Homework 5

Olga Lyashevskaya, George Moroz, Alla Tambovtseva and Ilya Schurov

Deadline: 3 March, 23:59

In this assignment there are several tasks on

- data manipulation with `dplyr`
- data visualisation with `ggplot2`

All functions that you need are covered in class material. If you have need more detailed description, you can see this and this courses on `dplyr` and this course on `ggplot2`.

Rmd template for this homework is here.

1. Noun class assignment

These datasets contain the results of the experiment that evaluates the inter speaker variation in noun class assignment among speakers of the Zilo dialect of Andi (a Nakh-Daghestanian language). In Zilo there are two classes (b-class and r-class) for inanimate objects with no obvious semantic distinction between them.

There are two datasets:

- zilo_class_experiment_data.csv
- `w_id` — word id
- `stimulus` — stimuli word in Zilo
- `translation_en` — translation to English
- `stimulus_source` — the source of the stimulus: native or loan
- `1 ... 16` — columns that contain answers of 16 speakers of Zilo: `b` or `r` (class markers)
- zilo_class_experiment_informants.csv
- `s_id` — speaker id (corresponds to numbers in the previous dataset)
- `sex` — sex of the speaker
- `age_2017` — age of the speaker on the moment of the interview

1.1 ratio of b-words vs. sex and stimulus source

We are interested in the following question: how often speakers choose b-words depending on their sex and stimulus source? To give quantitative answer we want to calculate the ratio of b-words ($\frac{b\text{-words}}{b\text{-words}+r\text{-words}}$) used by all speakers of given sex for all stimulus of given source (e.g. for all female speakers and native stimulus source, and so on; four numbers in total). Then we want to visualize the resulting four numbers with point-plot, x-axis corresponding to `sex` and color to `stimulus_source`, see the picture in subproblem 1.1.6.

For your convenience, we splitted the problems into several subproblems. If you feel yourself brave, you can skip solving 1.1.1 — 1.1.6 and solve the problem as a whole, then write your full solution into 1.1.7. But if you are new to `dplyr`, we recommend step-by-step approach.

The plan

In `zilo_class_experiment_data.csv`, we have columns with labels 1 to 16 that contains class markers (`r` or `b`) for answers (words) used by a particular speaker for a particular stimulus. For every answer we want to know sex of the corresponding speaker and source of corresponding stimulus. Then we want to group all answers based on sex and stimulus source and calculate ratio of b-words in each group separately.

1.1.0 Read the data

Use `read_csv` to read the data from csv-files into dataframes.

1.1.1 Wide to long

In `zilo_class_experiment_data.csv` we have different columns for different speakers. We first need to convert this table to *long* format, i.e. replace every row of initial table with several rows, one for each speaker, and adding a new variable (column) that contains speaker id (call it `s_id`), and a new column that contains the class marker for answer (call it `answer`).

Your new table should look like this:

```
    w_id stimulus      translation_en stimulus_source s_id  answer
 1     1 milki         hous           native          1     b
 2     2 "va\u0261on"  train wagon    loan            1     b
 3     3 "in\u0261ur"  window         native          1     b
```

We will refer to this table as `long_class`.

Hint. We have columns which *names* are numbers. If we just use number to refer to a column, this number is considered as column number (i.e. 1 is a first column), not a name. One have to put quotes or backticks to refer to columns by names: `"1"` or `` `1` `` instead of `1`.

1.1.2 Join with informants dataset

Now every row in `long_class` corresponds to an answer of a particular informant for particular stimulus. We want to group these answers by informant's sexes, but information on a sex of a particular informant is stored in different dataset, `zilo_class_experiment_informants.csv`. We have to join `long_class` with `zilo_class_experiment_informants.csv` in such a way that every row will contain information about the informant (including sex). This can be done by a family of *join* commands: which one is needed here: `left_join`, `right_join`, `full_join`, `inner_join` or `anti_join`? On which variable should you join (i.e. which variable should have the same value for rows to be aligned?)

If you proceed directly with join, the following problem can occur: you can't join on `s_id` because of incompatible types (numeric / character). This is due to the fact that `s_id` in `zilo_class_experiment_informants.csv` can be considered as numeric variable by `read_csv`, but in `long_class`, `s_id` is character variable. You have to convert this variable in either of dataframe to match the other one. Use `mutate` with `as.numeric` or `as.character`.

Now your table should look like this:

```
    w_id stimulus      translation_en stimulus_source  s_id answer sex   age_2017
 1     1 milki         hous           native              1 b      f           15
 2     2 "va\u0261on"  train wagon    loan                1 b      f           15
 ...
```

1.1.3 Group and count

Now we want to count the number of b- and r-words within each stimulus_source for each speaker's sex. Use `count` here.

Your table should look like this:

```
  stimulus_source sex   answer   n
1 loan            f     b        254
2 loan            f     r        138
...
```

1.1.4 Long to wide

Now we want to find a ratio of r-answers for every stimulus_source and sex. Currently, for each pair of stimulus_source and sex, we have two rows, one corresponds to r-answer and another to b-answer. We want to convert our table to *wide* format, in such a way that these two rows are converted to one row with two new columns: count for r-answers and count for b-answers.

Your table should look like this:

```
  stimulus_source sex       b     r
1 loan            f       254   138
2 loan            m       238   154
3 native          f       215   241
...
```

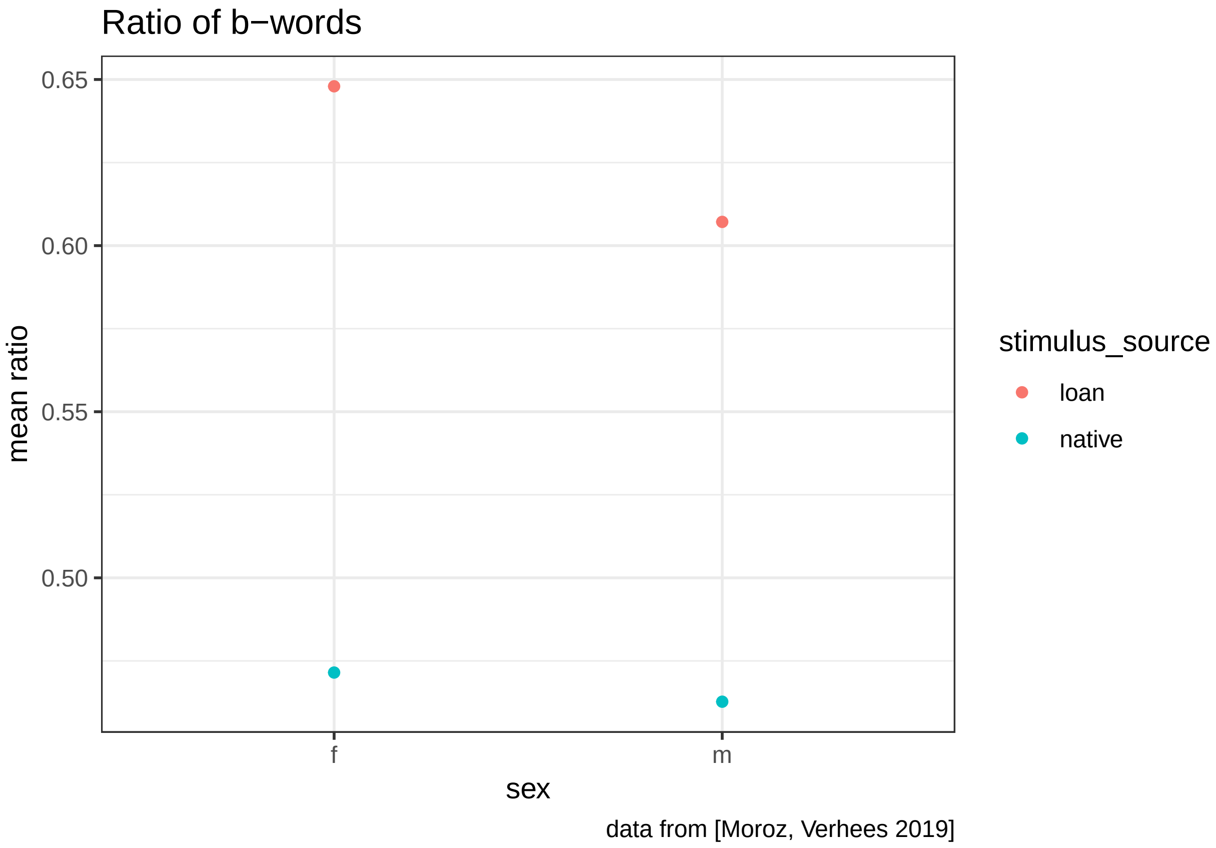
1.1.5 Find ratio

Now add a column called ratio that contains a ratio of b-words among all words for each row. Use mutate.

1.1.6 Visualize

Use ggplot to visualise obtained dataframe. Provide aes that use sex as values for x-axis, ratio as values for y-axis and stimulus_source as color value, than use geom_point layer. Also, use labs to add titles for axes and the whole image.

You have to obtain the following graph.

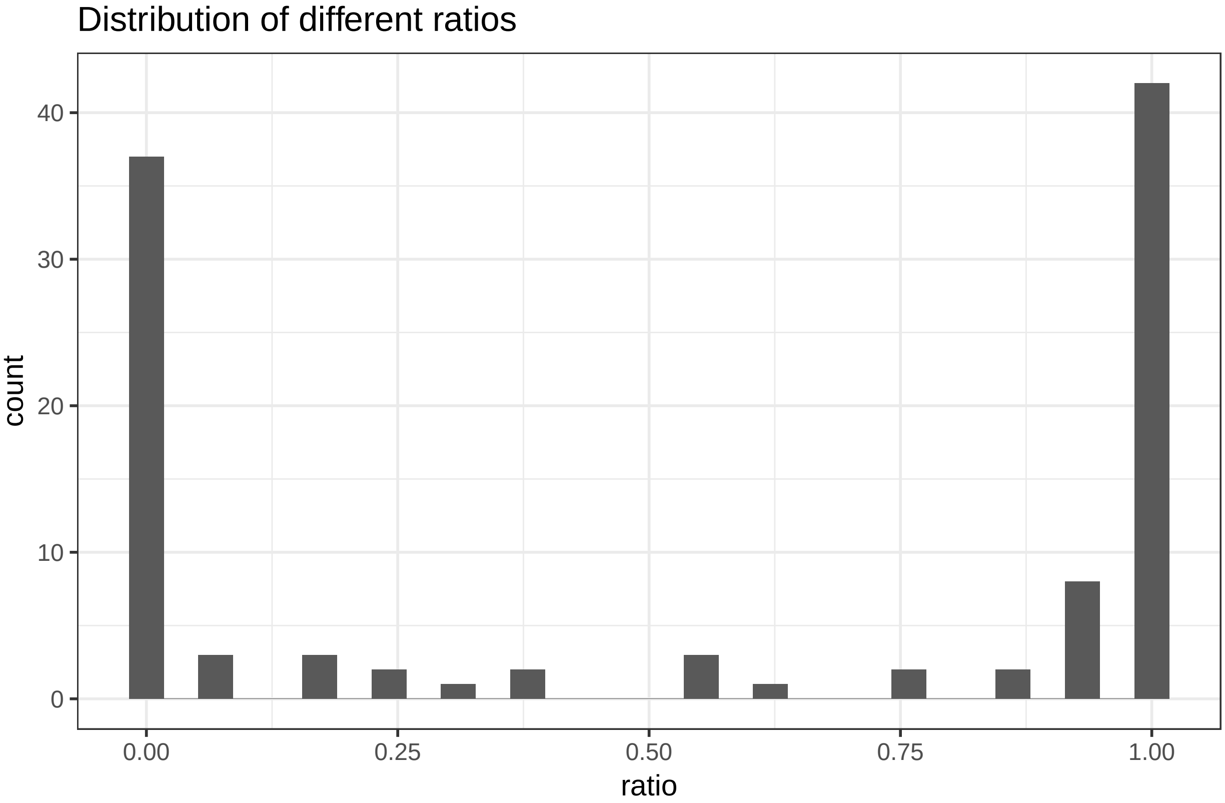


1.1.7 All together

You can write down all transformation needed to get the answer as a sequence of pipes. Do it here (optional):

1.2 ratio of b-words for each stimulus

Calculate the ratio of b-words ($\frac{b\text{-words}}{b\text{-words}+r\text{-words}}$) for each stimulus (i.e. each word) and visualise ratio distribution using `geom_histogram`. You should get this picture:



data from [Moroz, Verhees 2019]

1.3 stimuluae with given ratio

Calculate the ratio of b-words ($\frac{b\text{-words}}{b\text{-words}+r\text{-words}}$) for each stimulus. Which stimulae have value between 0.5 and 0.6? Use `filter`.

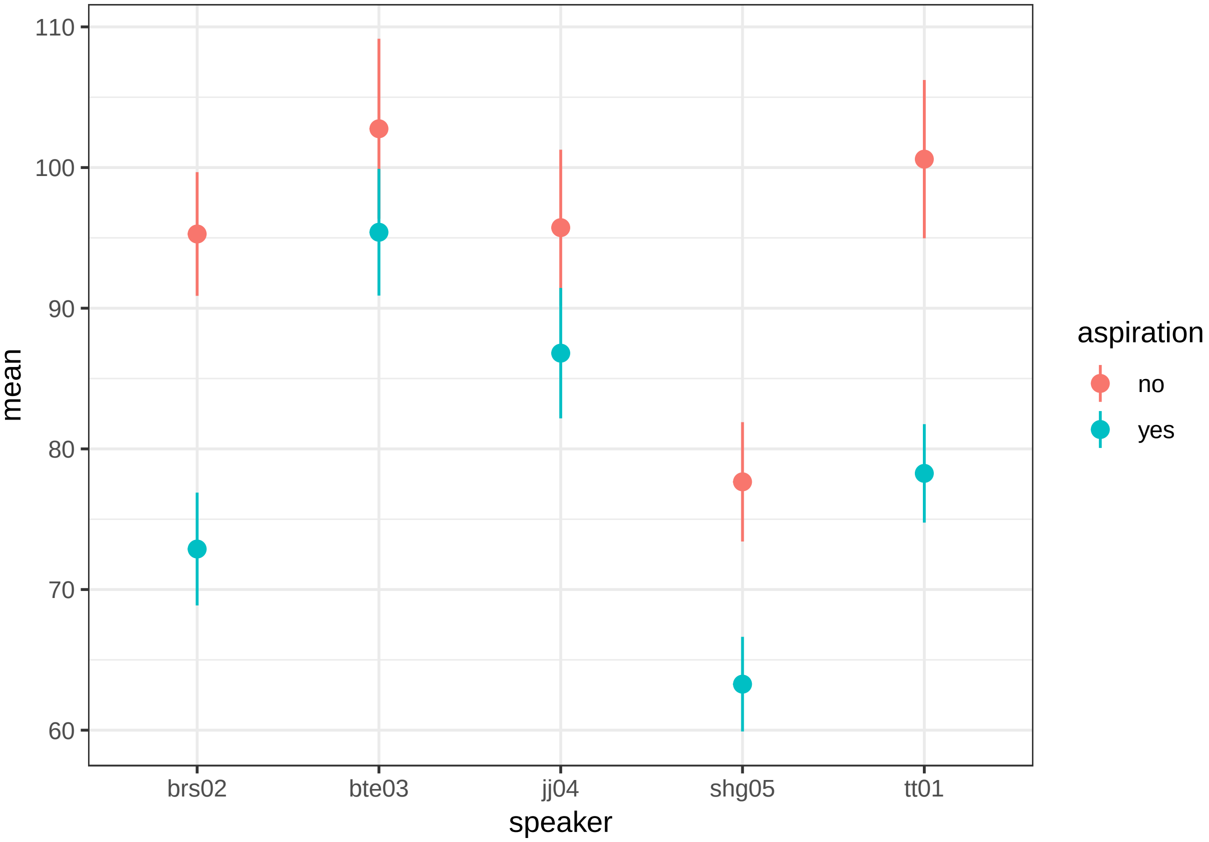
2. Vowel duration and aspiration

This dataset is based on (Coretta 2017, https://goo.gl/NrfgJm). This dissertation dealt with the relation between vowel duration and aspiration in consonants. Author carried out a data collection with 5 natives speakers of Icelandic. Then he extracted the duration of vowels followed by aspirated versus non-aspirated consonants.

Data are here.

2.1 Confidence intervals

Calculate a 95% confidence interval and a mean vowel duration (variable `vowel.dur`) for different contexts (variable `aspiration`) for all speakers (variable `speaker`) and visualise them using `geom_pointrange`. Use explicit formula (with `1.96` magic constant) we discussed in HW4 to find the upper and lower bounds of confidence intervals, then use `geom_point` and `geom_pointrage` to make a visualisation. You have to get this picture.


mean
brs02
bte03
jj04
shg05
tt01
speaker
aspiration
no
yes