Linguistic Theory, 2018-2019
Linguistic Data: Quantitative Analysis and Visualisation
Seminar 1. Descriptive statistics. (12 January)
Olga Lyashevskaya, George Moroz, Alla Tambovtseva and Ilya Schurov

Problem 1. Consider a sample:

$$15,\ 20,\ 5,\ 8,\ 12,\ 7,\ 3$$

If you do not want to work with abstract numbers, suppose that we have a small research on languages in danger of extinction and this sample includes numbers of speakers that are very small.

(a) Find the average of this sample (the sample mean).
(b) Find the median of this sample.
(c) Find the sample variance and the standard deviation of this sample.
(d) Add a new value, 100, to this sample. Find the mean and the median of the updated sample. Make conclusions about properties of these two statistics.

Problem 2. Consider a small sample of real data on the Ethnic Fractionalization Index (by Fearon, for convenience values of the index were multiplied by 100):

Tanzania	95.3
DR Kongo	93.3
Ghana	84.6
South Africa	88.0
Uganda	93.0
Togo	88.3
India	81.1
Papua New Guinea	100
Malawi	82.9

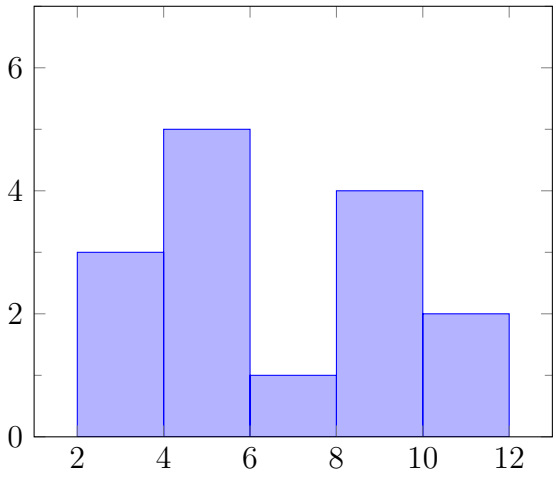
Take 81 as a starting point (the minimum value in the sample rounded for convenience) and plot a histogram for this sample:

(a) with the interval of width 1;
(b) with the interval of width 4;
(c) with the interval of width 10.

Which of three histograms seems to be the most sensible and adequate?

Note: To avoid ambiguity with values 88 and 93, assume that intervals are left-open: $(a, b]$. For example, 88 belongs to $(87, 88]$ and not to $(88, 89]$. This assumption is inherited in R histograms by default.

Problem 3. Look at the following histogram and answer the questions below.



(a) Find the proportion of values that are more than 10.
(b) Find the proportion of values that are more than 6, but less or equal 10.
(c) Show (approximately) where the median of such a sample can lie.