What are survey weights?

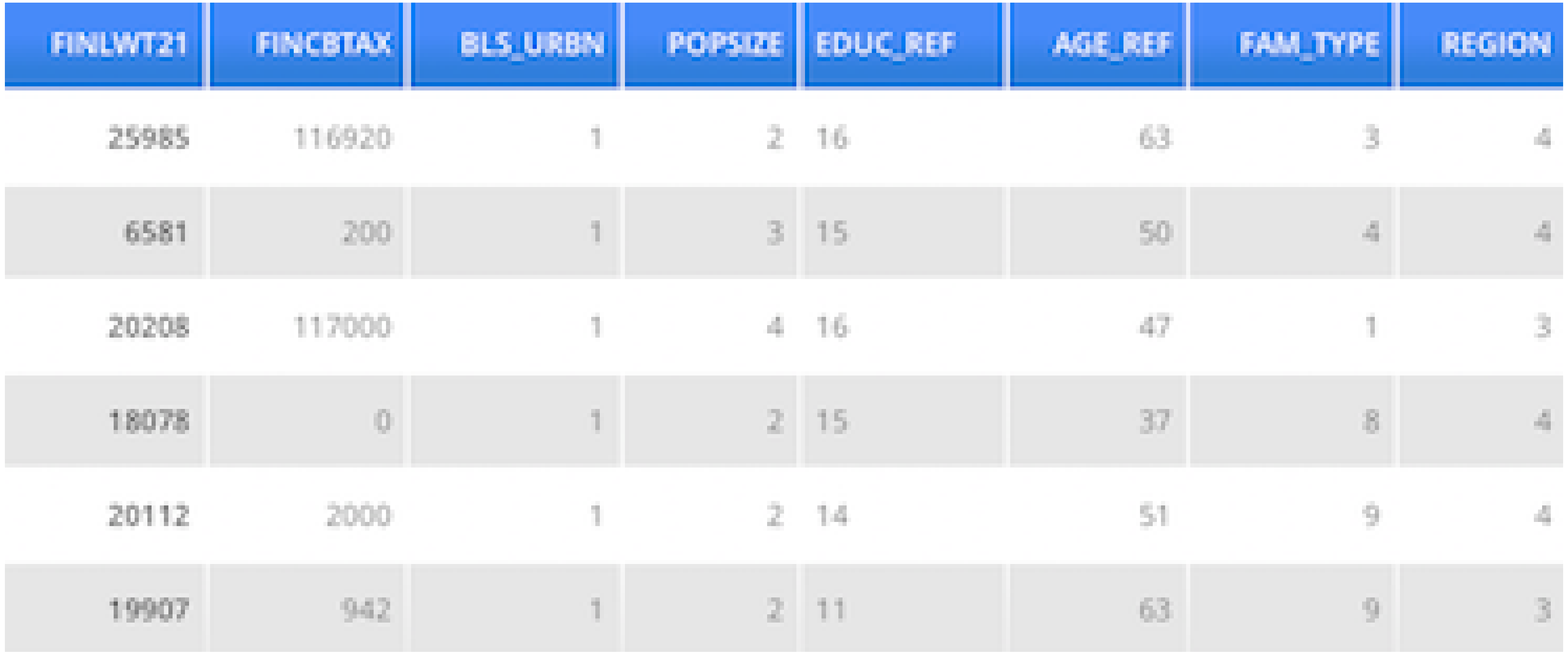
ANALYZING SURVEY DATA IN R

Kelly McConville

Assistant Professor of Statistics

Survey data

- Have you ever found yourself analyzing a dataset that contained a column of weights and wondered what they were?



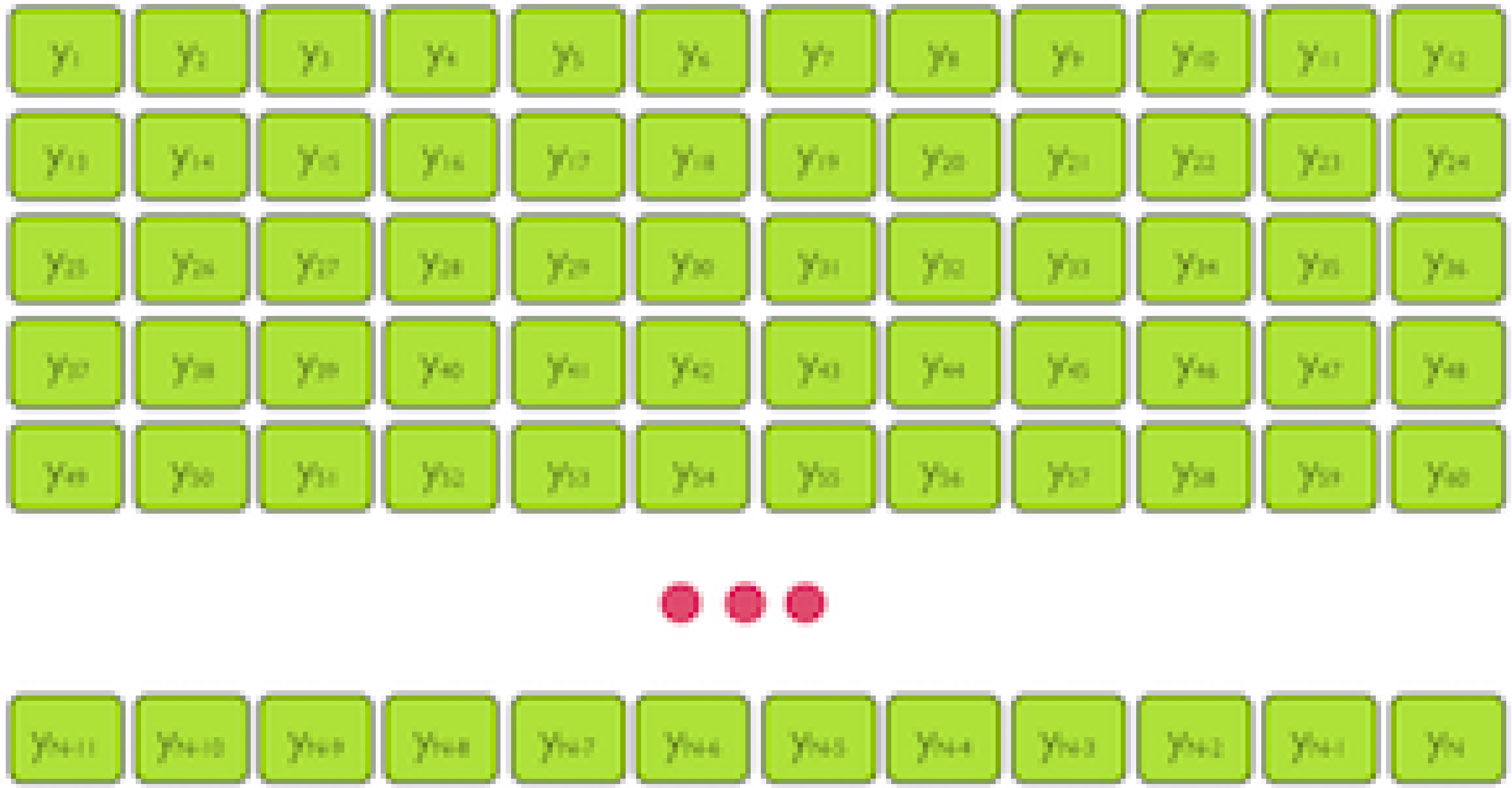
FINLWT21	FINCBTAX	BLS_URBN	POPSIZE	EDUC_REF	AGE_REF	FAM_TYPE	REGION
25985	116920	1	2	16	63	3	4
6581	200	1	3	15	50	4	4
20208	117000	1	4	16	47	1	3
18078	0	1	2	15	37	8	4
20112	2000	1	2	14	51	9	4
19907	942	1	2	11	63	9	3

Survey weights

- What are survey weights?
  - They are the result of using a **complex sampling design** to select a sample from a population.
  - Roughly, the survey weight translates to the number of units in the population that a sampled unit represents.
    - First weight in BLS sample = 25,985 households
    - Second weight in BLS sample = 6,581 households
- How do survey weights **impact** my analyses?

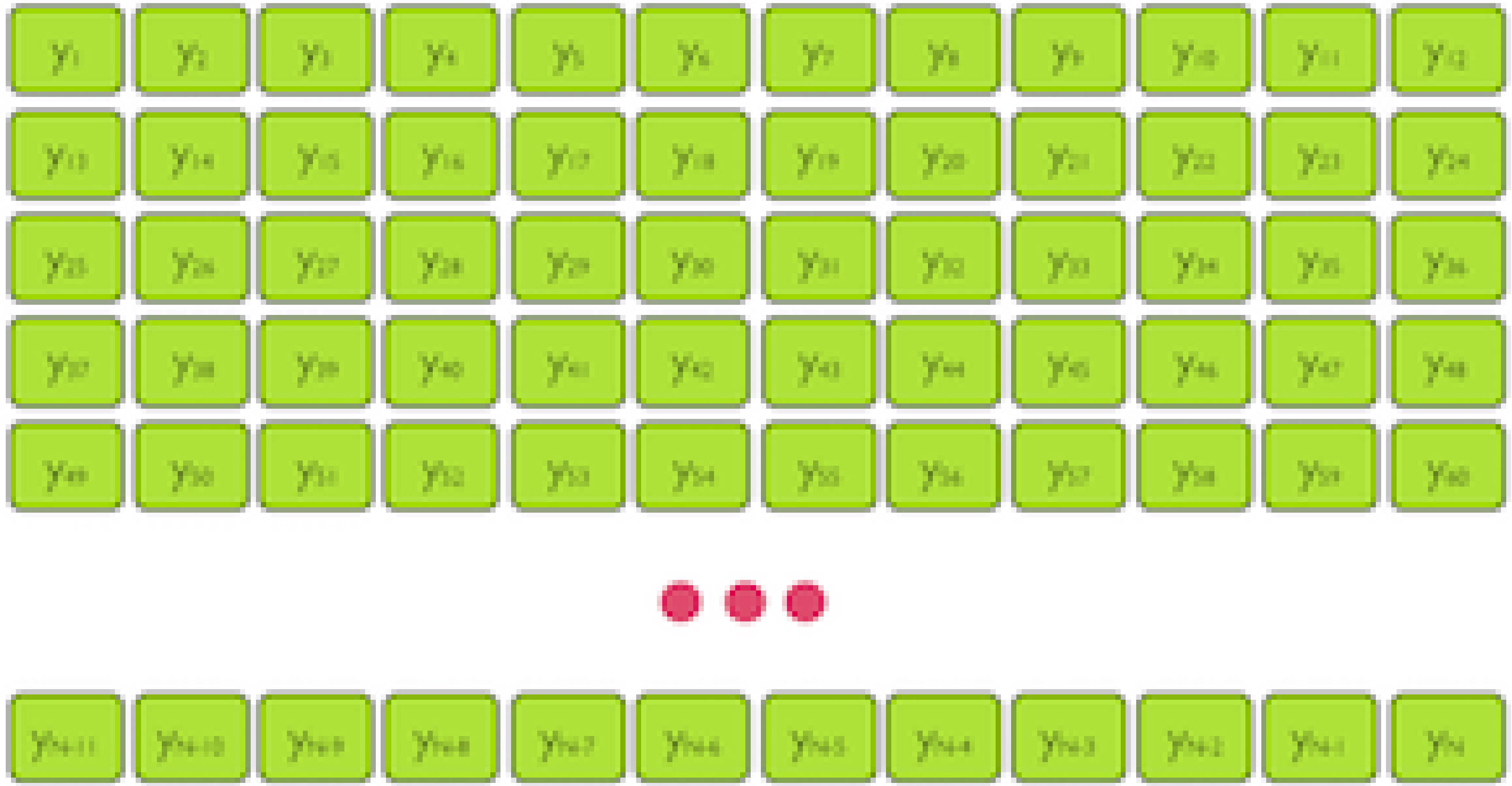
Survey estimation

- Survey data are commonly used to estimate a finite population quantity.



Survey estimation

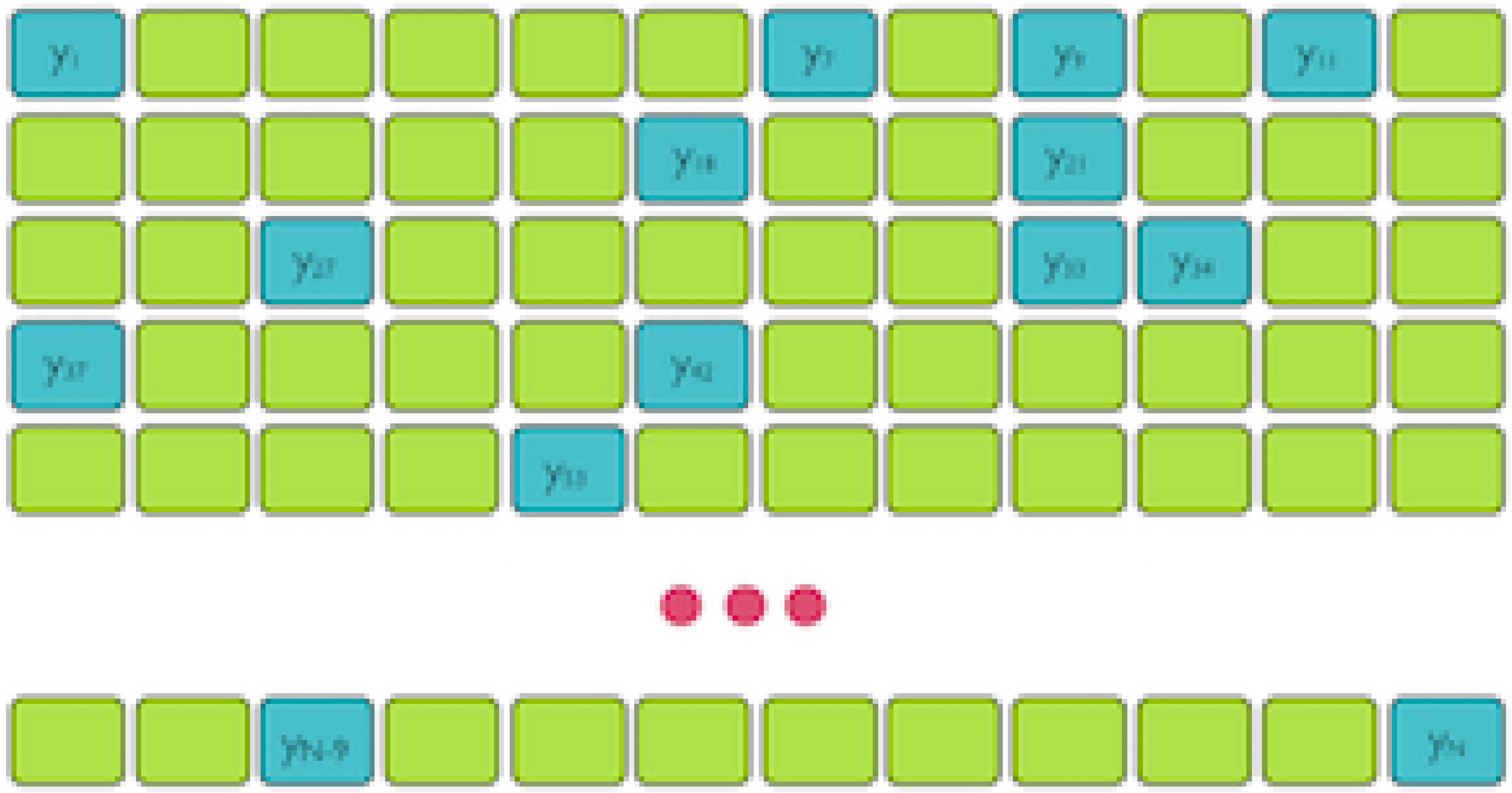
- Estimate the average household income in the U.S.:
  $\mu = \frac{1}{N} \sum_{i \in U} y_i.$



y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}	y_{21}	y_{22}	y_{23}	y_{24}
y_{25}	y_{26}	y_{27}	y_{28}	y_{29}	y_{30}	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}	y_{36}
y_{37}	y_{38}	y_{39}	y_{40}	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}	y_{46}	y_{47}	y_{48}
y_{49}	y_{50}	y_{51}	y_{52}	y_{53}	y_{54}	y_{55}	y_{56}	y_{57}	y_{58}	y_{59}	y_{60}
y_{N-11}	y_{N-10}	y_{N-9}	y_{N-8}	y_{N-7}	y_{N-6}	y_{N-5}	y_{N-4}	y_{N-3}	y_{N-2}	y_{N-1}	y_N

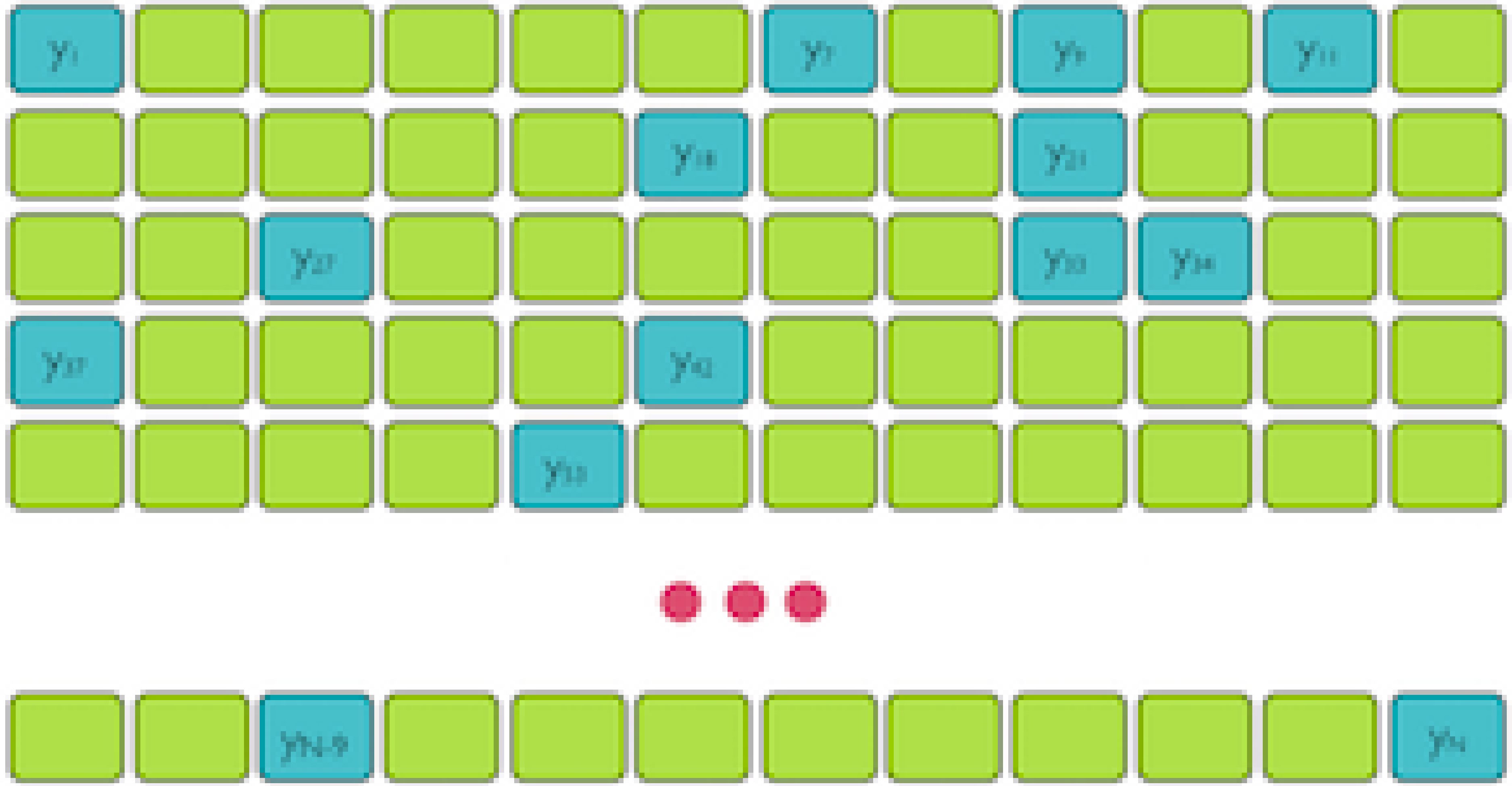
Survey estimation

- Using a complex sampling design, take a sample, called s, of n households.



Survey estimation

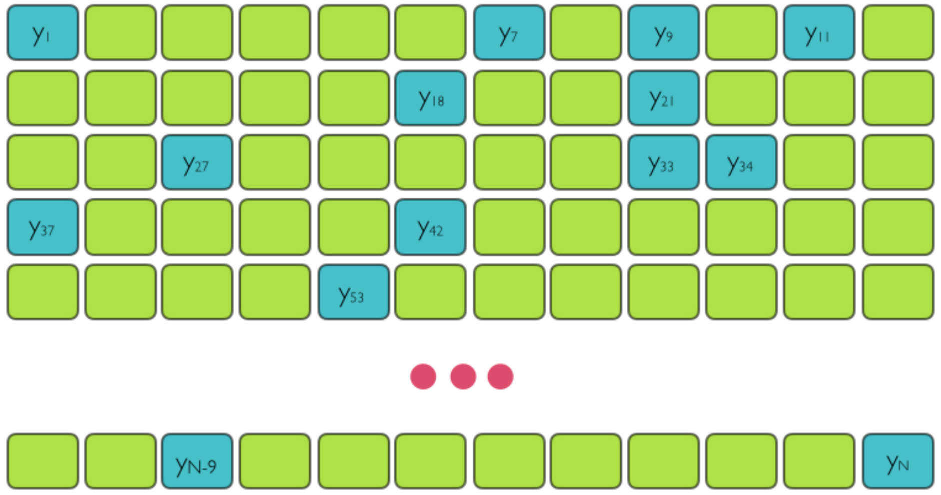
- Sample mean estimator: $\bar{y} = \frac{1}{n} \sum_{i \in s} y_i$.



Survey estimation

- Sample mean estimator:

$$\bar{y} = \frac{1}{n}\sum_{i \in s} y_i$$

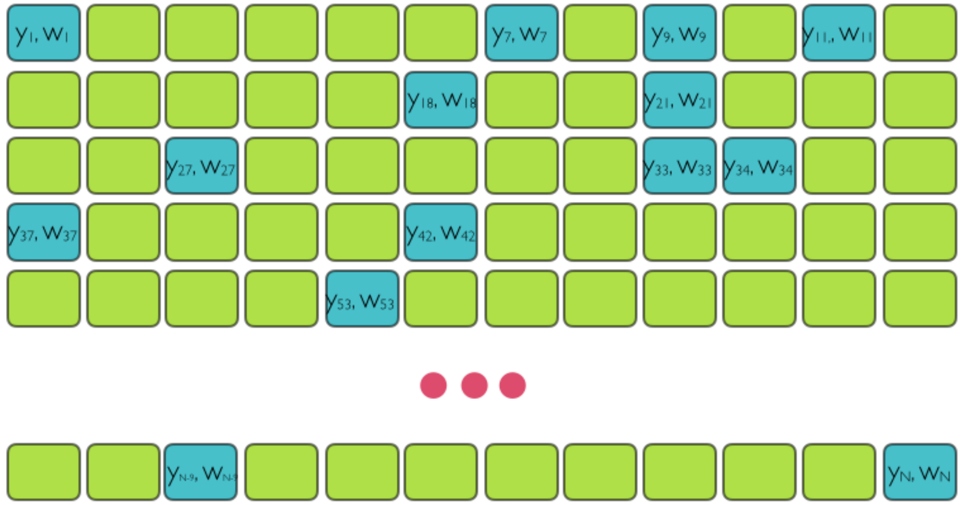


```r
mean(ce$FINCBTAX)
```

```
62480
```

Survey estimation

- For sampled units, we have the values and survey weights.



- How do I incorporate the weights?
- How do the weights impact my estimates? My graphics? My models?

Let's practice!

ANALYZING SURVEY DATA IN R

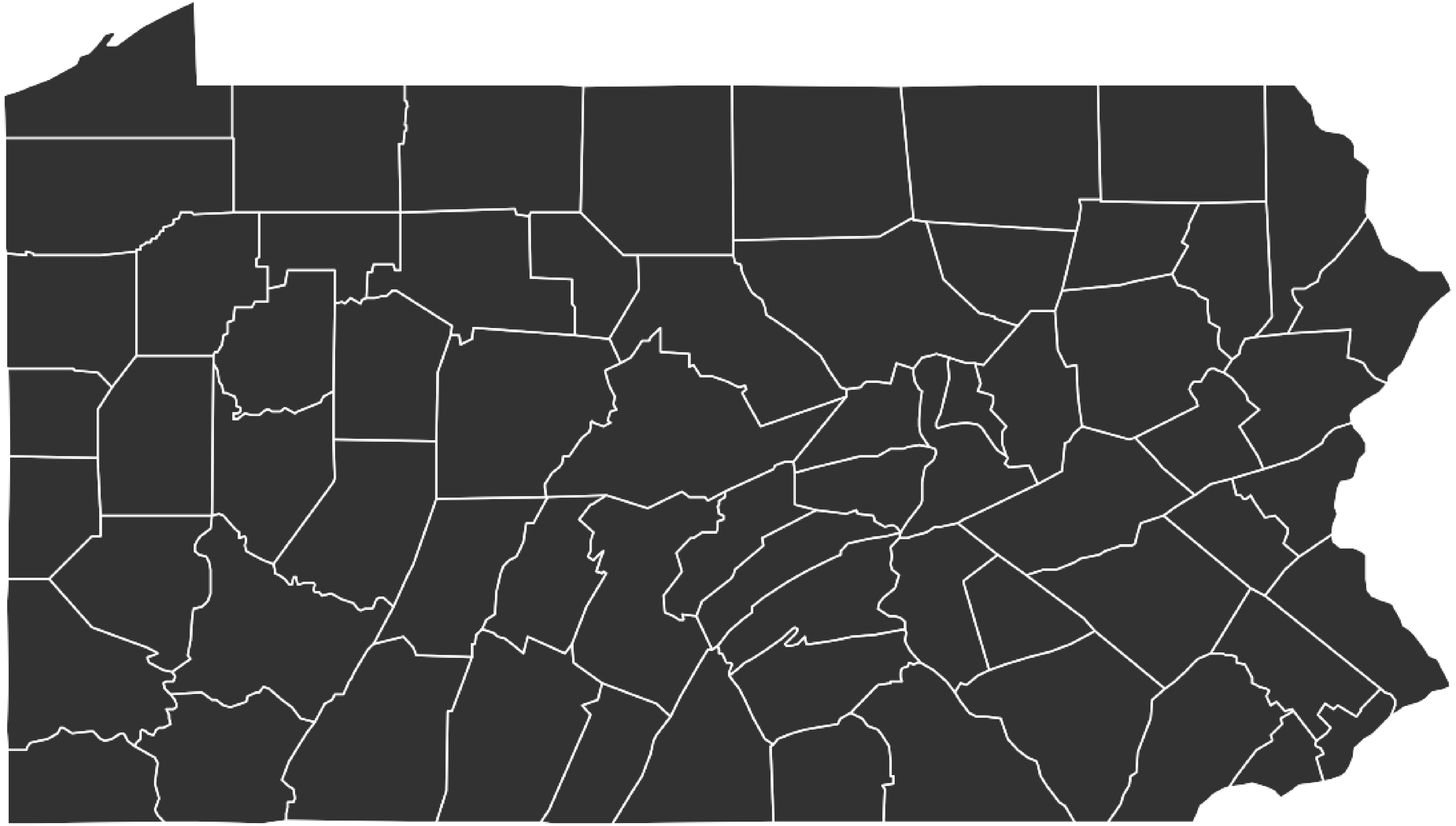
Elements of a sampling design

ANALYZING SURVEY DATA IN R

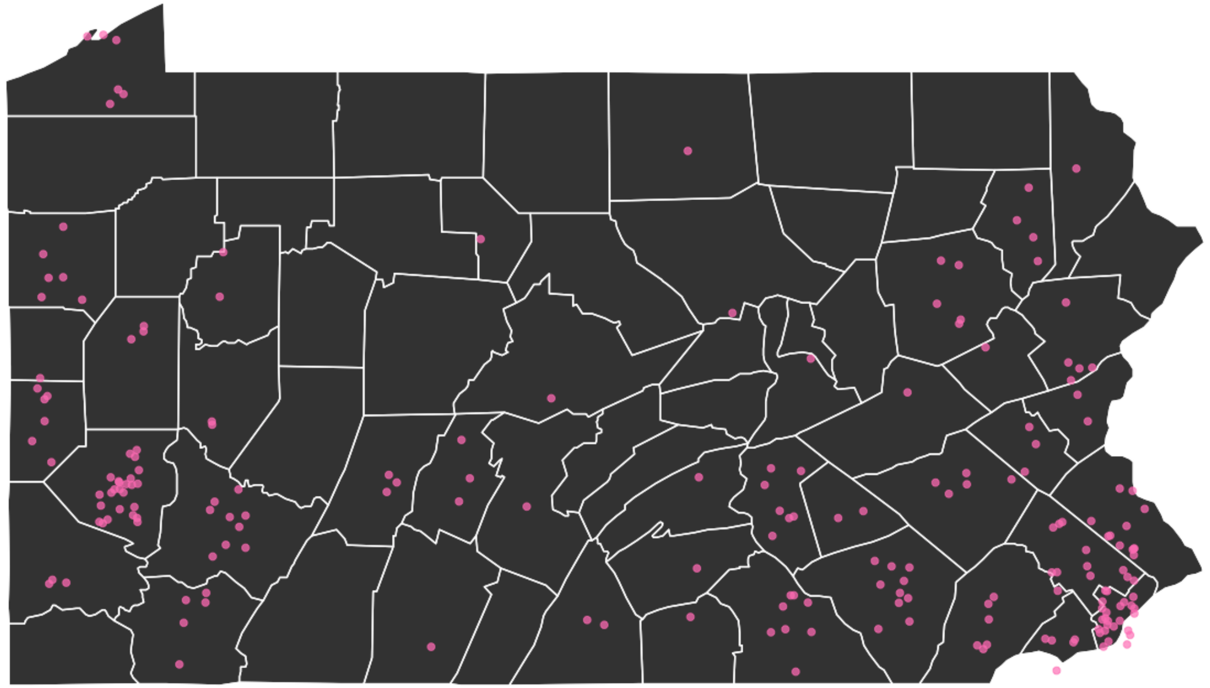
Kelly McConville

Assistant Professor of Statistics

Simple random sampling

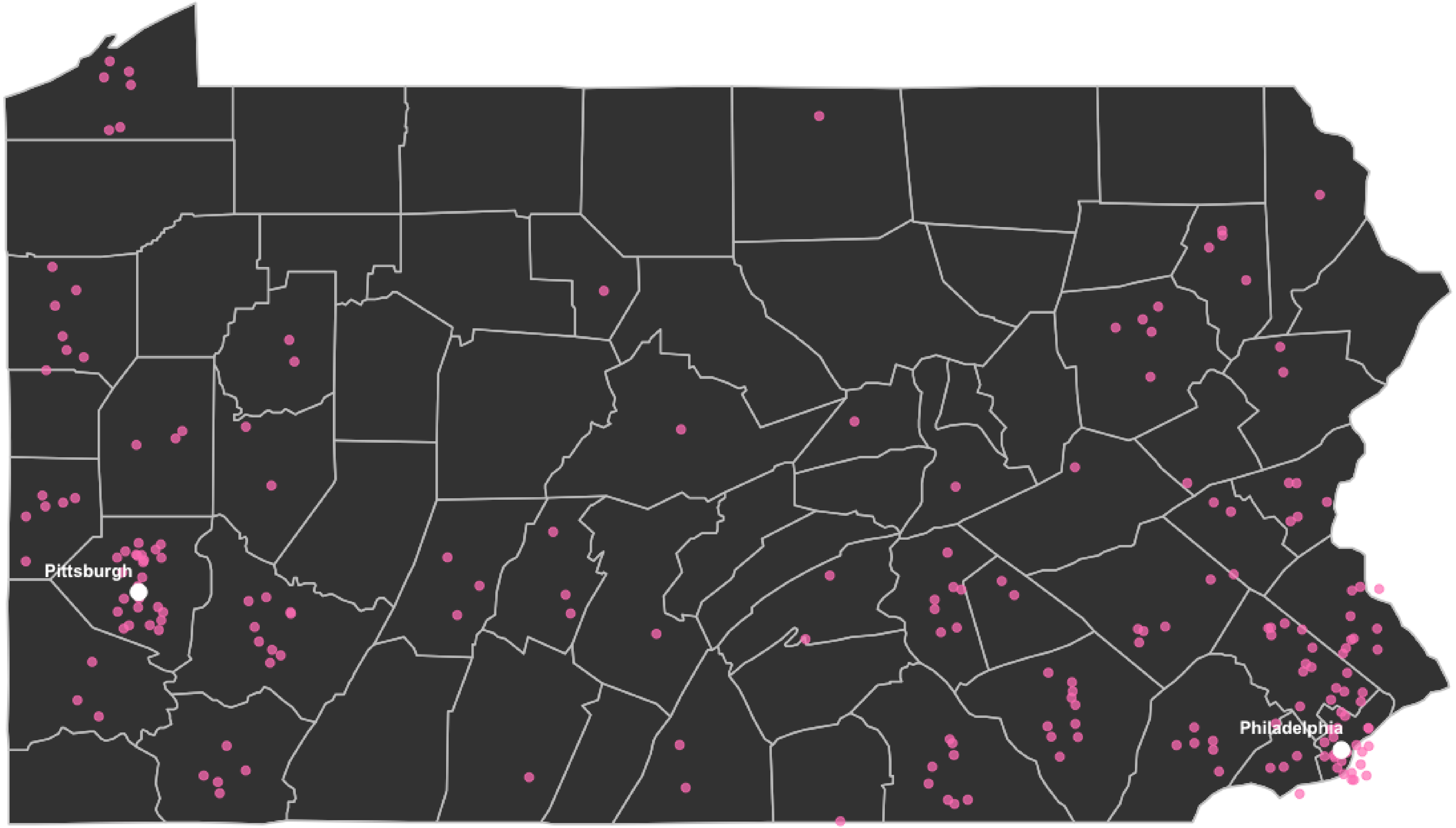


Simple random sampling

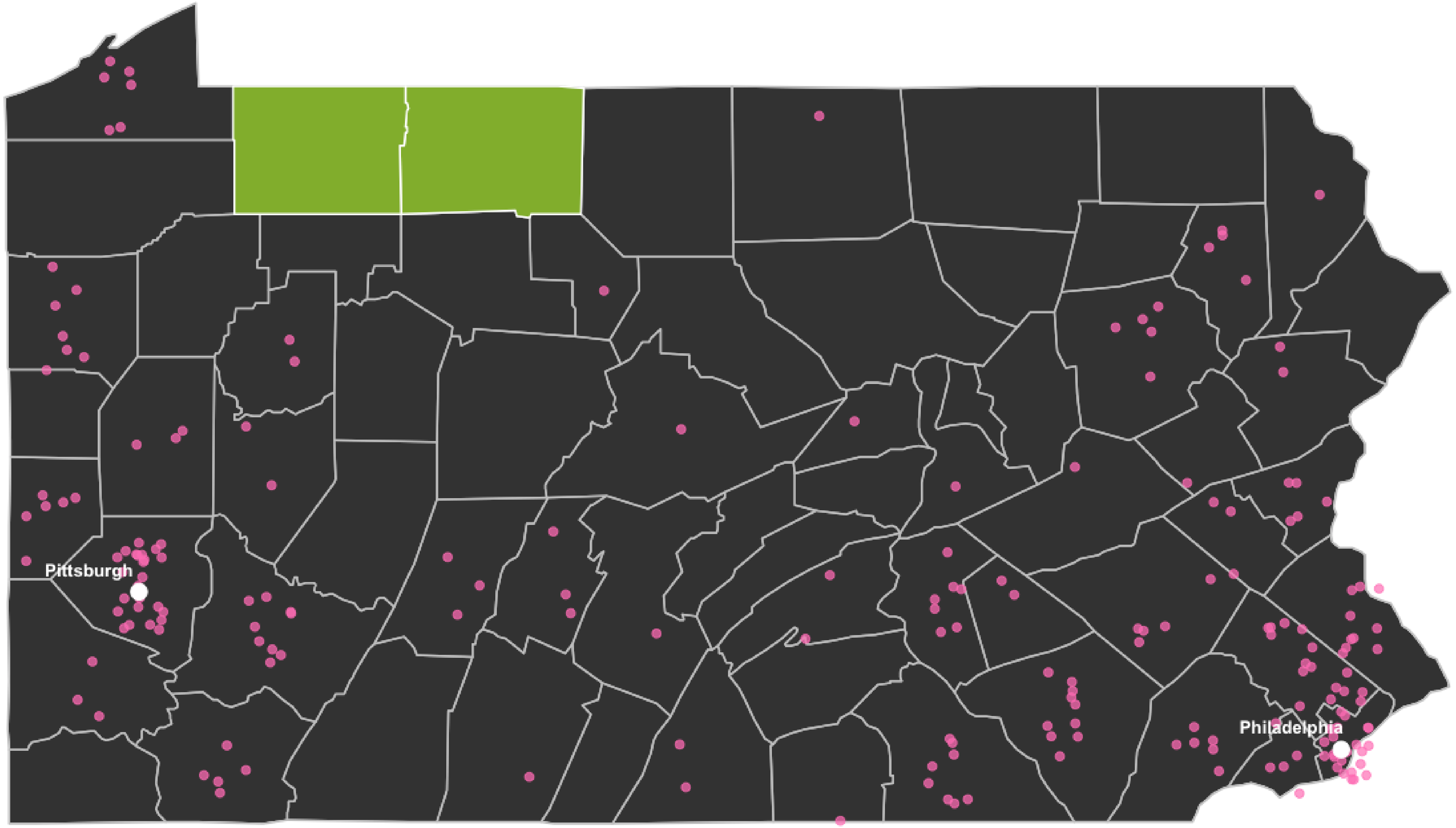


```
library(survey)
srs_design <- svydesign(data = paSample,
                        weights = ~wts,
                        fpc = ~N, id = ~1)
```

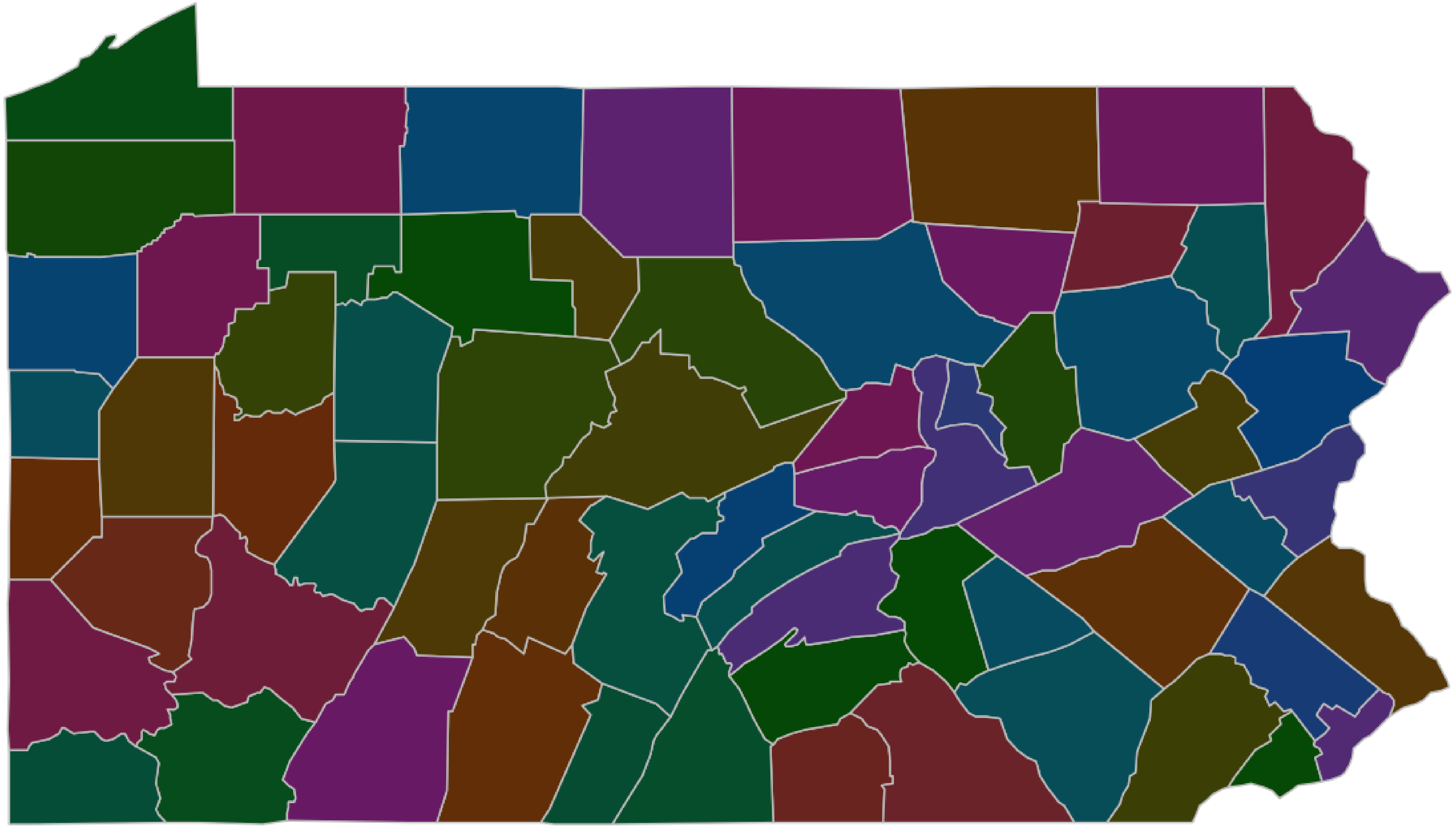
Simple random sampling



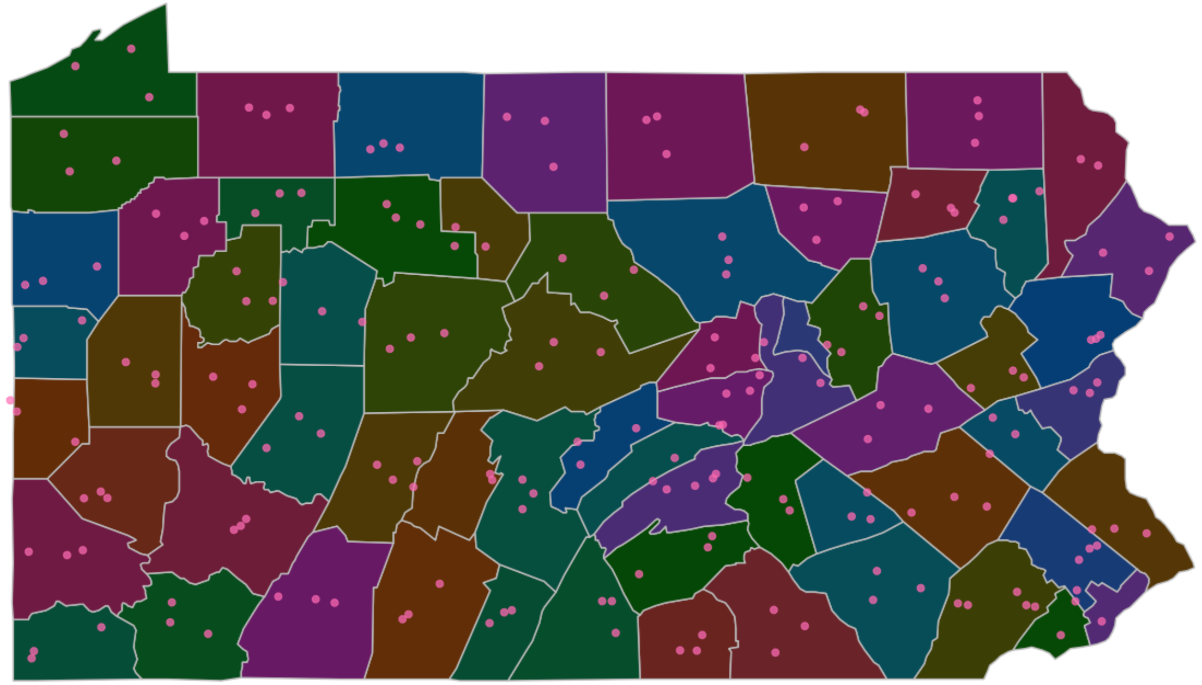
Simple random sampling



Stratified sampling

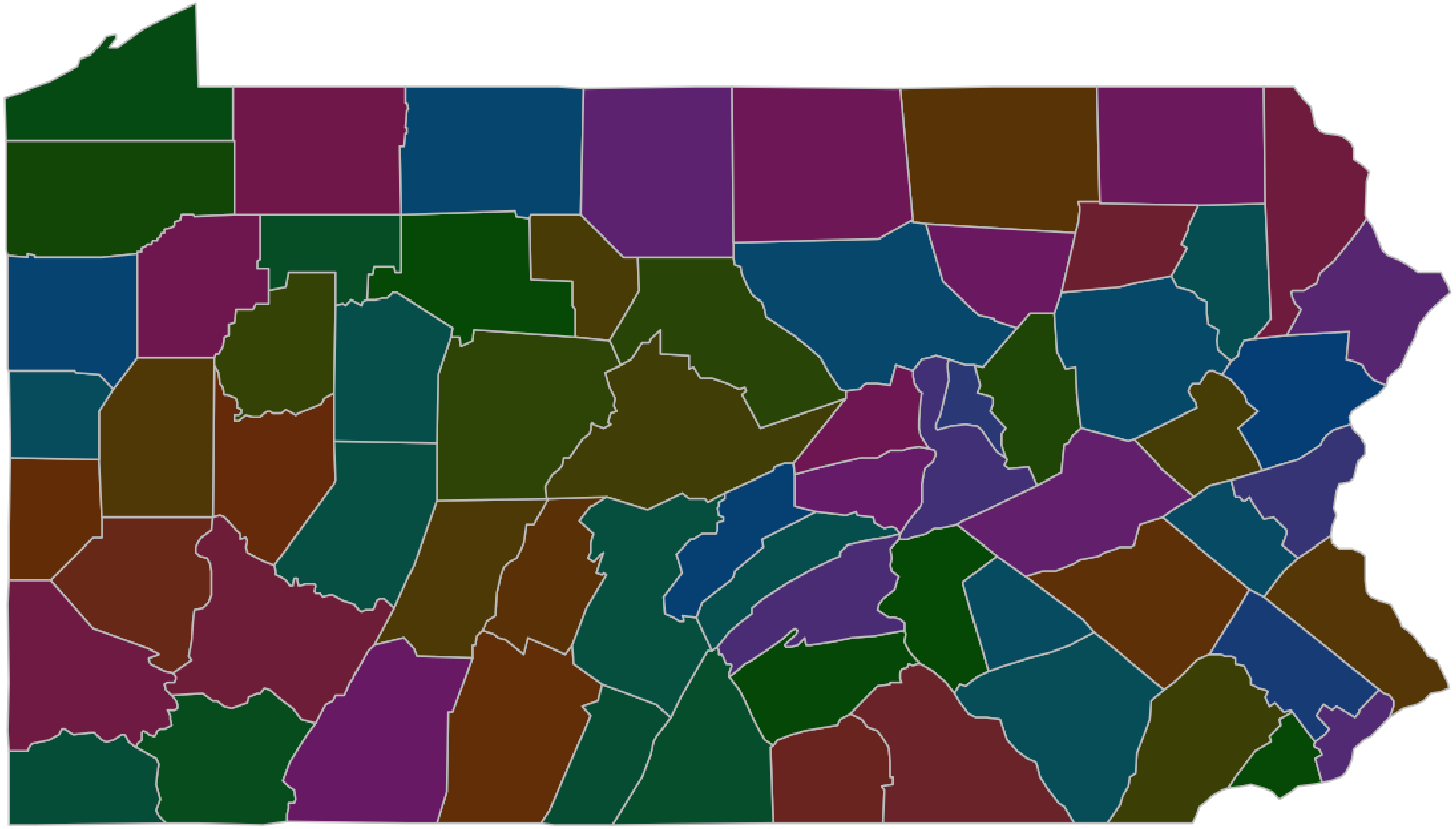


Stratified sampling

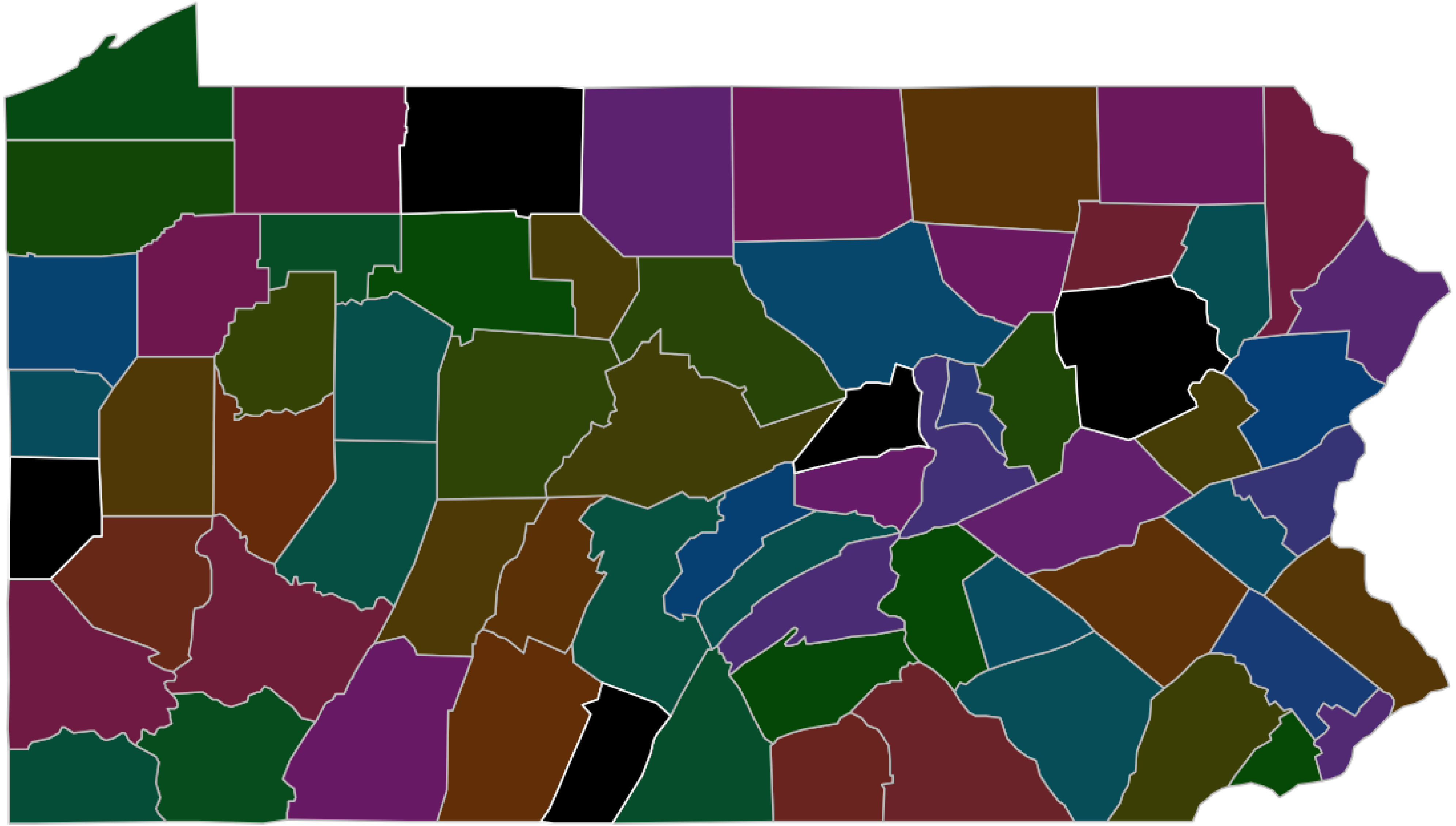


```r
library(survey)
stratified_design <- svydesign(data = paSample, id = ~1, weights = ~wts,
                               strata = ~county, fpc = ~N)
```

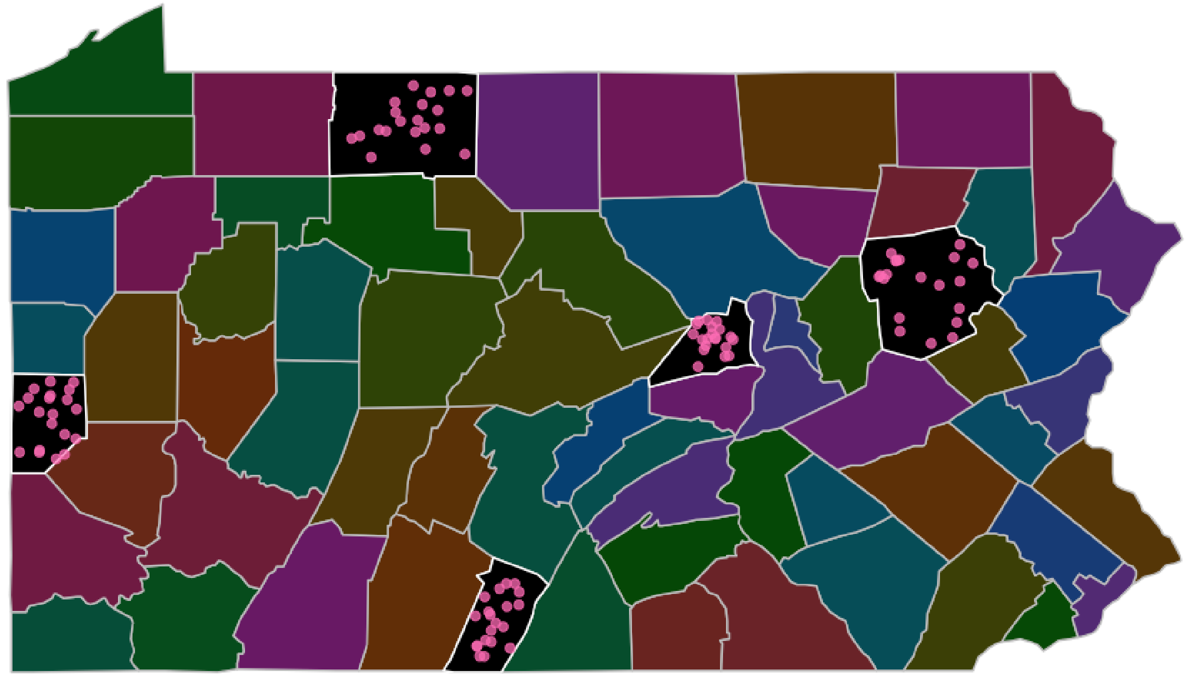
Cluster sampling



Cluster sampling



Cluster sampling



```r
library(survey)
cluster_design <- svydesign(data = paSample, id = ~county + personid,
                            fpc = ~N1 + N2, weights = ~wts)
```

Let's practice!

ANALYZING SURVEY DATA IN R

Impact of weights

ANALYZING SURVEY DATA IN R

Kelly McConville
Assistant Professor of Statistics

National Health and Nutrition Examination Survey (NHANES)

- Conducted by the U.S. National Center for Health Statistics.
- **Goal**: Understand the health of adults and children in the US.
- It is collected using a 4 stage design.
- **Stage 0**: The U.S. is *stratified* by geography and proportion of minority populations.
- **Stage 1**: Within strata, counties are randomly selected.
- **Stage 2**: Within counties, city blocks are randomly selected.
- **Stage 3**: Within city blocks, households randomly selected.
- **Stage 4**: Within households, people randomly selected.

NHANES

```r
library(NHANES)
dim(NHANESraw)
```

```
20293    78
```

```r
library(dplyr)
summarize(NHANESraw, N_hat = sum(WTMEC2YR))
```

```
# A tibble: 1 x 1
      N_hat
      <dbl>
 1 608534400
```

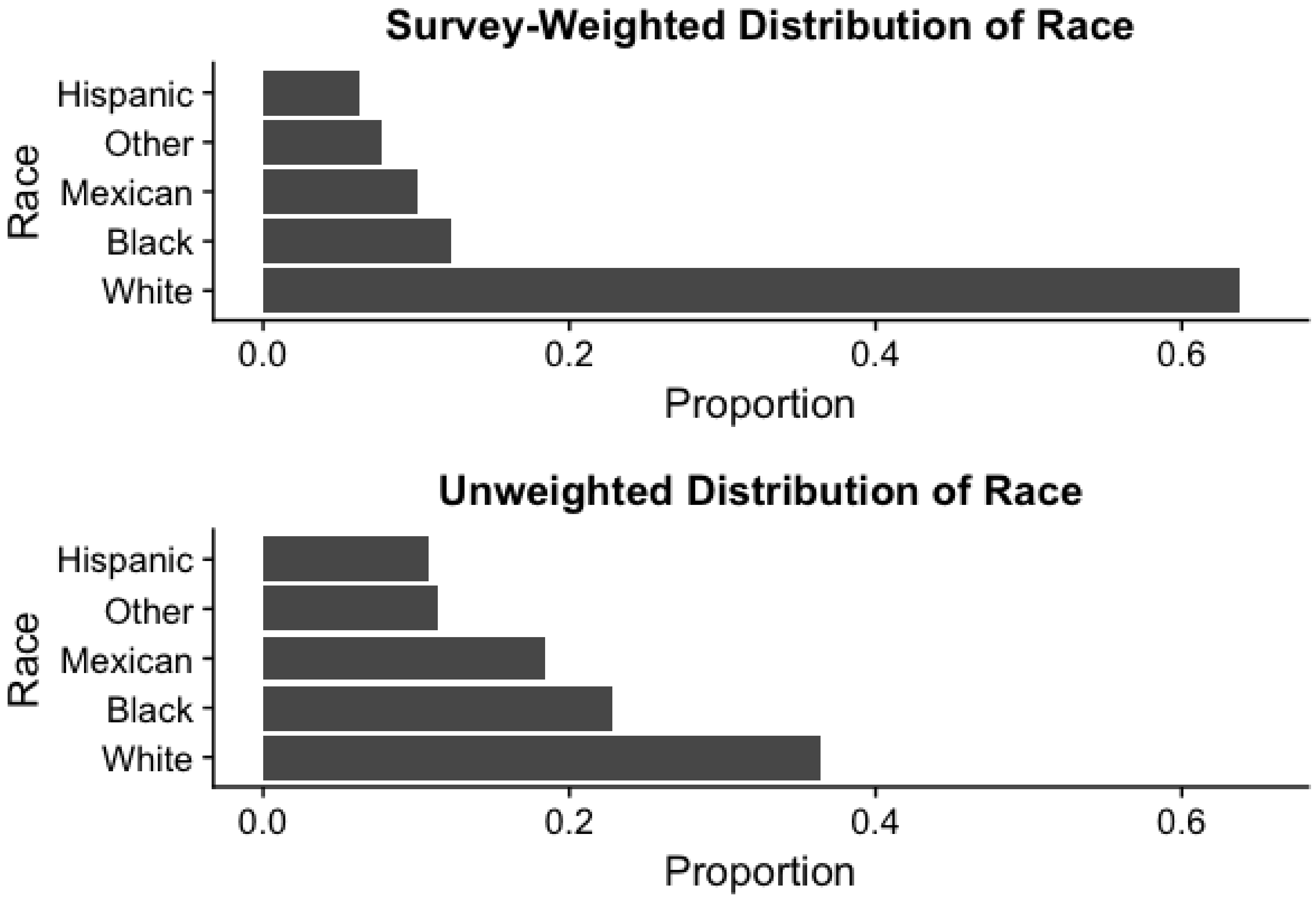
```r
NHANESraw <- mutate(NHANESraw, WTMEC4YR = WTMEC2YR / 2)
```

NHANES

```
NHANES_design <- svydesign(data = NHANESraw,
                           strata = ~SDMVSTRA,
                           id = ~SDMVPSU, nest = TRUE,
                           weights = ~WTMEC4YR)
distinct(NHANESraw, SDMVPSU)
```

```
# A tibble: 3 x 1
  SDMVPSU
    <int>
 1       1
 2       2
 3       3
```

Visualizing impact of weights



Let's practice!

ANALYZING SURVEY DATA IN R