Report on Survey Weights

Maksim Zubok

September 18, 2024

**Contents**

[Introduction 2](#_Toc43817)

[Sample to Population Comparison 2](#_Toc43818)

[Weights with Survey package 4](#_Toc43819)

[Weights created manually 8](#_Toc43820)

[Survey July Randomisation check 10](#_Toc43821)

[Survey July Missingness 14](#_Toc43822)

# Introduction

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

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

Table 1: Population Frame: Census 2020

|  |  |  |  |
| --- | --- | --- | --- |
| gender | age\_group | university\_education | Freq |
| Man | 18-24 | BA+ | 463546 |
| Man | 18-24 | BA- | 5164261 |
| Man | 25-34 | BA+ | 2707302 |
| Man | 25-34 | BA- | 7287296 |
| Man | 35-44 | BA+ | 3136411 |

# Sample to Population Comparison

To compare our samples to the population, we do two comparisons. First, we compare them to a nationally representative Levada omnibus survey which was fielded in March 2024. Second, we compare them to the 2020 Census.

The Table 2 compares the demographic composition of Qualtrics samples collected in August, February, and March to nationally representative Levada Omnibus survey.

Table 2: Comparison of Category Shares by Variable

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Values | Levada (N=1628) | Aug 23  (N=1600) | Feb 24 (N=1630) | March 24 (N=1630) | July 24 (N=1416) | Sept 24 (N=922) |
| Age | 18-24 | 0.08 | 0.07 | 0.09 | 0.09 | 0.09 | 0.20 |
| Age | 25-34 | 0.17 | 0.22 | 0.21 | 0.21 | 0.20 | 0.26 |
| Age | 35-44 | 0.23 | 0.28 | 0.18 | 0.18 | 0.19 | 0.23 |
| Age | 45-54 | 0.15 | 0.21 | 0.17 | 0.17 | 0.17 | 0.17 |
| Age | 55-64 | 0.18 | 0.10 | 0.25 | 0.20 | 0.21 | 0.09 |
| Age | 65+ | 0.19 | 0.12 | 0.10 | 0.15 | 0.15 | 0.04 |
| Gender | Man | 0.44 | 0.49 | 0.45 | 0.45 | 0.45 | 0.45 |
| Gender | Woman | 0.56 | 0.51 | 0.55 | 0.55 | 0.55 | 0.55 |
| Education | BA+ | 0.28 | 0.57 | 0.54 | 0.54 | 0.49 | 0.56 |
| Education | BA- | 0.72 | 0.43 | 0.46 | 0.46 | 0.51 | 0.44 |

The Figure 1 depicts strata shares compared to the 2020 Census.

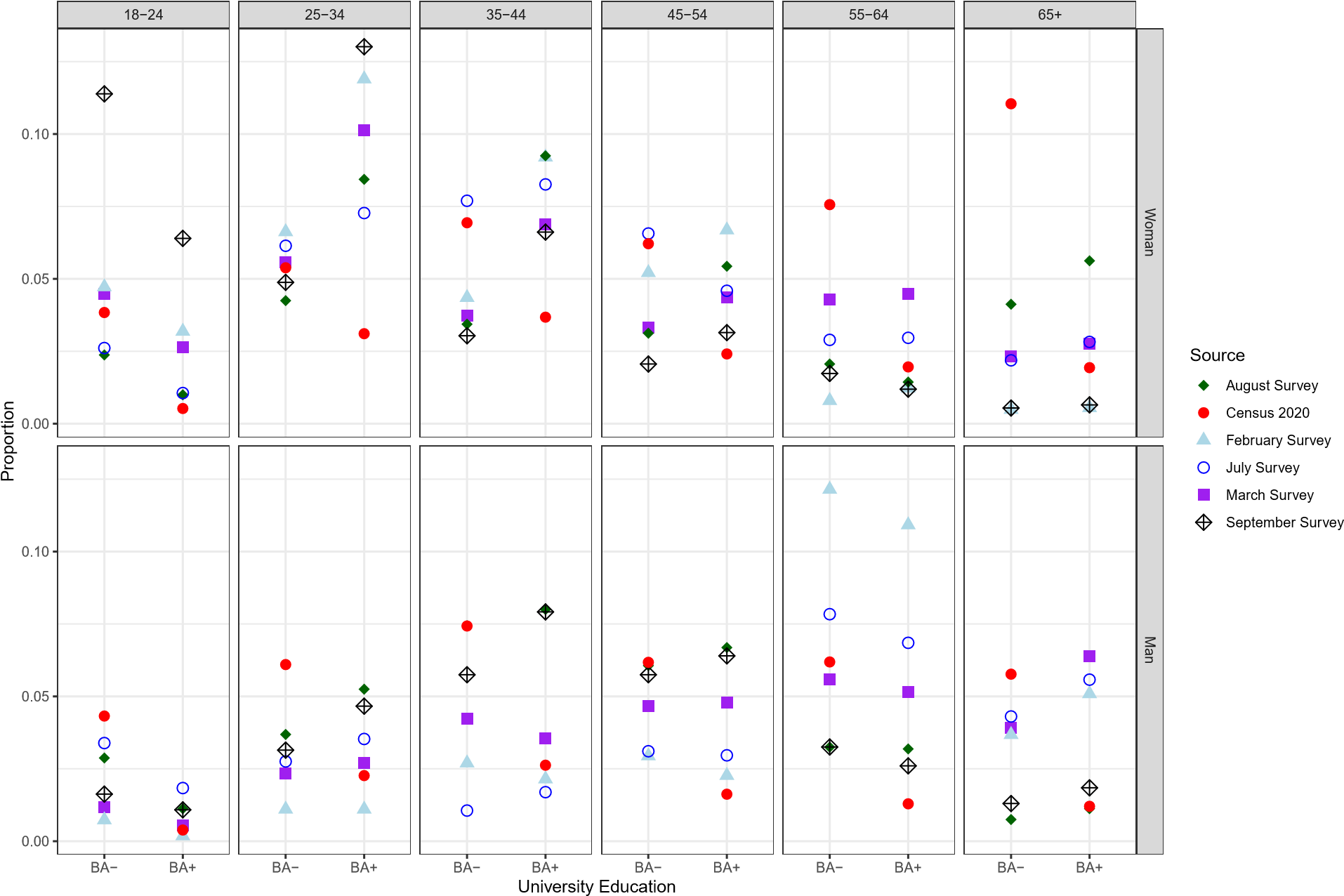


Figure 1: Demographic Comparison on Census 2020 The main observations are:

* The biggest disparities are is women 55-64 and 65+ without education. For instance, the February survey has 0 women in the latter category.
* We have oversampled young women, especially 25-34 years old with university education, and undersampled men without university education across all age categories except 55- and 65+.

# Weights with 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.**

## March

Table 3: March Survey PostStratify Weights Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| Value | 13863.25 | 24875.41 | 54076.05 | 73292.34 | 107672 | 347290.3 |

The table below explores shows strata that were assigned the highest weight.

Table 4: March Survey, Top Five Rows by Weight

|  |  |  |  |
| --- | --- | --- | --- |
| age\_group | gender | university\_education | weight\_poststratify |
| 65+ | Woman | BA- | 347290.3 |
| 18-24 | Man | BA- | 271803.2 |
| 25-34 | Man | BA- | 191770.9 |
| 45-54 | Woman | BA- | 137495.7 |
| 35-44 | Woman | BA- | 135884.6 |

## February

Table 5: February Survey PostStratify Weights Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| Value | 8670.35 | 19144.08 | 37348.11 | 73292.34 | 87350.22 | 1649629 |

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.

Table 6: February Survey, Top Five Rows by Unique Weight

|  |  |  |  |
| --- | --- | --- | --- |
| age\_group | gender | university\_education | weight\_poststratify |
| 65+ | Woman | BA- | 1649629.1 |
| 55-64 | Woman | BA- | 695189.1 |
| 18-24 | Man | BA- | 430355.1 |
| 25-34 | Man | BA- | 404849.8 |
| 65+ | Woman | BA+ | 256902.0 |

## August

Table 7: August Survey PostStratify Weights Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| Value | 18133.48 | 27510.75 | 33083.25 | 74666.57 | 112266.5 | 574250.5 |

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.

Table 8: August Survey, Top Five Rows by Weight

|  |  |  |  |
| --- | --- | --- | --- |
| age\_group | gender | university\_education | weight\_poststratify |
| 65+ | Man | BA- | 574250.5 |
| 55-64 | Woman | BA- | 273862.4 |
| 65+ | Woman | BA- | 199955.0 |
| 35-44 | Woman | BA- | 150708.4 |
| 45-54 | Woman | BA- | 148495.4 |

## July

Table 9: August Survey PostStratify Weights Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| Value | 15910.55 | 37565.66 | 66620.95 | 84369 | 79836.23 | 591827.9 |

As with previous surveys, the July survey’s distribution of weights is also skewed and the disparities between the bulk of the distribution and its tales are in the same orders of magnitude. The largest weight (and biggest disparity to population census 2020 is men without university education aged 35-44).

Table 10: July Survey, Top Five Rows by Weight

|  |  |  |  |
| --- | --- | --- | --- |
| age\_group | gender | university\_education | weight\_poststratify |
| 35-44 | Man | BA- | 591827.9 |
| 65+ | Woman | BA- | 425710.7 |
| 55-64 | Woman | BA- | 220425.8 |
| 25-34 | Man | BA- | 186853.7 |
| 45-54 | Man | BA- | 167718.7 |

## September

Table 11: August Survey PostStratify Weights Summary

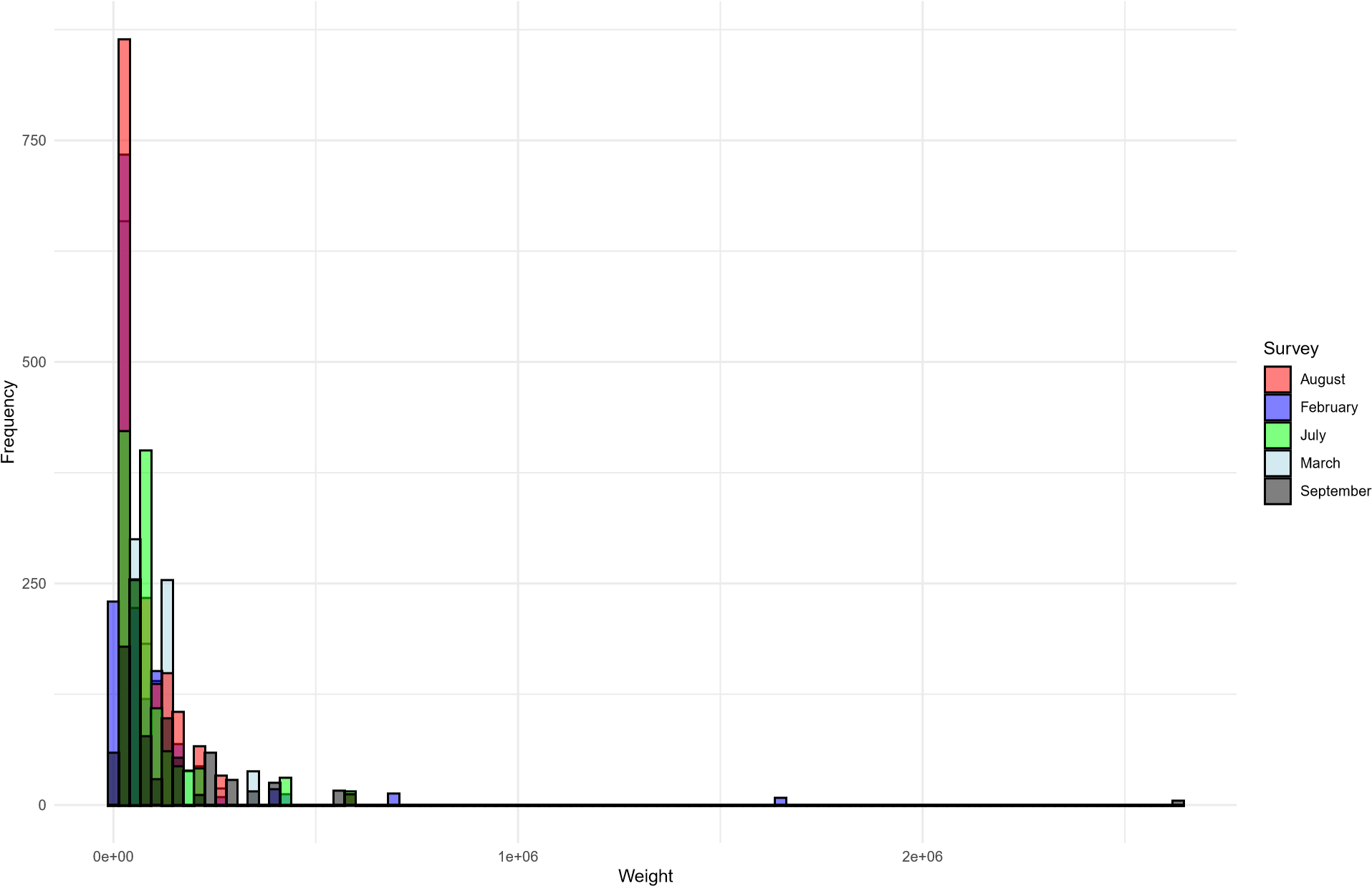
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| Value | 10652.44 | 32886.14 | 62960.51 | 129573.2 | 142934.5 | 2639407 |

The September survey is skewed towards younger people, the share of older people in the survey is much smaller than in the population as recorded by the 2020 census. People aged 65+ have the highest survey weight.

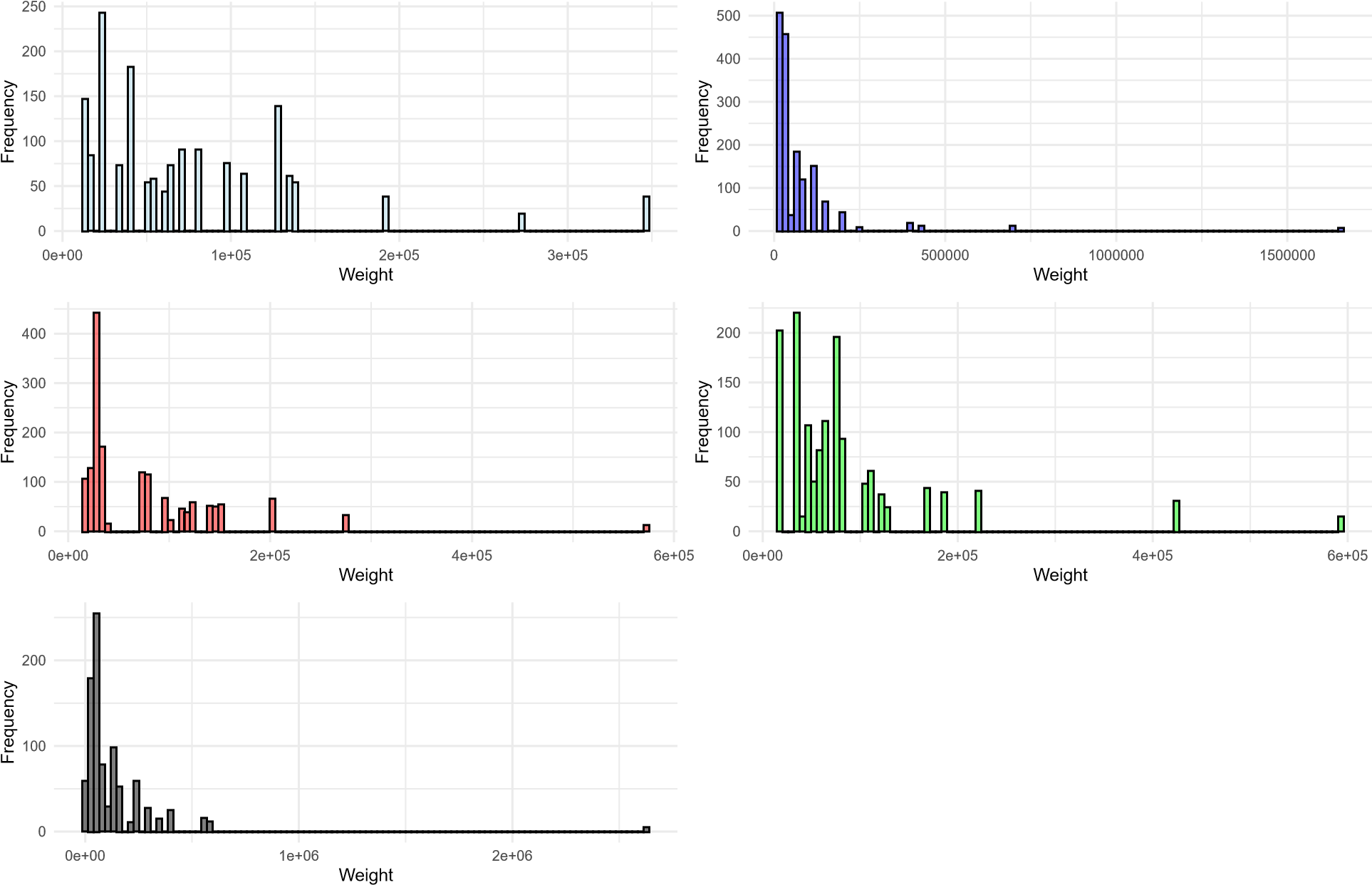
Table 12: September Survey, Top Five Rows by Weight

|  |  |  |  |
| --- | --- | --- | --- |
| age\_group | gender | university\_education | weight\_poststratify |
| 65+ | Woman | BA- | 2639406.6 |
| 65+ | Man | BA- | 574250.5 |
| 55-64 | Woman | BA- | 564841.1 |
| 45-54 | Woman | BA- | 390777.3 |
| 65+ | Woman | BA+ | 385353.0 |

Distribution of Weights by Survey



Distribution of Weights Across Months



# 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: population frequency weight𝑖 = sample frequency𝑖 𝑖

**March**

Table 13: March Survey Weights Summary

|  |  |  |
| --- | --- | --- |
| Min. 1st Qu. Median Mean 3rd Qu. | | Max. |
| Value 0.19 0.51 0.85 1.24 1.76 | | 4.74 |
| Table 14: February Survey Weights Summary | |  |
|  | Min. 1st Qu. Median Mean 3rd Qu. | Max. |
| Value | 0.12 0.48 1.39 2.81 2.27 | 22.51 |
|  | Table 15: August Survey Weights Summary |  |
|  | Min. 1st Qu. Median Mean 3rd Qu. | Max. |
| Value | 0.24 0.4 1.05 1.43 1.72 | 7.69 |
|  | Table 16: July Survey Weights Summary |  |
|  | Min. 1st Qu. Median Mean 3rd Qu. | Max. |
| Value | 0.19 0.52 0.83 1.38 1.49 | 7.01 |

## February

**August**

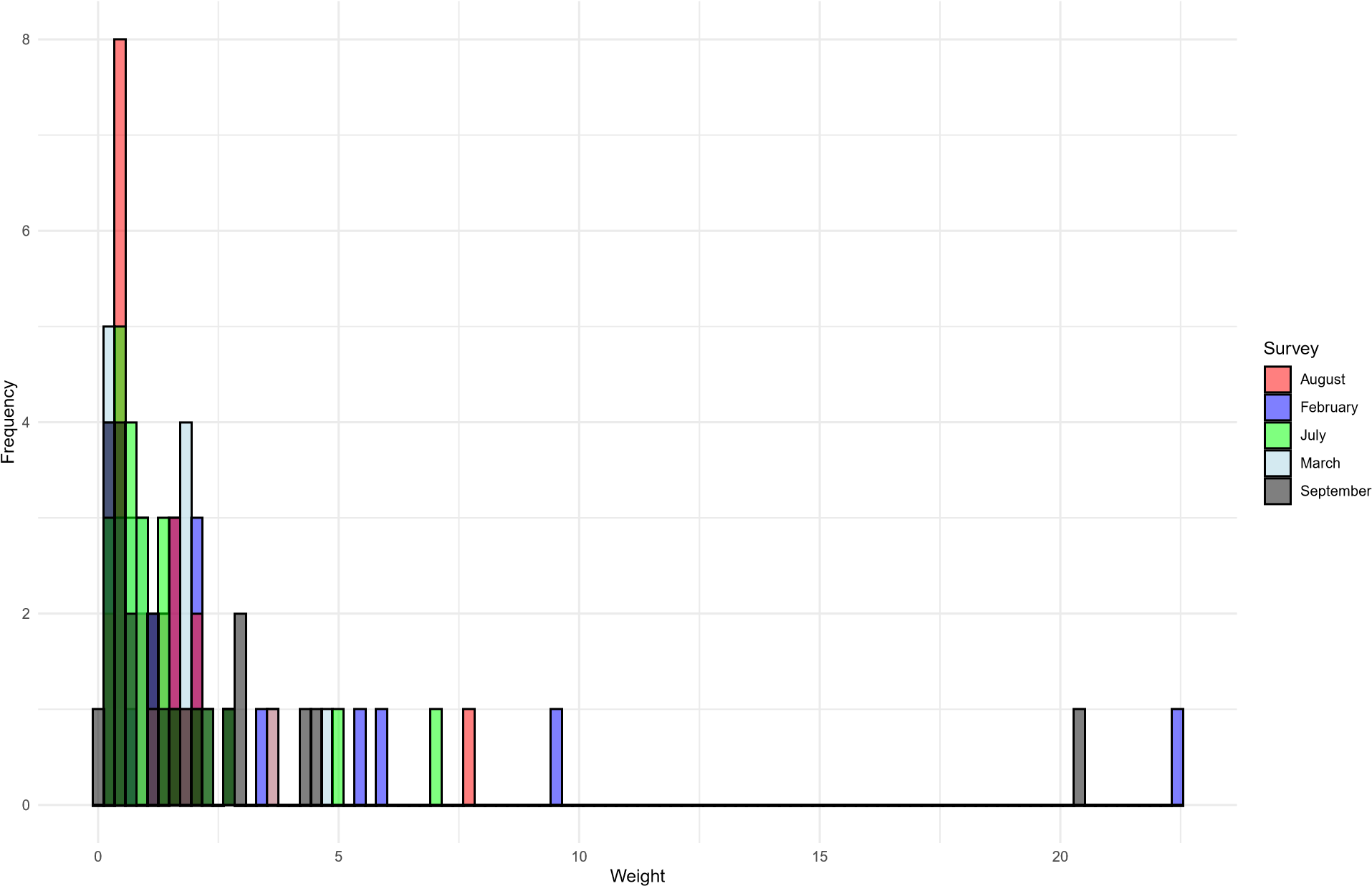
**July**

## September

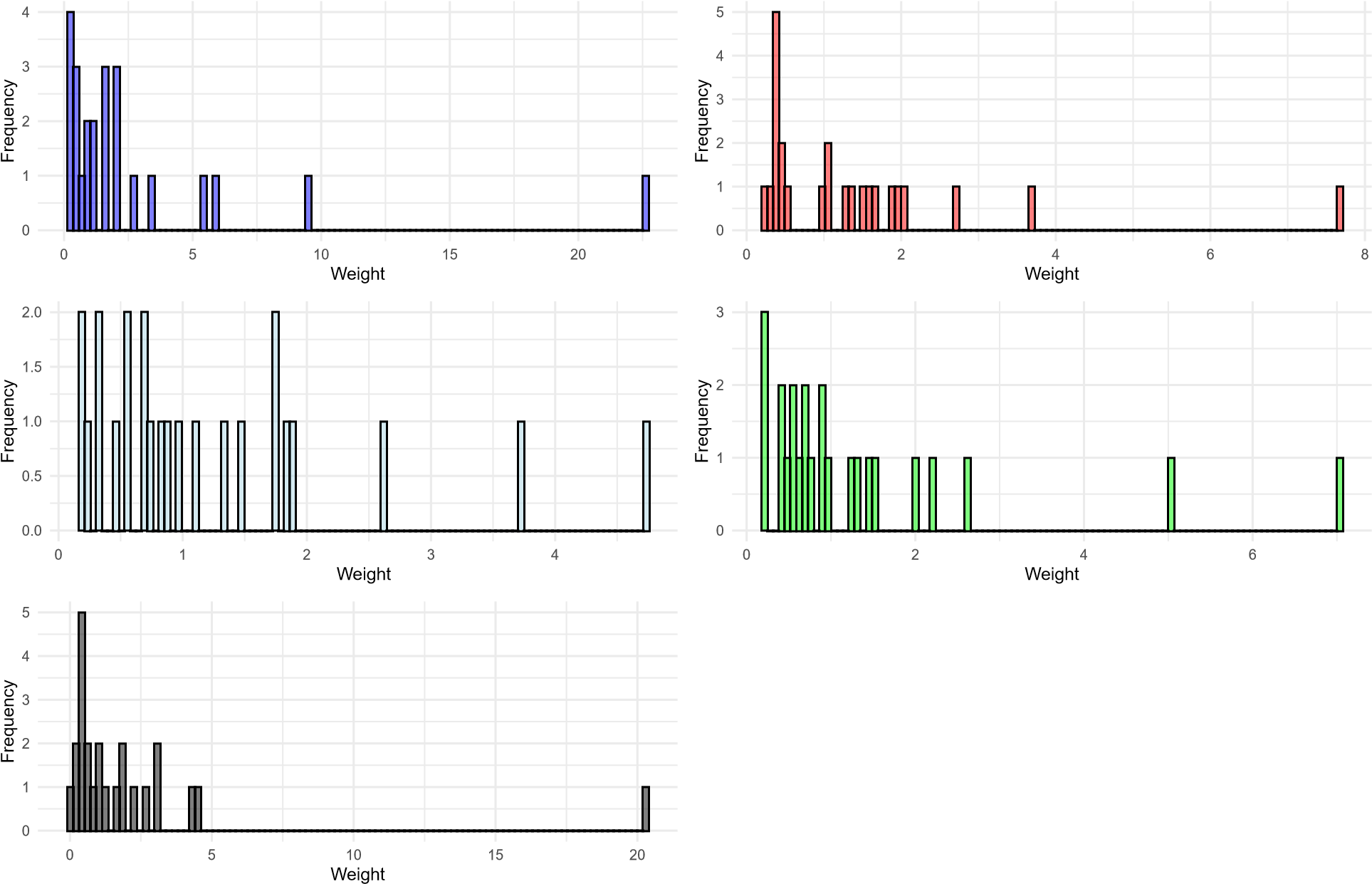
Table 17: September Survey Weights Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| Value | 0.08 | 0.45 | 1.09 | 2.23 | 2.38 | 20.37 |

Distribution of Weights by Survey



Distribution of Weights Across Months



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 **44.17**, but with simple manually created weights it is **16.19**. For **August** survey, the difference with postStratify is **17.36**, but with simple manually created weights it is **7.32**.

# Survey July Randomisation check

To check for randomisation, I first create a categorical variable that captures treatment assignment:

survey\_july <- survey\_july %>%

mutate(

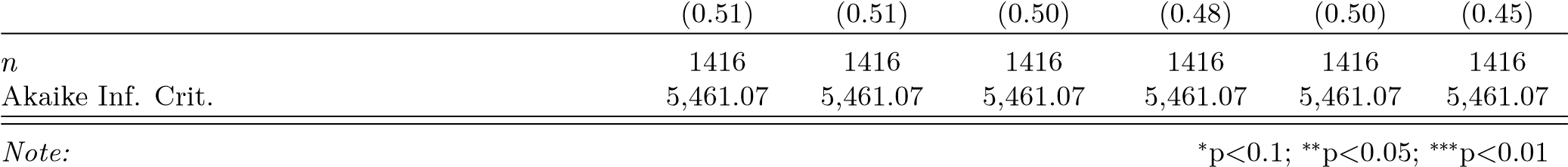
treatment1 = case\_when(!is.na(DV1A) ~ "A", !is.na(DV1B) ~ "B", !is.na(DV1C) ~ "C", !is.na(DV1D) ~ "D", !is.na(DV1E) ~ "E", !is.na(DV1F) ~ "F", !is.na(DV1H) ~ "I", !is.na(DV1I) ~ "I",

TRUE ~ NA\_character\_)

Then, I run a multinomial logit to see if the odds of a treatment category assignment relative to control (or first treatment category) can be predicted as a function of the base demographic controls, namely age, gender, university education, and size of a city of residence.

As I see from the tables below, we do not have any violation of random assignment, as indicated by the absence of any relationships between treatment assignment and demographic variables that can be statistically different from zero.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| age | −0.02∗∗ | −0.01 | −0.02∗∗ | 0.003 | −0.01 | −0.01 |
|  | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| genderMan | −0.31 | 0.11 | −0.05 | −0.59∗∗∗ | −0.08 | −0.24 |
|  | (0.22) | (0.22) | (0.22) | (0.23) | (0.22) | (0.19) |
| university\_educationBA+ | −0.39∗ | −0.41∗ | −0.09 | −0.29 | −0.30 | −0.18 |
|  | (0.22) | (0.22) | (0.22) | (0.22) | (0.22) | (0.19) |
| urbancity\_250\_500\_thousand | 0.01 | −0.10 | −0.48 | −0.93∗∗ | −0.57 | −0.13 |
|  | (0.46) | (0.45) | (0.46) | (0.43) | (0.46) | (0.40) |
| urbancity\_500\_thousand\_to\_1\_million | 0.12 | 0.18 | −0.08 | −0.57 | 0.15 | 0.28 |
|  | (0.46) | (0.45) | (0.45) | (0.42) | (0.44) | (0.40) |
| urbancity\_more\_than\_1\_million | −0.07 | −0.44 | −0.30 | −0.91∗∗ | −0.55 | −0.12 |
|  | (0.43) | (0.43) | (0.42) | (0.39) | (0.42) | (0.37) |
| urbancity\_or\_village\_less\_than\_100\_thousand | −0.18 | −0.47 | −0.21 | −0.78∗∗ | −0.50 | 0.05 |
|  | (0.43) | (0.43) | (0.42) | (0.39) | (0.42) | (0.37) |
| urbandifficult\_to\_answer | −1.13 | 0.33 | −0.44 | −1.64 | 0.41 | −0.20 |
|  | (1.28) | (0.94) | (1.07) | (1.27) | (0.91) | (0.93) |
| urbanMoscow | 0.11 | 0.03 | −0.25 | −0.96∗∗ | −0.18 | 0.01 |
|  | (0.47) | (0.46) | (0.47) | (0.45) | (0.46) | (0.41) |
| urbanno\_answer | −0.55 | −12.56∗∗∗ | −12.87∗∗∗ | −13.48∗∗∗ | −0.57 | −0.94 |
|  | (1.47) | (0.0000) | (0.0000) | (0.0000) | (1.47) | (1.46) |
| urbanSaint\_Petersburg | −0.91 | −0.17 | −0.38 | −1.43∗∗ | −0.35 | −0.12 |
|  | (0.65) | (0.55) | (0.56) | (0.59) | (0.54) | (0.48) |
| Constant | 1.21∗∗ | 0.83 | 1.14∗∗ | 1.01∗∗ | 0.93∗ | 1.25∗∗∗ |

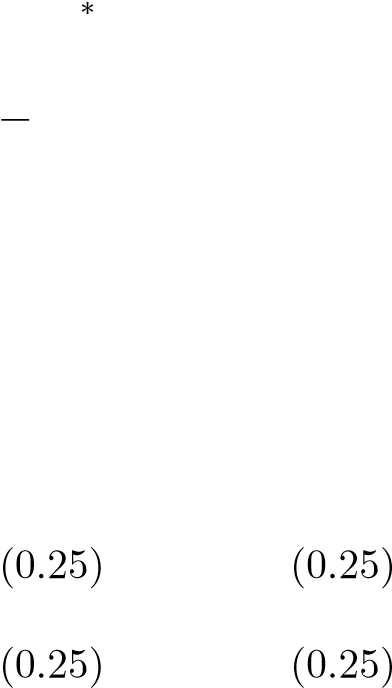
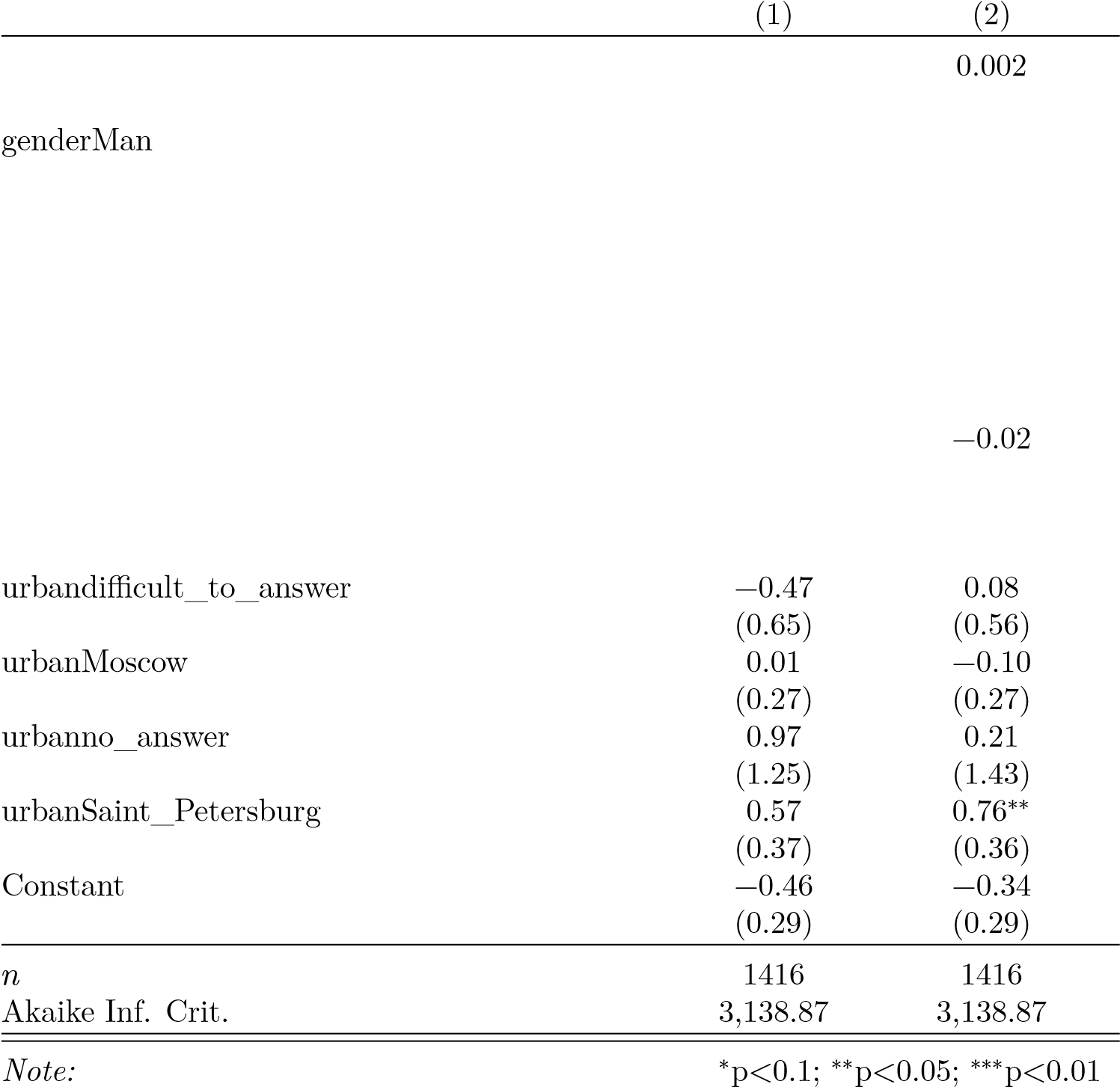
Table 18: Randomisation Check Protest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| B |  | *Dependent variable:* |  |  |
| C | D E | F | I |
| (1) | (2) | (3) (4) | (5) | (6) |

Table 19: Randomisation Check Education Experiment 1

*Dependent variable:*

B C

 age 0.01

(0.005) (0.004)

0.16 0.05

(0.14) (0.14)

university\_educationBA+ 0.07 0.20

(0.13) (0.13)

urbancity\_250\_500\_thousand 0.05 0.13

(0.27) (0.27)

urbancity\_500\_thousand\_to\_1\_million 0.11 0.01

(0.26) (0.26)

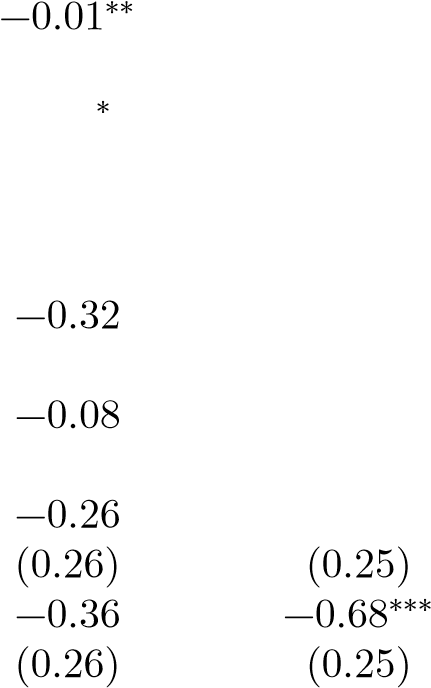
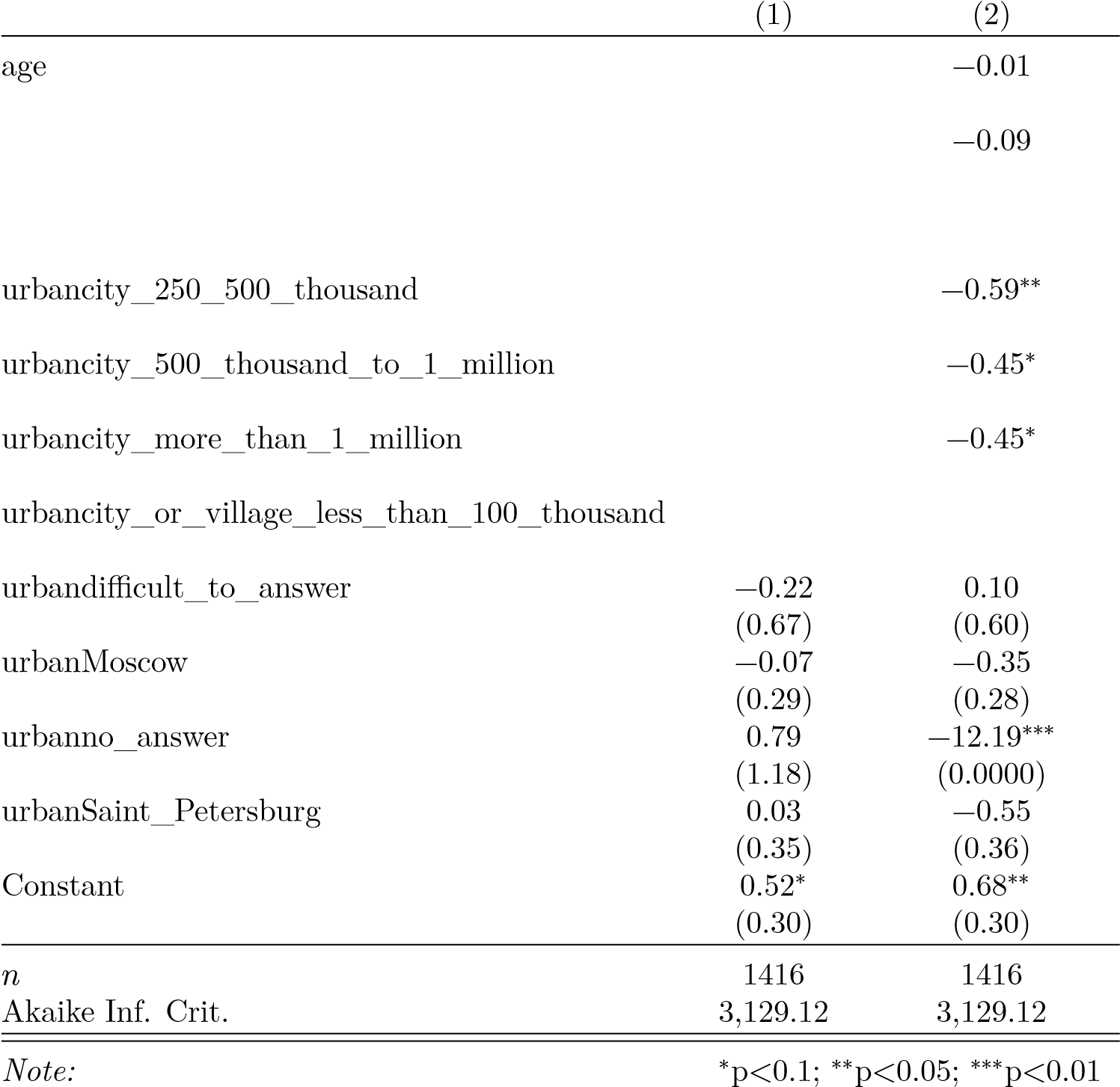
urbancity\_more\_than\_1\_million 0.22

urbancity\_or\_village\_less\_than\_100\_thousand 0.30 0.32

Table 20: Randomisation Check Education Experiment 2

*Dependent variable:*

B C

 (0.004) (0.005)

genderMan 0.25

(0.14) (0.14)

university\_educationBA+ 0.06 0.13

(0.13) (0.13)

(0.28) (0.27)

(0.27) (0.26)

# Survey July Missingness

To check for missing responces, I pull together all treatment variables of the same experiment and treat answers “I do not know” and “refuse to answer” as missing. Then, I estimate if base control demographic variables predict missingness in the experiment. The table below shows the results.

As I see from the table, younger people, women, and people who could not answer about the population size of their place of residence are less likely to answer. People with university education are more likely to provide an answer to the experiments questions.

Table 21: Missings Check Experiments

*Dependent variable:*

missing\_treatment1 missing\_treatment2 missing\_treatment3

