February Survey Report

Maksim Zubok

2024-03-15

## Intro

To create weights, I am working with the 2020 [census data](https://github.com/antndlcrx/nonviolent-repression/blob/main/data/surveys/Tom3_tab1_VPN-2020.xlsx), particulary the cross tabbed gender, age, and university education file [here](https://github.com/antndlcrx/nonviolent-repression/blob/main/data/surveys/ru_population_frame.csv).

I did not include information about region of residence, even though we could do it after harmonising census data with the survey.

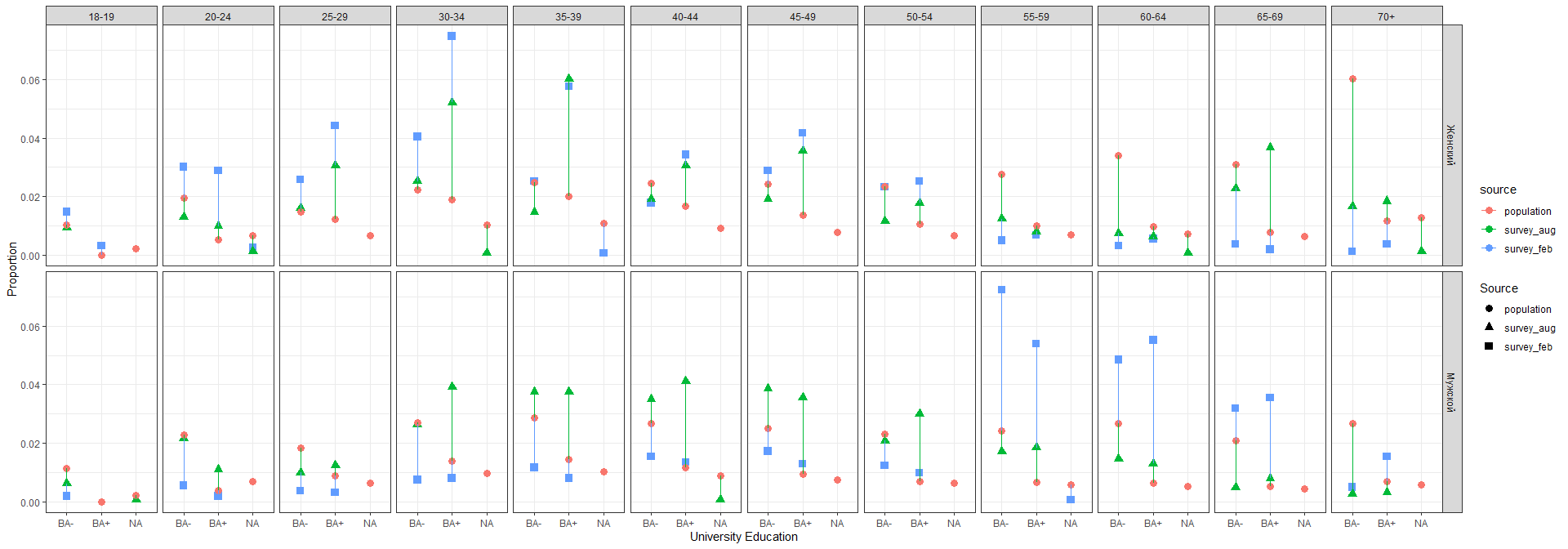
head(ru\_population\_frame, 5)

## # A tibble: 5 × 4  
## gender age\_group university\_education Freq  
## <fct> <fct> <chr> <dbl>  
## 1 Мужской 18-19 BA+ 1  
## 2 Мужской 18-19 BA- 1367504  
## 3 Мужской 18-19 NA 247138  
## 4 Мужской 20-24 BA+ 463546  
## 5 Мужской 20-24 BA- 2727980

Note that because in the survey we have a handful of people aged 18-19 who reported having BA education and because “survey” package does not permit 0 in the population frame, I put 1 in Freq for 18-19 men and women BA+ intersection.

## Sample to Population Comparison

## `summarise()` has grouped output by 'gender', 'age\_group'. You can override  
## using the `.groups` argument.  
## `summarise()` has grouped output by 'gender', 'age\_group'. You can override  
## using the `.groups` argument.



The main disparities between Survey Feb and Population are:

* oversampled young women and undersampled young men, especially age 30-39 for women with university education.
* the opposite for old people. Oversampled older men, undersampled older women without university education.

## Weights witth Survey package

## survey library ##  
unweighted\_data <- svydesign(ids = ~1, data = survey\_feb)

## Warning in svydesign.default(ids = ~1, data = survey\_feb): No weights or  
## probabilities supplied, assuming equal probability

weighted <- postStratify(unweighted\_data, ~age\_group + gender + university\_education,  
 ru\_population\_frame, partial=TRUE)

## Warning in postStratify.survey.design(unweighted\_data, ~age\_group + gender + :  
## Some strata absent from sample: ignored

summary(weights(weighted))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 18588 33196 62445 52625 3604783

Note: some stratas had no observations in the survey (NA on education for some age gender groups). This means we had to ignore them in producing weights.

Also the weights for women with no uni education aged 70+ seem to be too large. It is not surprising given the disparity, but noteworthy.

