Homework 1

Linguistic Data: Quantitative Analysis and Visualisation

Olga Lyashevskaya, George Moroz, Ilya Schurov and Alla Tambovtseva

Deadline: 27 January, 23:59

The solutions should be submitted via Google forms.

A link to the form: https://goo.gl/forms/TBx0wLPofFUfZFrI3.

Part 1

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

Problem 1

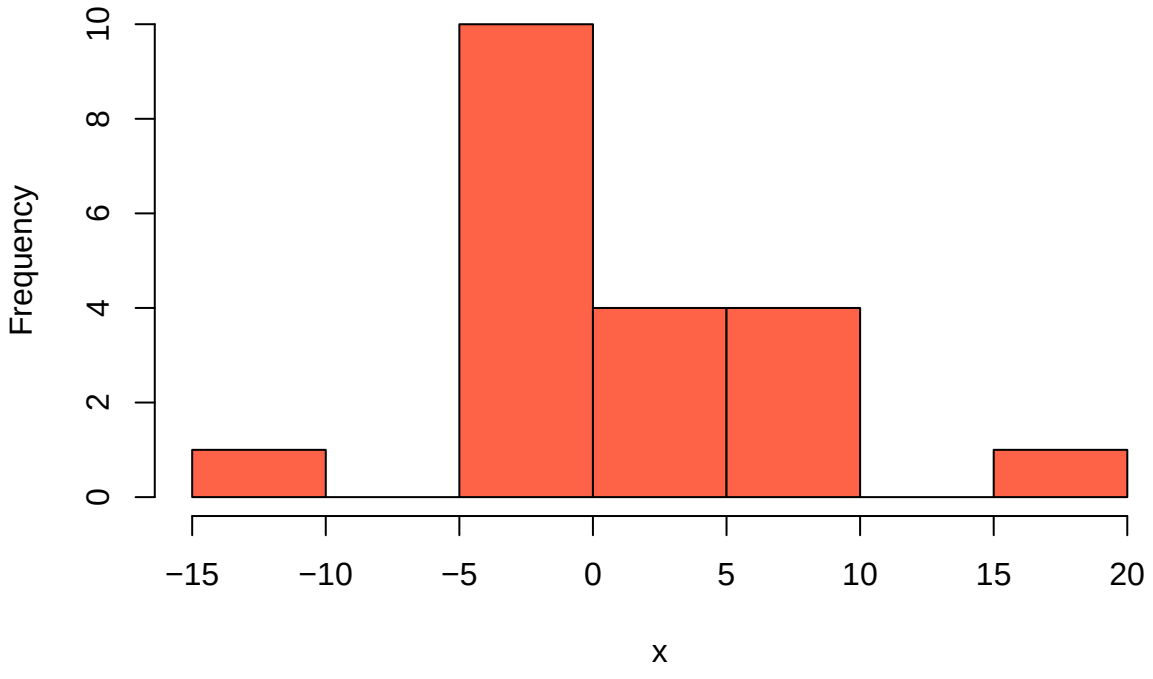
Consider the following sample:

2, 8, 7, 8, 1, 12, 6, 4, 10, 2

a. Find the median of this sample.
b. Find the sample mean. Provide your calculations.
c. Find the sample variance (corrected, unbiased one) and its standard deviation. Provide your calculations.
d. Add elements -20 and 80 to the sample. Calculate the median and the mean of the updated sample. Which of the statistics has changed more dramatically? Provide both your calculations and answers.

Problem 2

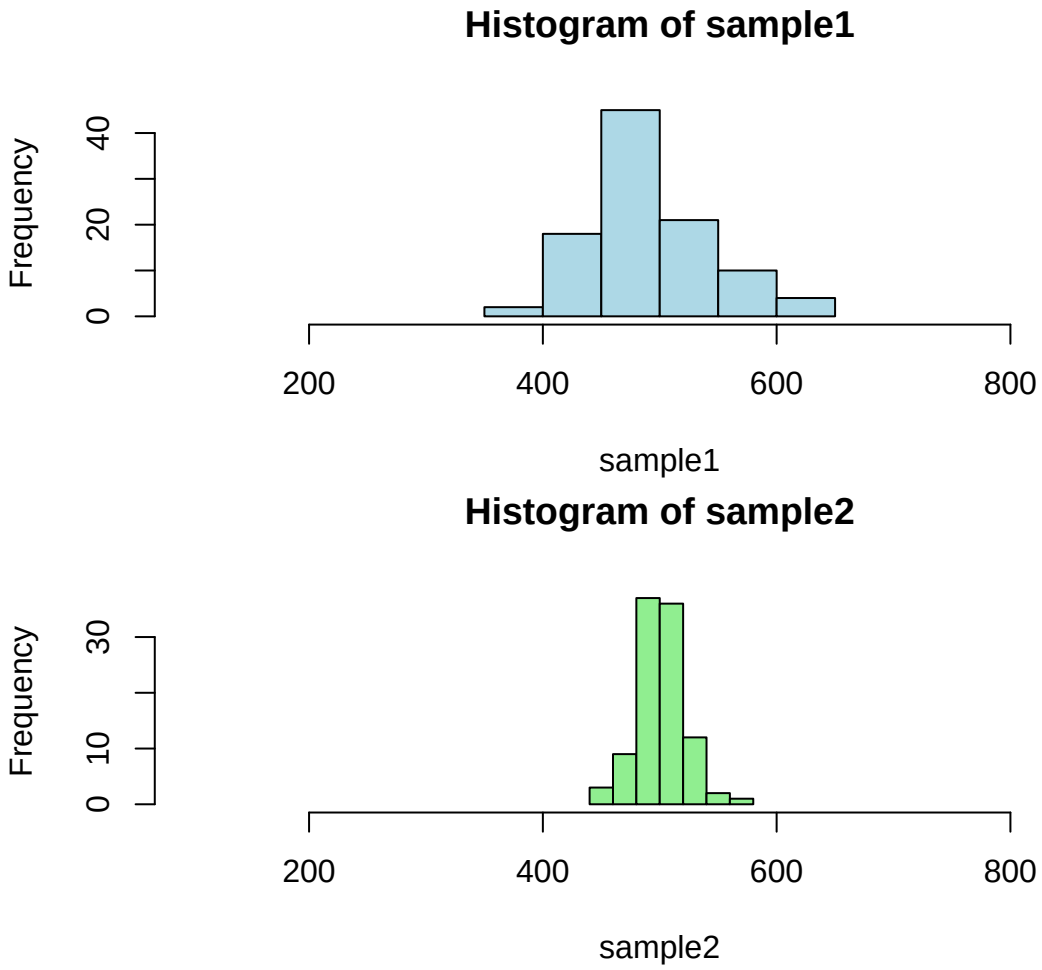
Look at the following histogram and answer the questions.



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.

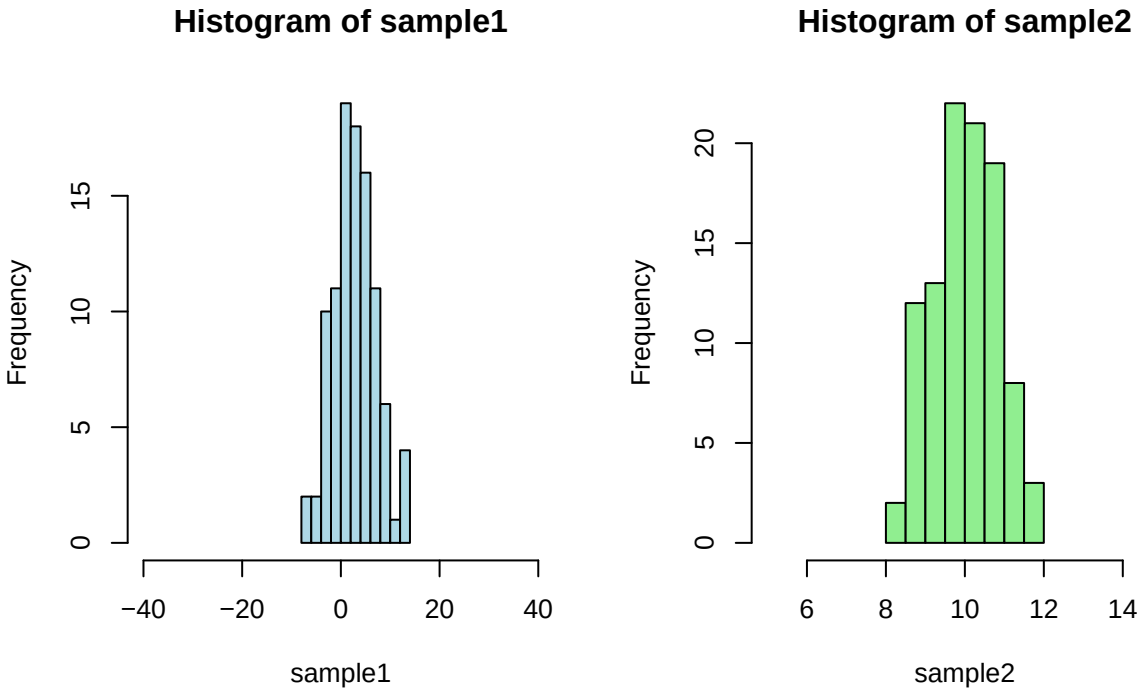
Problem 3

a. Look at 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.

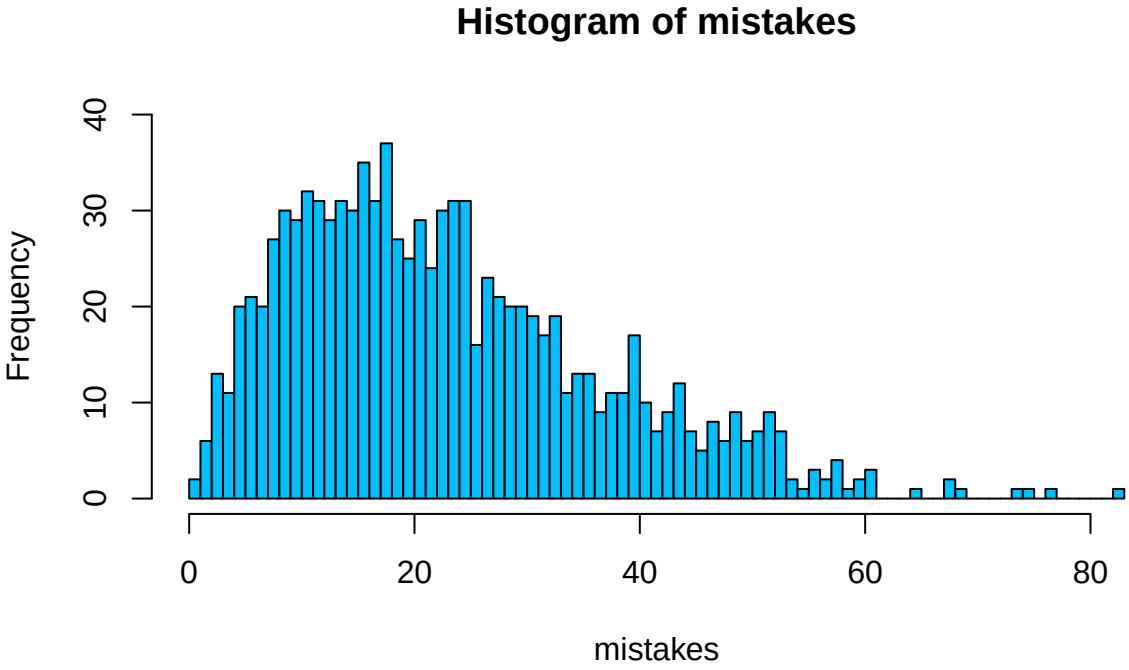
b. Look at histograms of two samples.



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

Problem 4

Below is the histogram of the number of mistakes students made in a writing paper in English exam. Look at the histogram and answer the questions.



a. Is it true that the 50% of students made more than 35 mistakes? Explain your answer.

b. Is it true that most students made no more than 10 mistakes? Explain your answer.

c. Which of the following values is closer to be the median of `mistakes`: 10, 20, 30, 40? Explain your answer.

Part 2

You should use R (RStudio) to solve problems in Part 2.

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.

a. Arrange them in a vector and call it `age`. Examine its type (numeric, character, etc). Provide your code as well as R outputs.

b. Calculate the following descriptive statistics for `age`: sample mean, sample median, sample variance and standard deviation. Provide your code as well as R outputs.

Problem 6

Here is a series of words:

pie, bar, bar, pie, pie, bar, bar, chart.

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

b. Calculate the relative frequences of values in `words` measured in percent. Provide both R code and outputs.

Problem 7

Plot a histogram of the vector `age` from Problem 5. It should contain 5 bins. Change a color to any you want. Provide your R code.