Homework 1

Linguistic Data: Quantitative Analysis and Visualisation. Linguistic theory group.

Deadline: 9 February, 23:59:59

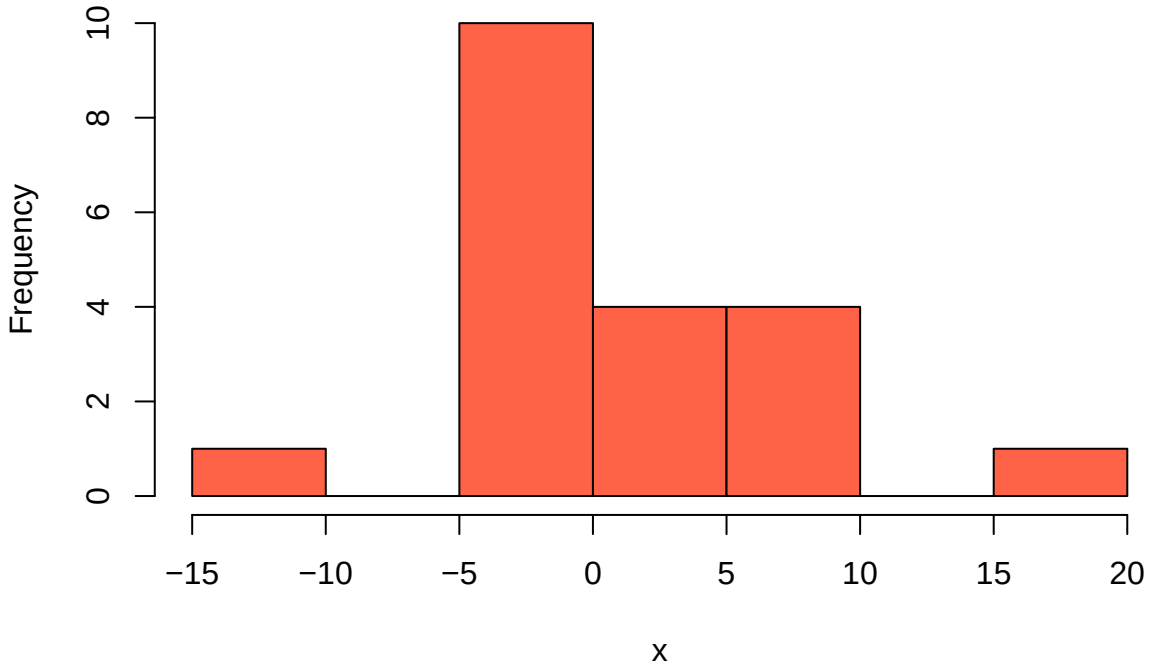
The solutions should be submitted via GitHub.

Part 1. A preliminary training

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

Problem 1

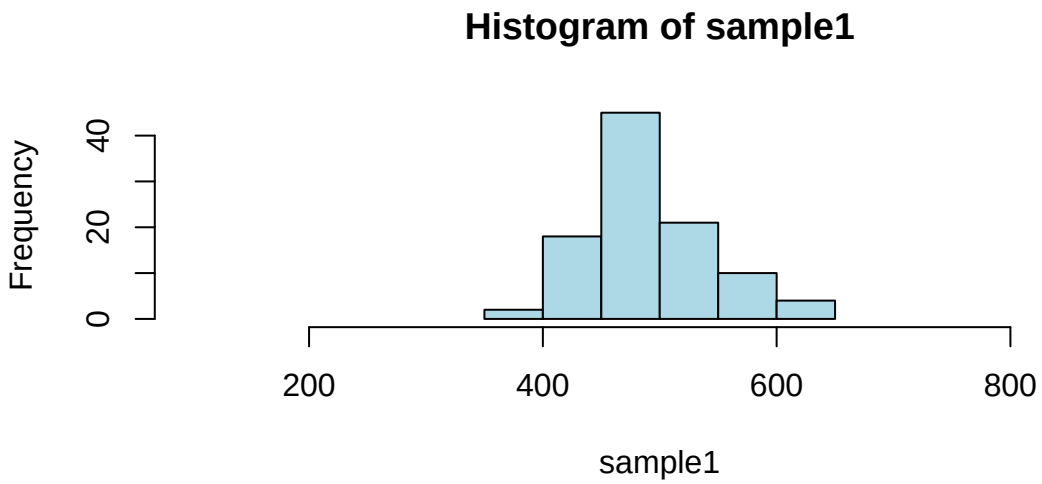
Look at the following histogram and answer the questions.

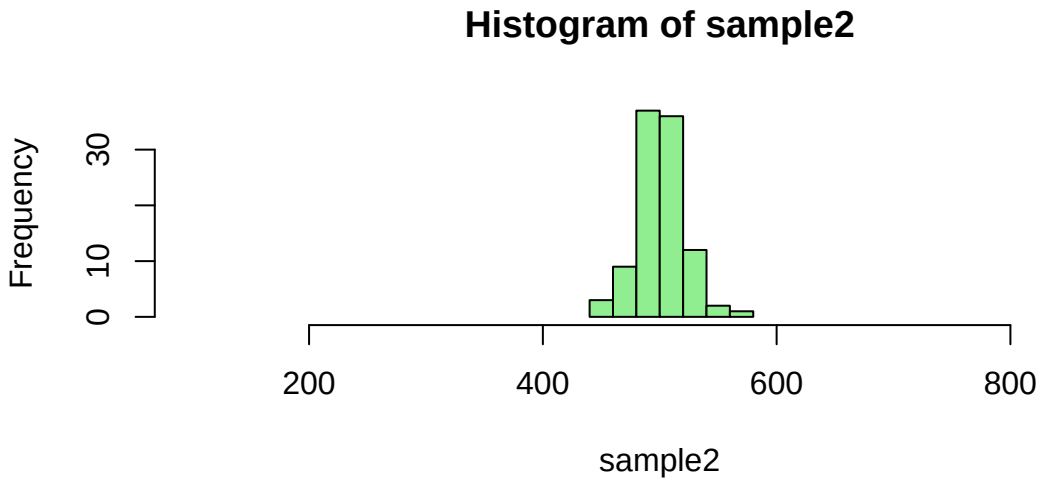


a. What is the proportion of values in the sample that exceed 5? Explain your answer.
b. Indicate the interval where the median of this sample can lie. Explain your answer.
c. How the histogram will change if we add an element 7 to the sample? Explain your answer.

Problem 2

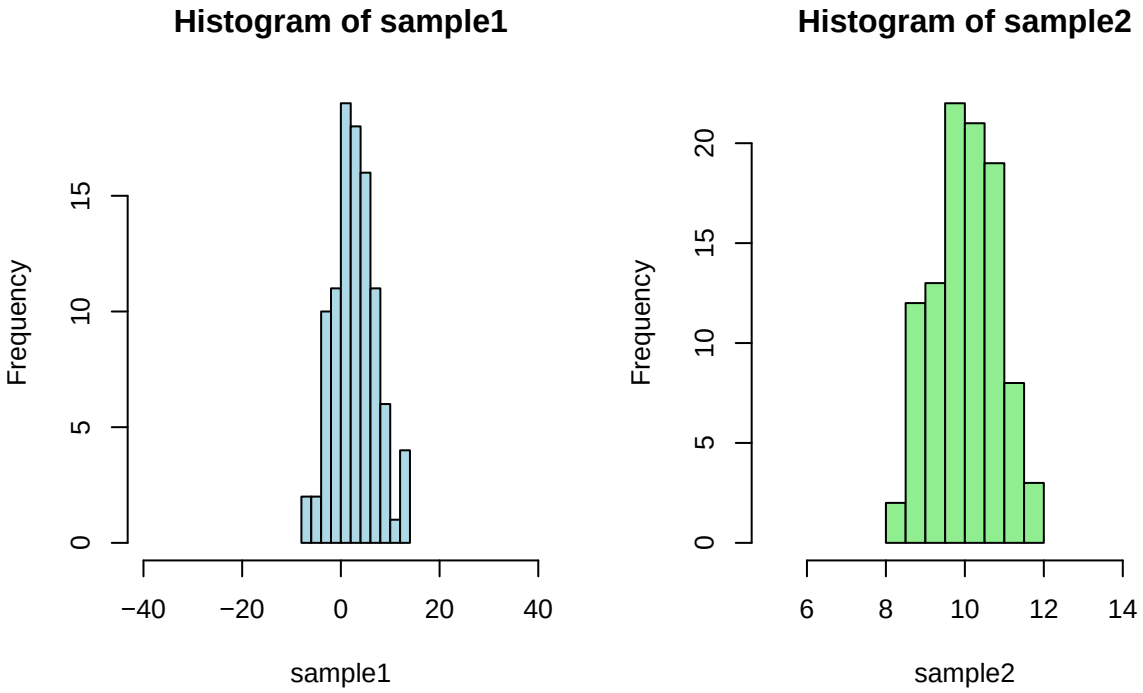
a. Look at the histograms of two samples. They illustrate the distribution of normalized average reaction time to frequent words (in ms) in two groups of people.





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

b. Look at the histograms of two samples.



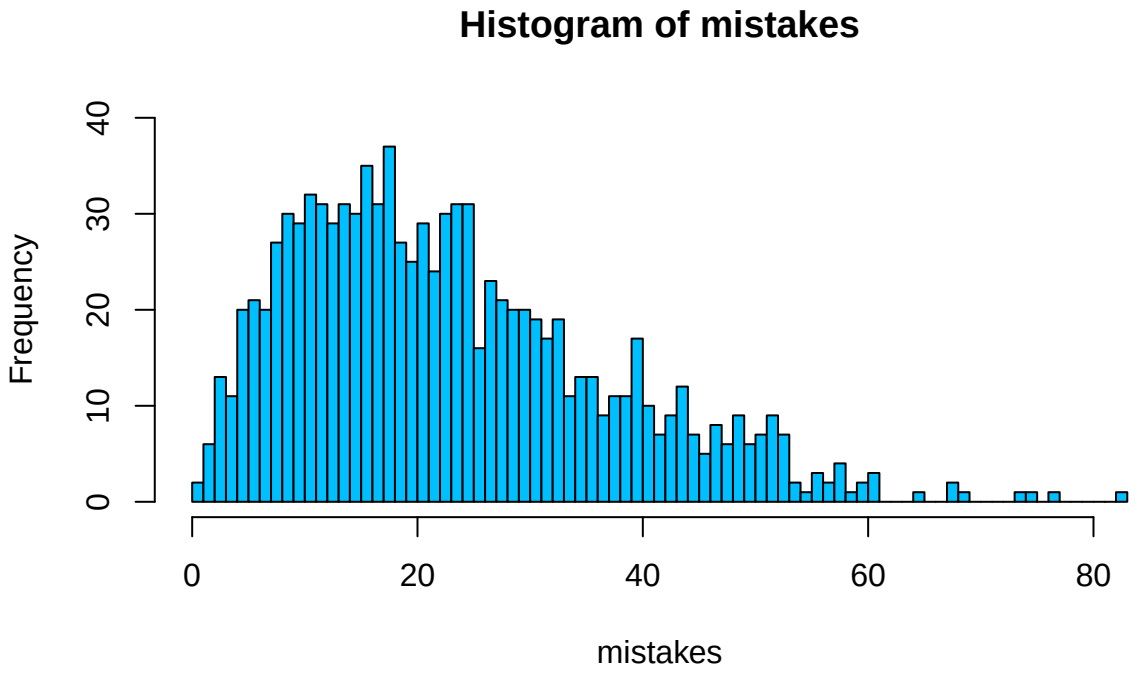
Which of the samples has a larger variance? Explain your answer.

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 the median of `mistakes`: 10, 20, 30, 40?

Explain your answer below:

Problem 4. 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.

4.1 Hypothesis

State the null hypothesis and alternative.

H_0:

H_1:

4.2 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.

Hint. There is a function `dbinom(k, n, p)` in R that calculates probability to get `k` heads if you toss a coin `n` times and probability to obtain a head in one tossing is `p`. You can use this function here.

4.3 Conclusion

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

4.4 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?

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.

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.

On measures of central tendency:

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