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.

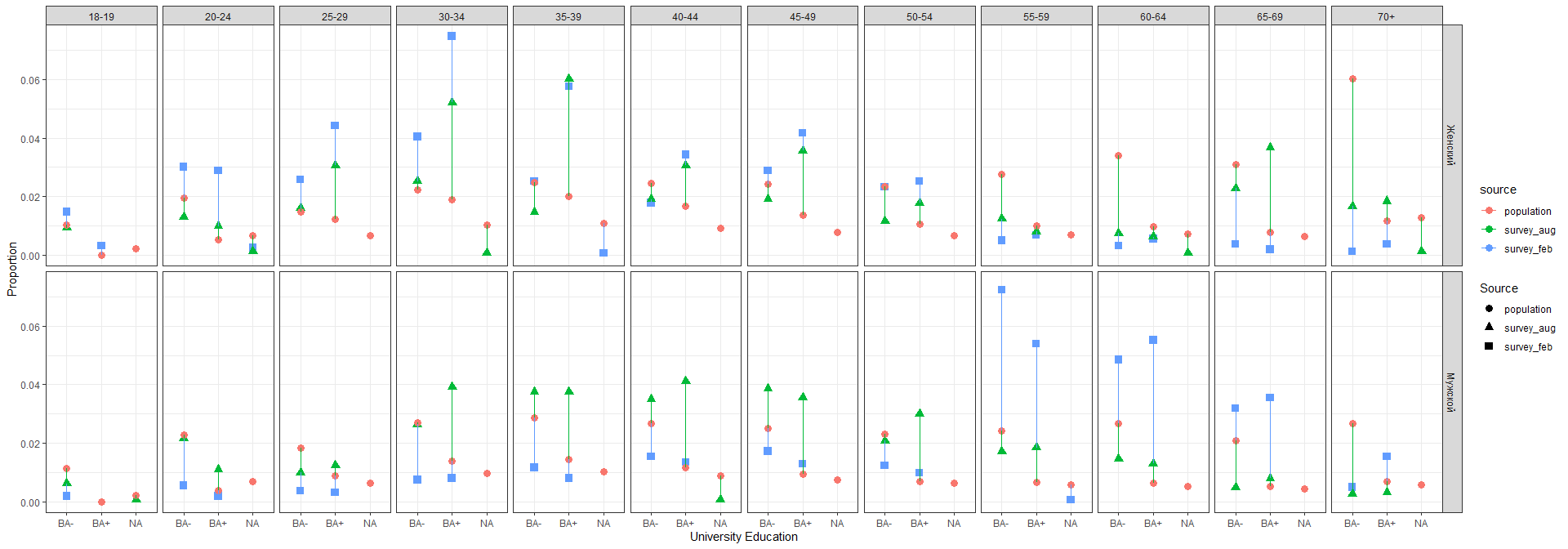
head(ru\_population\_frame, 5)

## # A tibble: 5 x 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

To compute post-stratification weights we rely on the postStratify function from the survey package. The function adjusts the sampling and replicate weights so that the joint distribution of a set of post-stratifying variables matches the known population joint distribution. **However, the package documentation does not describe how exactly the adjustment is implemented.**

## survey library ##  
unweighted\_data <- svydesign(ids = ~1, data = survey\_feb)  
  
  
weighted <- postStratify(unweighted\_data, ~age\_group + gender + university\_education,  
 ru\_population\_frame, partial=TRUE)  
  
# save weights   
survey\_feb$weight <- weights(weighted)  
sum\_feb <- summary(weights(weighted))  
sum\_feb

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

Note: some strata 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.

## survey library ##  
unweighted\_data <- svydesign(ids = ~1, data = survey\_aug)  
  
  
weighted <- postStratify(unweighted\_data, ~age\_group + gender + university\_education,  
 ru\_population\_frame, partial=TRUE)  
  
# save weights   
survey\_aug$weight <- weights(weighted)  
sum\_aug <- summary(weights(weighted))  
sum\_aug

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15425 25753 43376 65425 77942 1216293

For August, we also see that some weights are much larger than others. As you can see in the graphs below, the distribution of weights is similarly skewed and the disparities between the bulk of the distribution and its tales are in the same orders of magnitude. However, the largest weight in Feb survey is three times bigger than the largest weight in Aug survey.

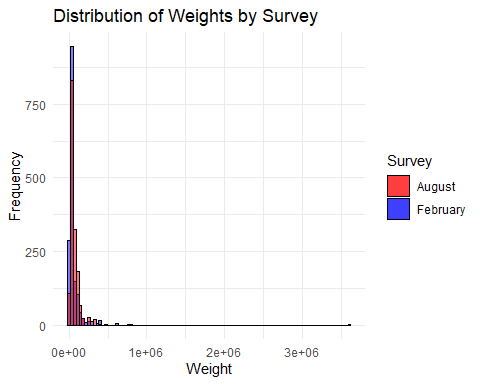
The largest weights in both surveys relate to different population groups.

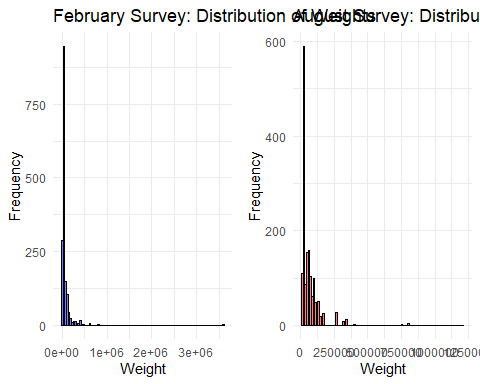
February Survey, Top Five Rows by Weight

| age\_group | gender | university\_education | weight |
| --- | --- | --- | --- |
| 70+ | Женский | BA- | 3604783.0 |
| 70+ | Женский | BA- | 3604783.0 |
| 35-39 | Женский | NA | 1302931.0 |
| 60-64 | Женский | BA- | 811229.8 |
| 60-64 | Женский | BA- | 811229.8 |

August Survey, Top Five Rows by Weight

| age\_group | gender | university\_education | weight |
| --- | --- | --- | --- |
| 30-34 | Женский | NA | 1216293 |
| 40-44 | Мужской | NA | 1052747 |
| 60-64 | Женский | NA | 856270 |
| 70+ | Мужской | BA- | 797236 |
| 70+ | Мужской | BA- | 797236 |





## Weights created manually

To check the plausibility of resulting weights, we create alternative weights based on the population frequencies of the combination of the same strata (Yana’s approach). The weights are calculated for each category:

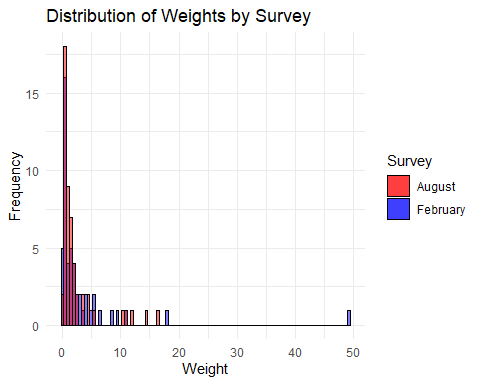
# calculate weights and print the df with weights  
weights\_aug\_strata <- left\_join(survey\_aug\_strata,  
 pop\_strata,  
 c("gender", "age\_group", "university\_education")) %>%   
 rename(population\_proportion = proportion.y,  
 sample\_proportion = proportion.x) %>%   
 # calculate weights as popul prop/sample prop  
 mutate(weight = population\_proportion / sample\_proportion)  
  
sum\_man\_aug <-summary(weights\_aug\_strata$weight)  
sum\_man\_aug

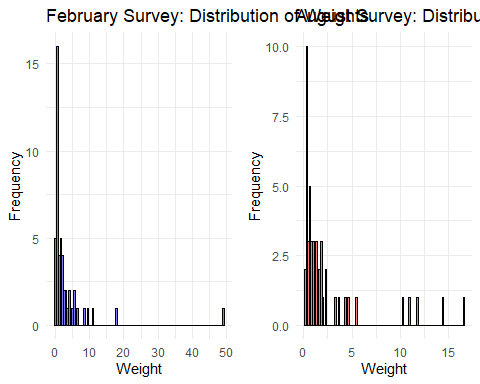
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2105 0.5255 1.0878 2.4029 1.9037 16.5951

# calculate weights and print the df with weights  
weights\_feb\_strata <- left\_join(survey\_feb\_strata,  
 pop\_strata,  
 c("gender", "age\_group", "university\_education")) %>%   
 rename(population\_proportion = proportion.y,  
 sample\_proportion = proportion.x) %>%   
 # calculate weights as popul prop/sample prop  
 mutate(weight = population\_proportion / sample\_proportion)  
  
# head(weights\_feb\_strata)  
sum\_man\_feb <- summary(weights\_feb\_strata$weight)  
sum\_man\_feb

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.5014 1.1922 3.3786 3.0920 49.1836

## Adding missing grouping variables: `gender`, `age\_group`  
## Adding missing grouping variables: `gender`, `age\_group`





Calculating weights this way gives us the same overall picture. The main difference is in the orders of magnitude between the median weight and the tail (highest weight).

For **February** survey, the difference (max weight / median weight) with postStratify is **108.59**, but with simple manually created weights it is **41.26**. For **August** survey, the difference with postStratify is **28.04**, but with simple manually created weights it is **15.26**.

Note that some observations in February survey have a weight of zero because they do not exist in the population census. These are women 18-19 years old with university education. My suspicion on why we have these observations is that some respondents might have misreported their education attainment.