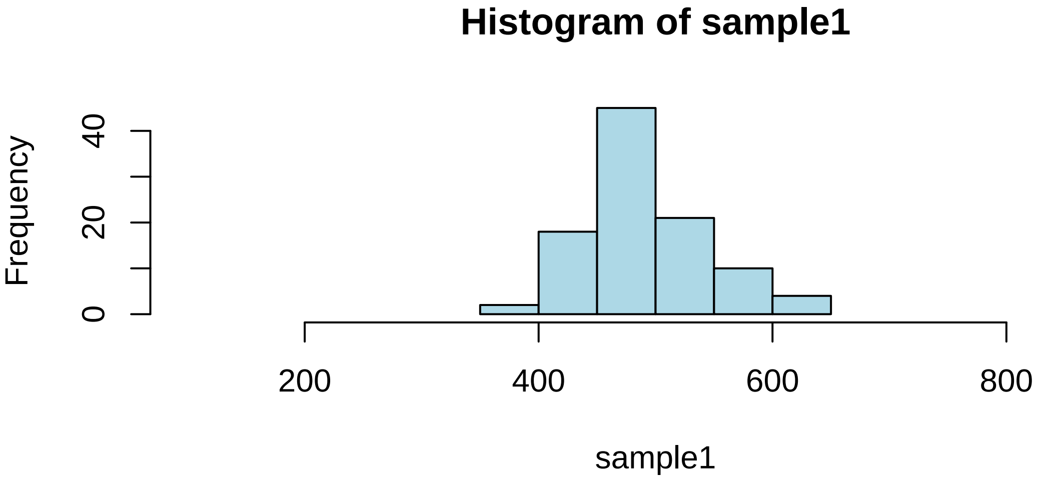
How the histogram will change if we add an element 7 to the sample? Explain your answer.

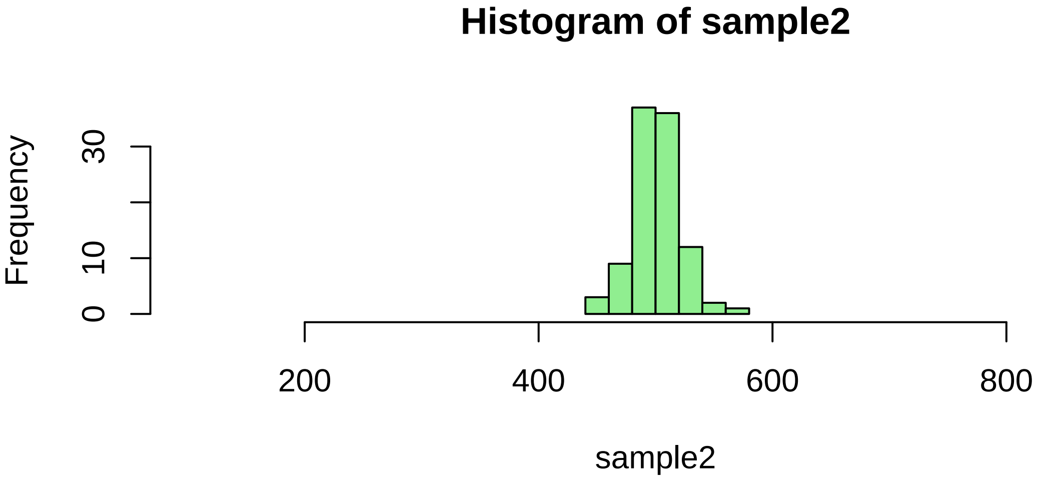
Your answer and explanation here

Problem 2

2.1

Look at the histograms of two samples. They illustrate the distribution of normalized average reaction time to frequent words (in ms) in two groups of people.



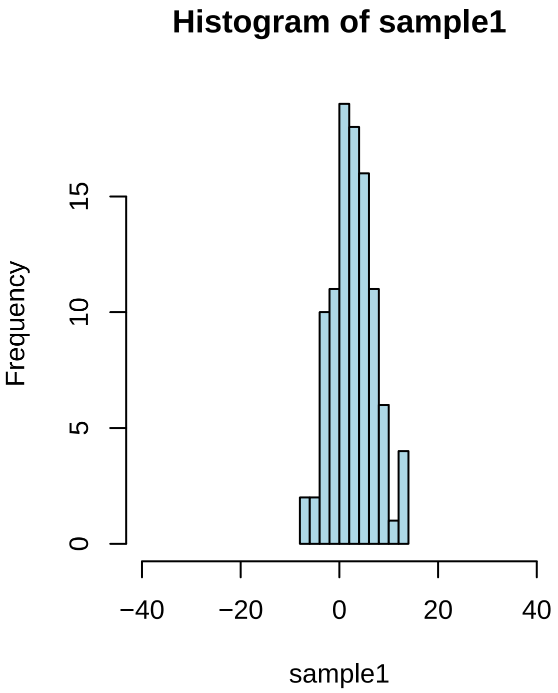


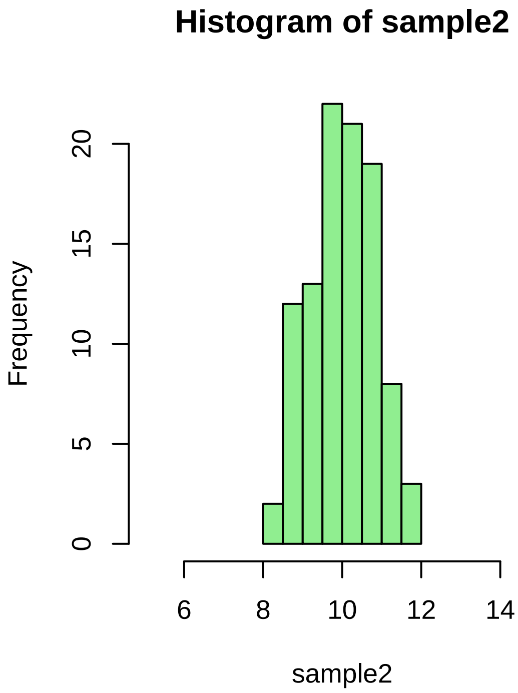
Which of the samples has a larger variance? Explain your answer.

Your answer and explanation here

2.2

Look at the histograms of two samples.



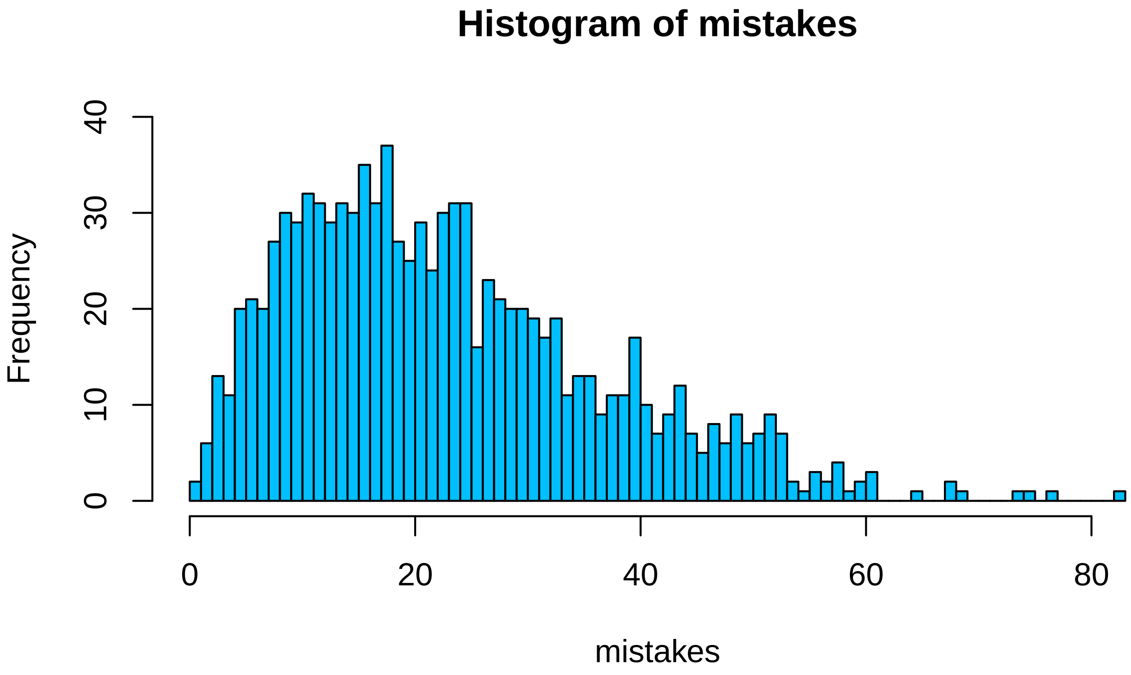


Which of the samples has a larger variance? Explain your answer.

Your answer and explanation here

Problem 3

Below is the histogram of the number of mistakes students made while writing an examination essay in English. Look at the histogram and answer the questions.



3.1

Is it true that 50% students made more than 35 mistakes?

Explain your answer below.

Your answer and explanation here

3.2

Is it true that most students made no more than 10 mistakes?

Explain your answer below.

Your answer and explanation here

3.3

Which of the following values is closer to the median of `mistakes`: 10, 20, 30, 40?

Explain your answer below.

Your answer and explanation here

Part 2. Exact binomial test

Problem 4. Type I and type II errors

A magician claims that he can predict the future. To test his abilities, the following experiment is conducted. A fair coin is tossed several times, every time magician tries to guess the outcome before he can see it. Sometimes his guess is correct, sometimes it is not. We allow for the magician to make mistakes from time to time (nobody is perfect, even magicians).

We toss a coin n times and denote number of correct guesses by X. If X is large enough, we say that the magician proved that he has paranormal abilities. Otherwise, we say he didn't prove it.

Denote probability that magician guesses the result of one coin-tossing by p. Then our hypotheses are as follows:

- $H_0\colon p = 1/2$ (i.e. random guessing occurs)
- $H_1\colon p > 1/2$ (i.e. the magician's guessing abilities are better than random guess, though probably not perfect)

4.1

Assume that the magician cannot predict future, thus H_0 holds true (but we don't know it). We conduct an experiment as follows. We toss a fair coin three times, i.e. $n = 3$. If he guesses correctly results of all three tossings, i.e. $X = 3$, we conclude he can predict future. Otherwise, we say that we don't have enough evidence to claim the magician predict future.

4.1.1

How likely we will claim that the magician can predict future? Find exact probability.

Your answer and explanation here

4.1.2

How likely we will say that we don't have enough evidence to claim the magician predict future. Find exact probability.

Your answer and explanation here

4.1.3

Consider outcomes that happen in 4.1.1 and 4.1.2. Which of them constitute type I error? Type II error?

Your answer and explanation here

4.1.4

Find probability to make type I error. Type II error.

Your answer and explanation here

4.2

Assume that the magician can predict future, namely, his actual probability to correctly guess the result of one tossing is 2/3. Solve problem 4.1 under this new assumption. Highlight the differences between two settings.

Your answer and explanation here

Problem 5. p-value

Assume that the magician from the previous problem cannot guess future. Let $n = 11$, i.e. we toss a coin 11 times. We claim the magician can predict future (i.e. reject H_0) if the number of correct guesses X is as large as some predefined number X_{crit}.

Your answer and explanation here

5.1

Let $X_{crit} = 11$, i.e. we need to have 11 correct guesses out of 11 to claim that the magician can predict future. What is probability to make Type I error in this case?

Your answer and explanation here

5.2

Let $X_{crit} = 10$, i.e. we need to have at least 10 correct guesses out of 11 to claim that the magician can predict future. What is probability to make Type I error in this case? You can use function `pbinom` to find answer.

Your answer and explanation here

5.3

Find the smallest integer number such that if X_{crit} is equal to that number, the probability to make Type I error is smaller than the significance level 0.05.

Your answer and explanation here

5.4

Assume that we obtained X_{obs} correct guesses out of 11. We don't know X_{obs}, but we know that the corresponding p-value (i.e. probability to obtain X_{obs} correct guesses or more provided that null hypothesis holds) is less than 0.01. Can you conclude, without finding exact value of X_{obs}, which value is larger: X_{obs} or the value for X_{crit} that you found in 5.3? Use to explain your answer.

Your answer and explanation here

5.5

Assume we actually obtained 7 correct guesses out of 11. Find p-value, i.e. probability to obtain 7 correct guesses or more provided that null hypothesis holds true.

Your answer and explanation here

Problem 5. Exact binomial test

In a certain language there are two forms of a word "go": normal and dialectical. Whe know that if we select random person from the Country, this person will use normal form with probability 2/3 and dialectical form with probability 1/3. (One person uses only one form all the time.) Researcher suggests that the percentage of people who use dialectical form in a particular City is higher than in the Country. To prove this point, she proceed with the following experiment. Random person from the City is selected and his/her usage of word "go" is recorded. This is repeated n times (the same person theoretically can be chosen more than one time, but the City is large comparatively to n, so it rarely occurs in practice).

The results are as following: 20 informants where selected ($n = 20$), 17 of them use dialectical form.

Is it enough to say that the percentage of people who use dialectical form in the City is higher than in the Country?

5.1 Hypothesis

Denote the probability that randomly selected person from the City uses dialectical form by p. State the null hypothesis and alternative.

- H_0: *Your answer here*
- H_1: *Your answer here*

5.2 Find p-value

Recall that p-value is a probability to get the data that we obtained or "more extreme" (more convincing to reject null hypothesis in favor of alternative) provided that null hypothesis holds. Find p-value for your data.

Your answer and explanation here

5.3 Conclusion

Will you reject null hypothesis? Use significance level of 5%.

Your answer and explanation here

5.4 Answer

Can we claim that we have enough evidence to say that the percentage of people who use dialectical form in the City is higher than in the Country?

Your answer and explanation here

Supplementary reading

Use of exact binomial test in linguistic research:

- Gries, Stefan Th. "Phonological similarity in multi-word units." Cognitive Linguistics 22.3 (2011): 491-510. Link
  Stefan Gries proves that alliteration is observed in multi-word expressions more often than in general.
- Harald Bayen (2008: 51-52) evaluates the probability of observing exactly one occurrence of the word *hare* in the corpus sample of 1 mln words given its estimated frequency of 8.23 words per million according to the SELEX frequency database.