Homework 3

Linguistic Data: Quantitative Analysis and Visualisation

Deadline: 9 February, 12:00

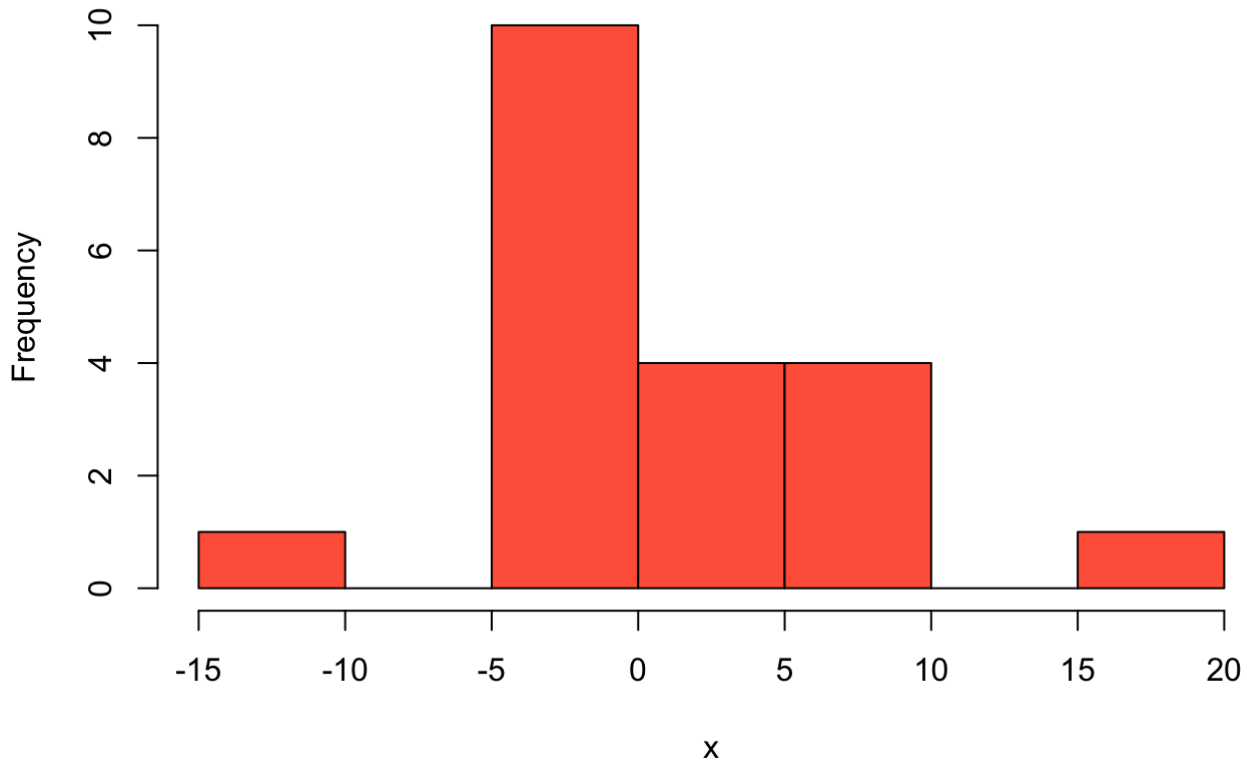
The solutions should be submitted via GitHub.

Part 1. A preliminary training

Do not use R (RStudio) to solve problems in Part 1. Answers won't be evaluated.

Problem 1

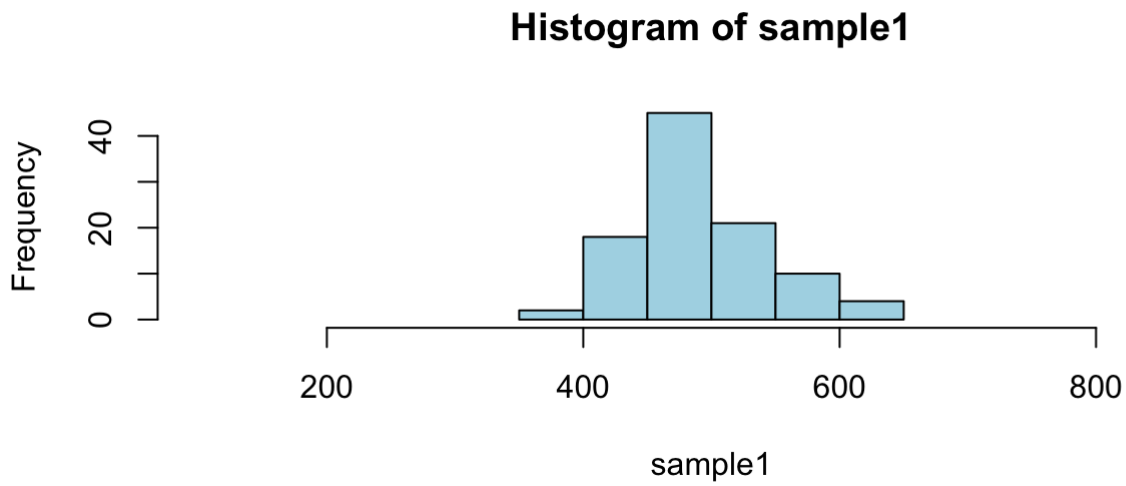
Look at the following histogram and answer the questions.

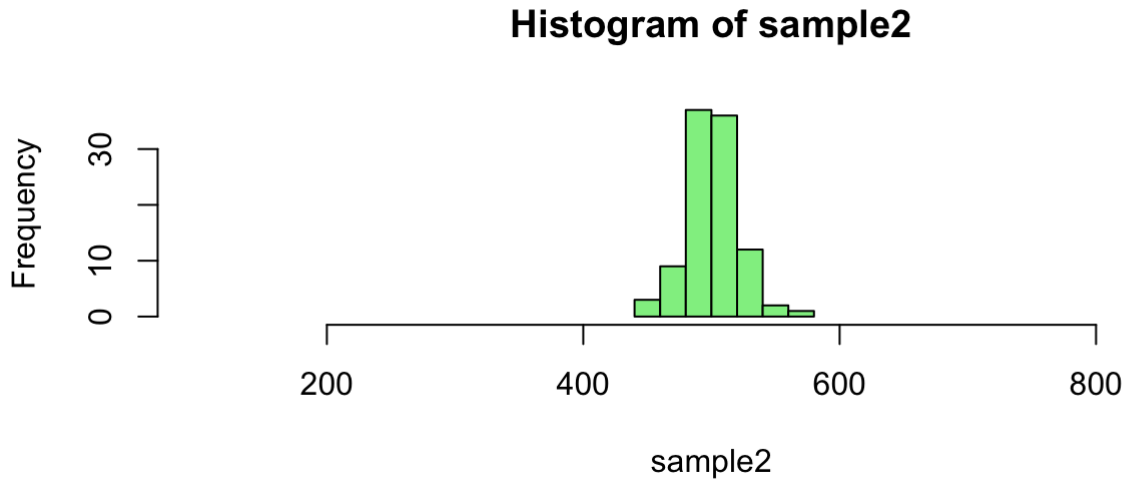


a. What is the proportion of values in the sample that exceed 5? Explain your answer.
b. Indicate the interval where the median of this sample can lie. Explain your answer.
c. How the histogram will change if we add an element 7 to the sample? Explain your answer.

Problem 2

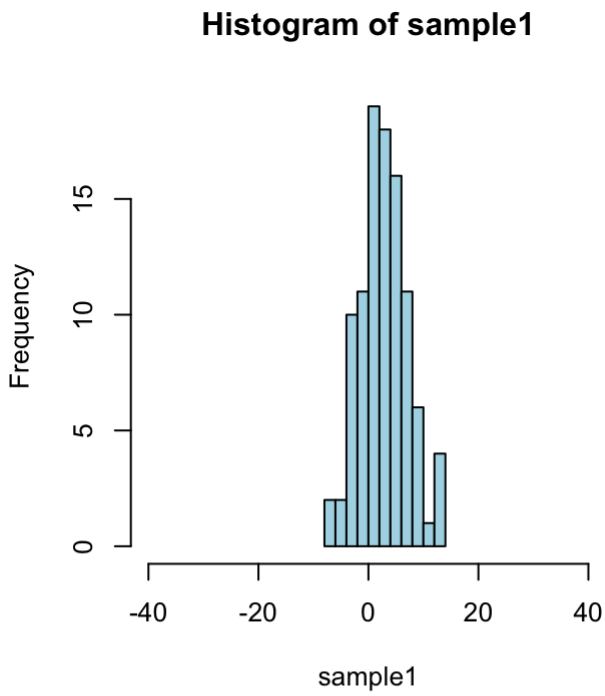
a. Look at the histograms of two samples. They illustrate the distribution of normalized average reaction time to frequent words (in ms) in two groups of people.

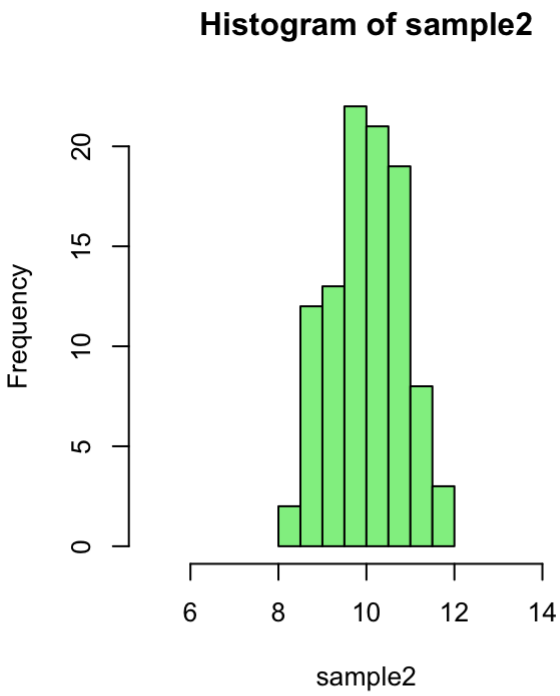




Which of the samples has a larger variance? Explain your answer.

b. Look at the histograms of two samples.





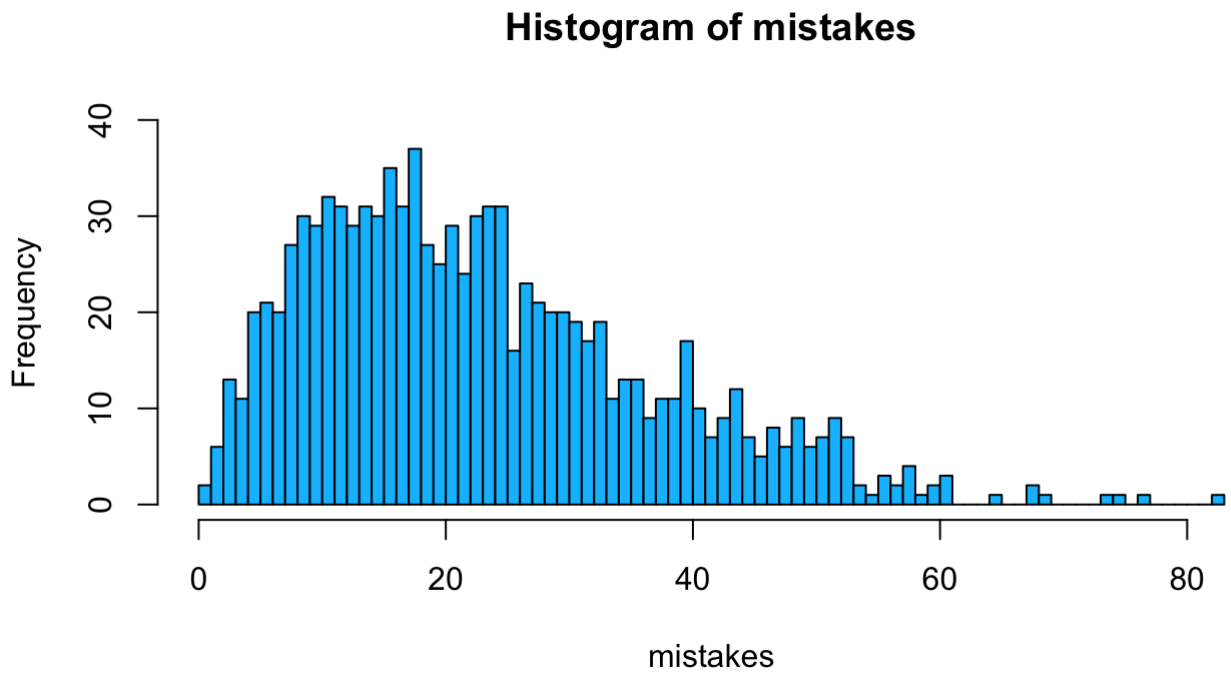
Which of the samples has a larger variance? Explain your answer.

Part 2

Do not use R (RStudio) to solve problems in Part 2. Answers for problem 3 will be evaluated. Please paste YES or NO into (empty) code blocks and explain you answer below the block.

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:

3.2

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

Explain your answer below:

3.3

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

Explain your answer below:

Problem 4. Exact binomial test

The null hypothesis is that $p = 0$ (i.e. no success is possible). In a dataset, there is only one success out of 1 000 000 observations. Will you reject the null hypothesis?

Explain your answer below:

Part 3

Use R (RStudio) to solve problems in Part 3. Your answers will be evaluated. Please paste R code into R code blocks and explain you answer below the block, if needed.

Problem 5

Here is a sample of respondents' age:

$44, 50, 42, 64, 66, 42, 72, 56, 72, 54, 46, 48, 48, 52, 50, 66, 84.$

5.1

Arrange them in a vector and call it `age`.

5.2

Examine the type of `age` variable (numeric, character, etc).

5.3

Plot the histogram of the vector `age` with 5 bins. Change its color to any you want. (Use either R basic or ggplot2 style for plotting.)

Problem 6

Here is a series of words:
pie, bar, bar, pie, pie, bar, bar, chart.

6.1

Arrange elements above in a vector and call it `words`

6.2

Calculate the relative frequences of values in `words` measured in percent.

Problem 7. Position of verbs in verses

The dataset "The last words in verses" (https://raw.githubusercontent.com/LingData2019/LingData2020/master/data/poetry_last_in_lines.csv) is based on texts written in the 1820s and 1920s (Corpus of Russian Poetry of the Russian National Corpus). Authors collected only one line per author to keep observations as independent from each other 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 (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 in the rhyming posision are used differently (more often or less often) than expected for verbs in general?

Let's assume that the probability for verbs to be used in any position ('in general') is 17% (according to (http://www.ruscorpora.ru/new/corpora-stat.html)).

7.1 State hypothesis

What is your null hypothesis H_0 and what is the alternative hypothesis H_1?

```
H0:
H1:
```

7.2

Read the dataset "The last words in verses" (https://raw.githubusercontent.com/LingData2019/LingData2020/master/data/poetry_last_in_lines.csv). Filter out the relevant observations from 1920s, calculate the number of verbs observed in the sample, and the sample size.

7.3

Use an exact binomial test to calculate p-value.

7.4 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?

7.5

(A bonus problem, extra points in evaluation). Repeat 2.3 for verses written in the 1820s.

Write down your general conlusions about data provided for both 1920s and 1820s data.

Problem 8. One-sample t-test

Using Icelandic data on vowel duration from seminar Link (https://raw.githubusercontent.com/LingData2019/LingData2020/master/data/icelandic.csv) test the null hypothesis that the population mean of vowel duration in speaker `shg05` equals 73 (ms). To perform a one-sample t-test, you can use the following example of R code:

```
t.test(sample, mu = 7725) # mu is a population mean
```

8.1

Write down a two-tailed alternative hypothesis.

```
H1:
```

8.2

Perform a one-sample t-test.

8.3

Interpret results.

Supplementary reading

Use of exact binomial test in lingiistic research:

- Gries, Stefan Th. “Phonological similarity in multi-word units.” Cognitive Linguistics 22.3 (2011): 491-510. Link (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.7412&rep=rep1&type=pdf) 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.

On measures of central tendency:

- Levshina 2015, Chapter 3 (p. 48); Gries 2009, Chapter 1.3 (p. 116).

On t-test:

- Gries 2009, Chapter 3 (p. 198).