ANALYZING SURVEY DATA IN R

What are survey weights?

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?

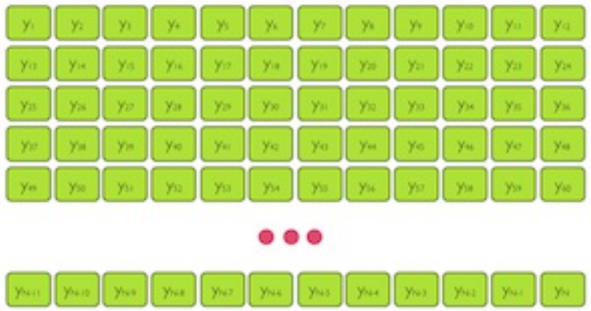
FINLWT21	FINCBTAX	BLS_URBN	POPSIZE	EDUC_REF	AGE_REF	FAM_TYPE	REGION
25985	116920	1	2	16	63	3	4
6581	200	1	3	15	50	4	4
20208	117000	1	4	16	47	1	3
18078	0	1	2	15	37	8	4
20112	2000	1	2	14	51	9	4
19907	942	1	2	11	63	9	3

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.

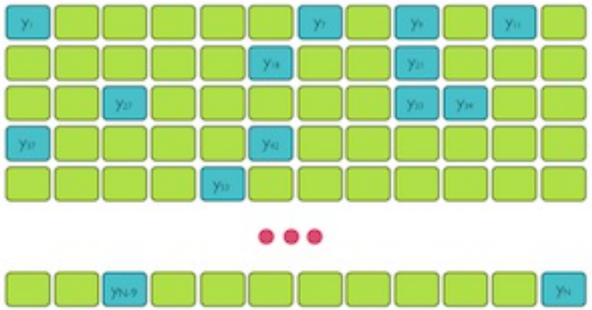


Survey estimation

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

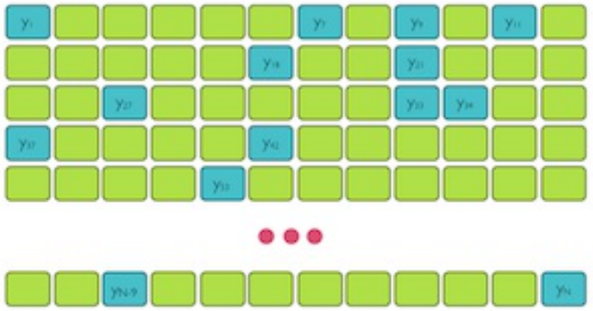
Survey estimation

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



Survey estimation

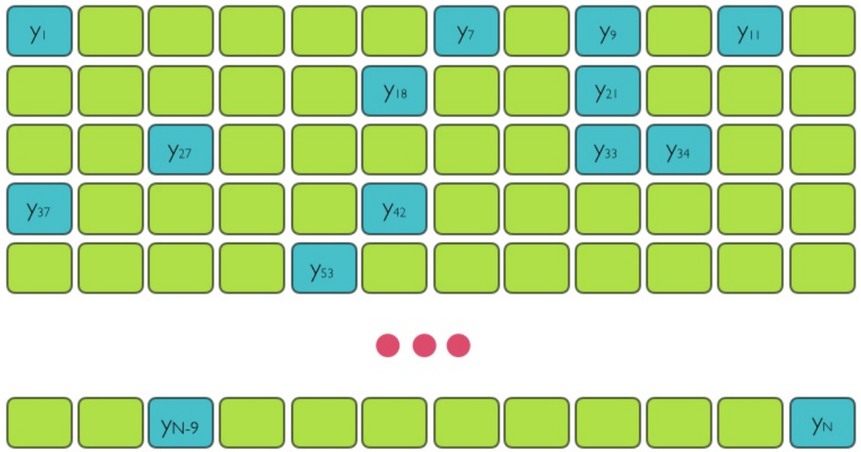
- Sample mean estimator: $\bar{y} = \frac{1}{n} \sum_{i \in s} y_i$.



Survey estimation

- Sample mean estimator:

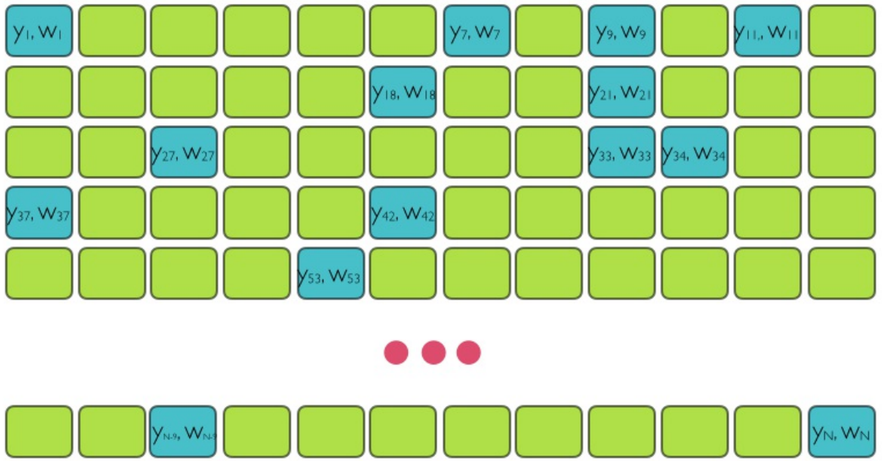
$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i$$




```
mean(ce$FINCBTAX)
[1] 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?

ANALYZING SURVEY DATA IN R

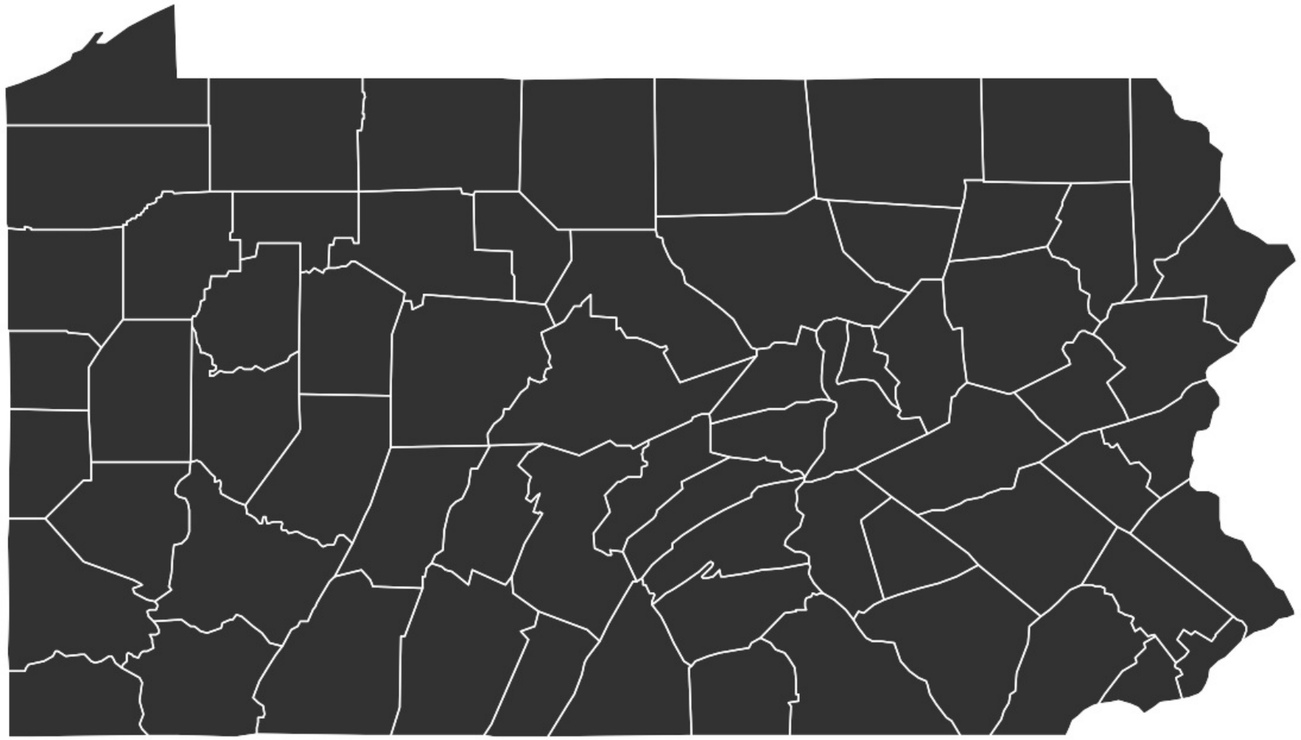
Let's practice!

ANALYZING SURVEY DATA IN R

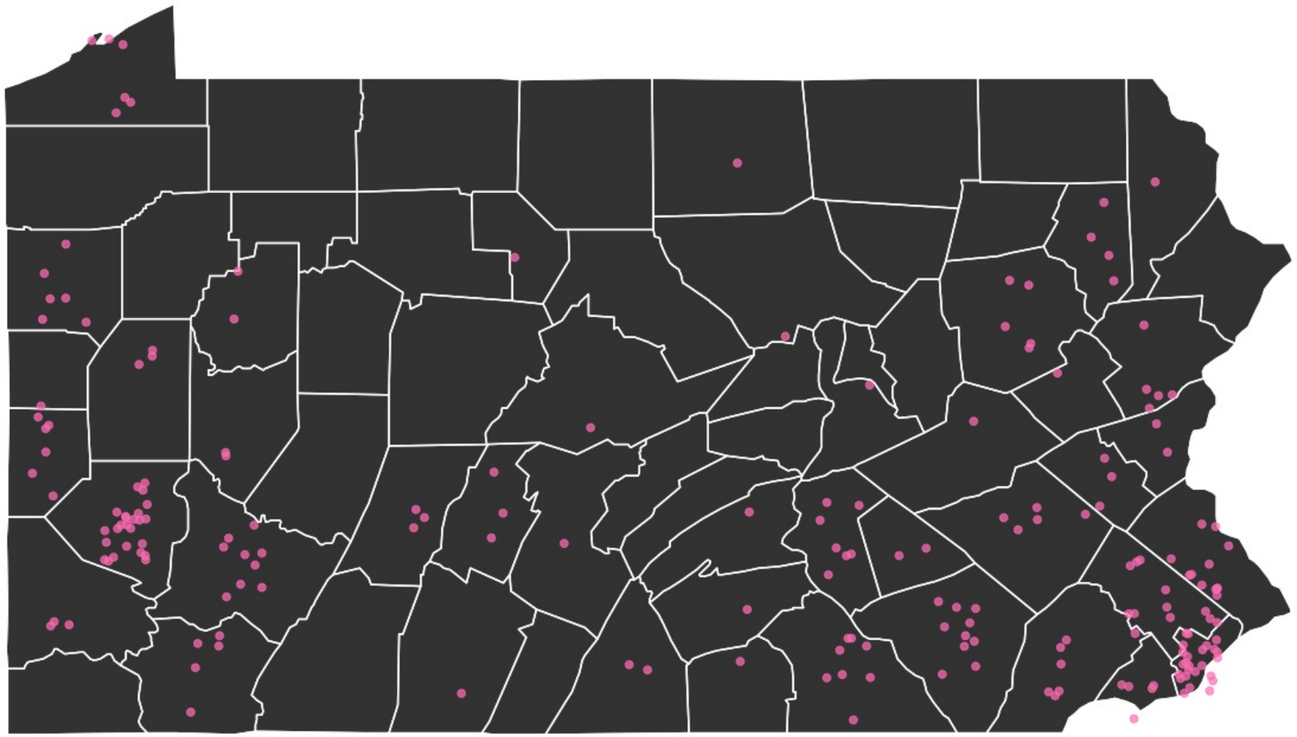
Elements of a sampling design

Kelly McConville
Assistant Professor of Statistics

Simple random sampling

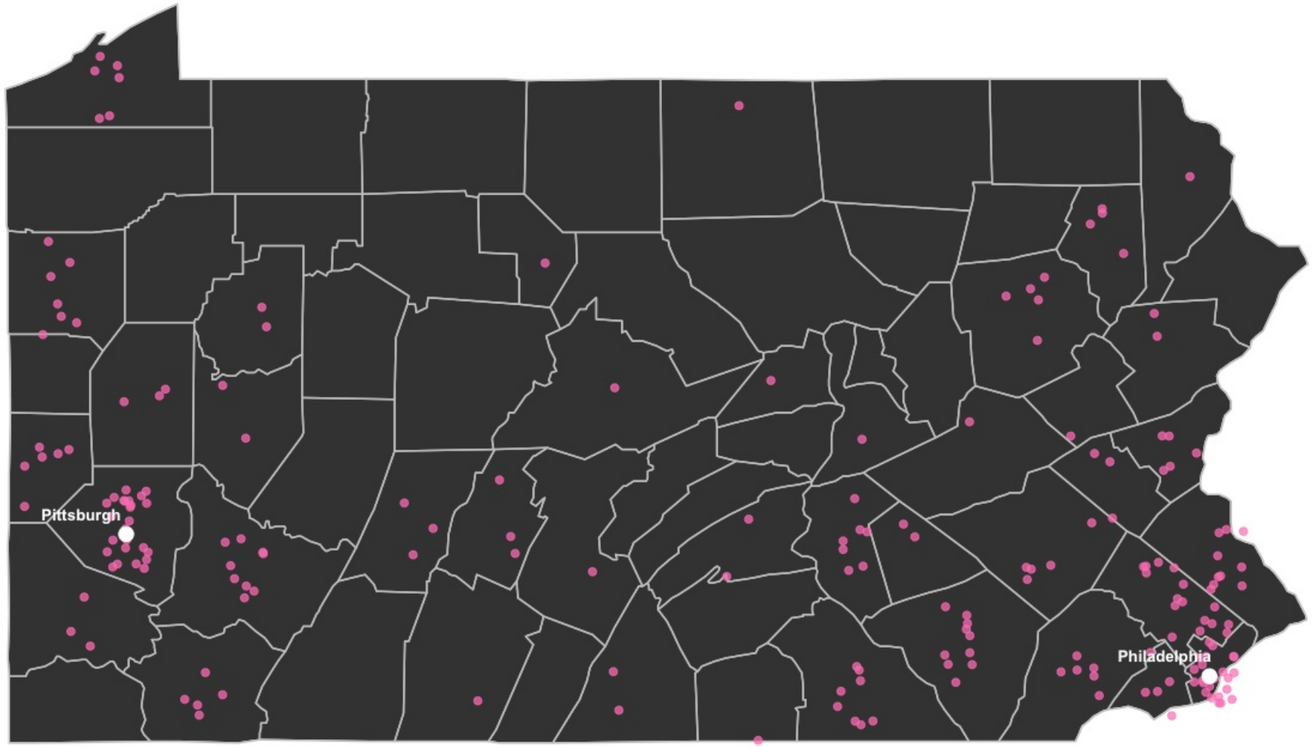


Simple random sampling

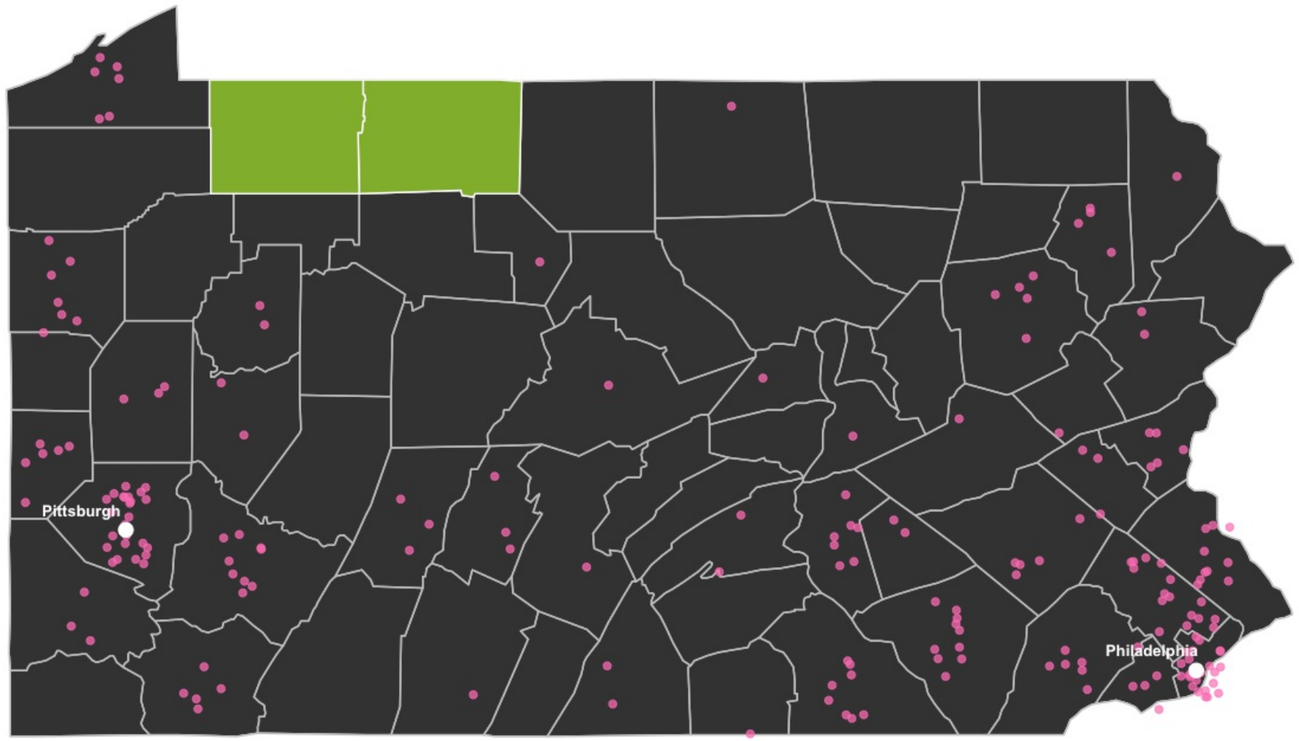


```
library(survey)
srs_design <- svydesign(data = paSample,
                        weights = ~wts,
                        fpc = ~N, id = ~1)
```

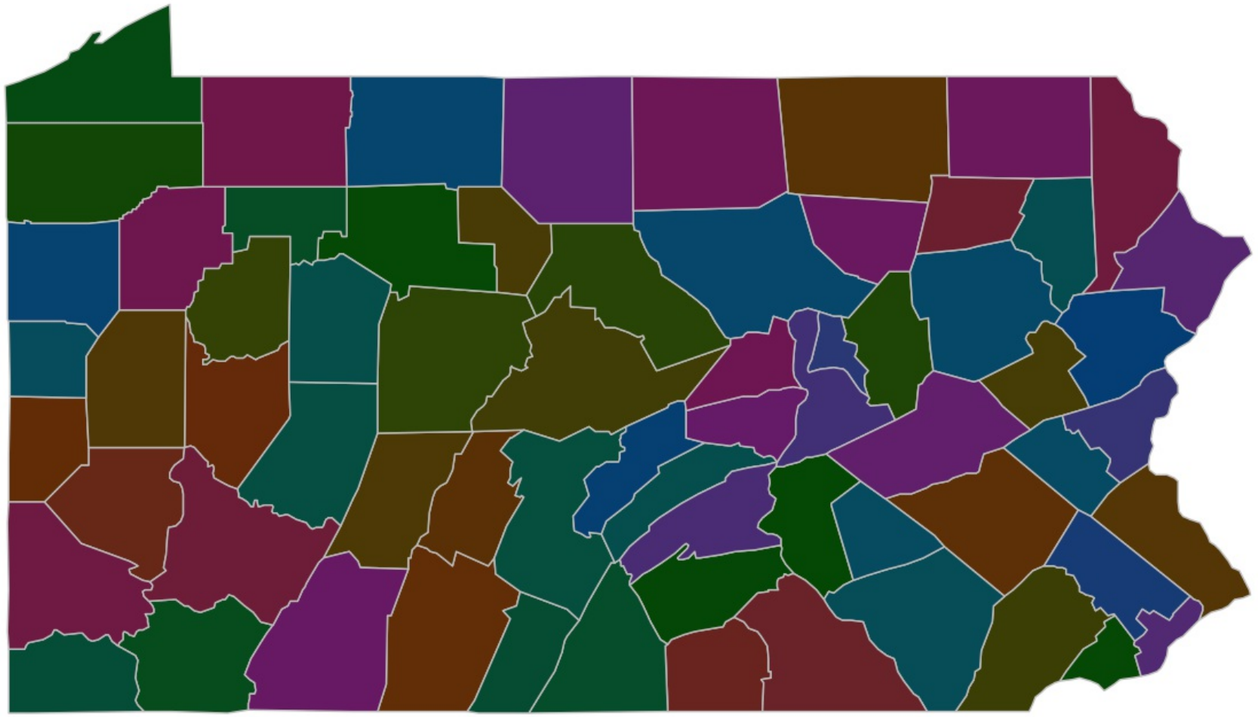
Simple random sampling



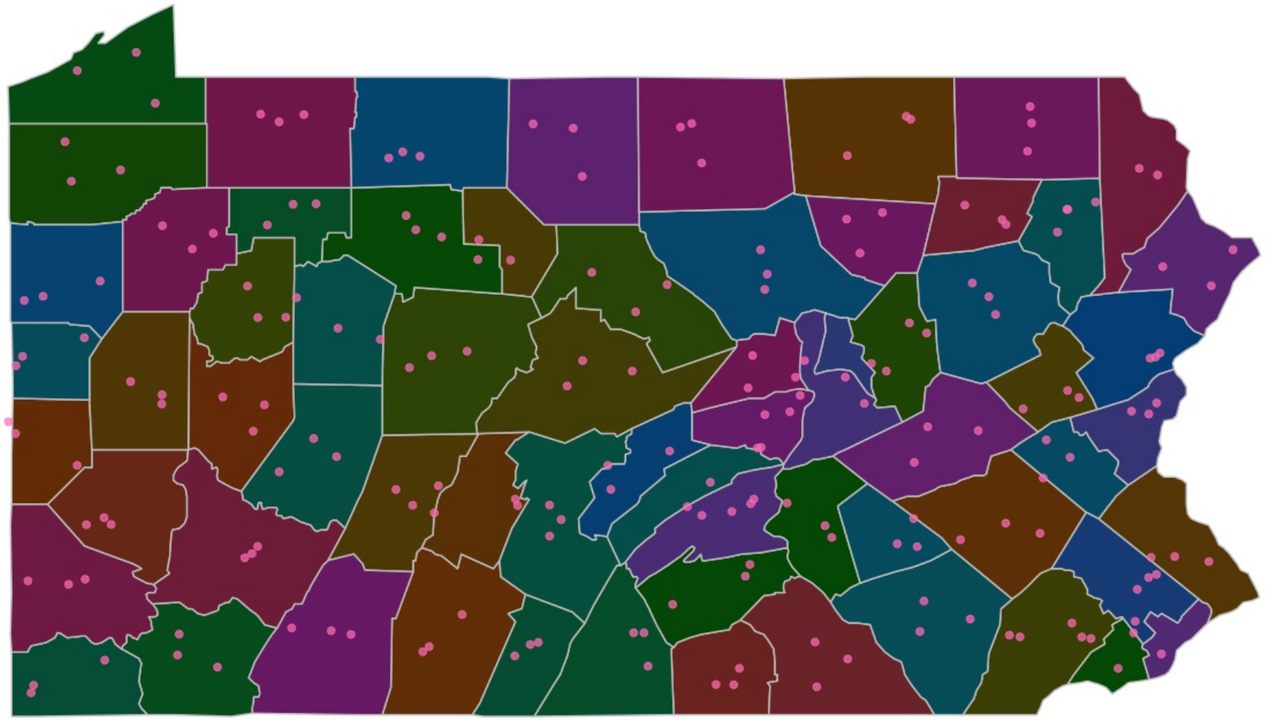
Simple random sampling



Stratified sampling

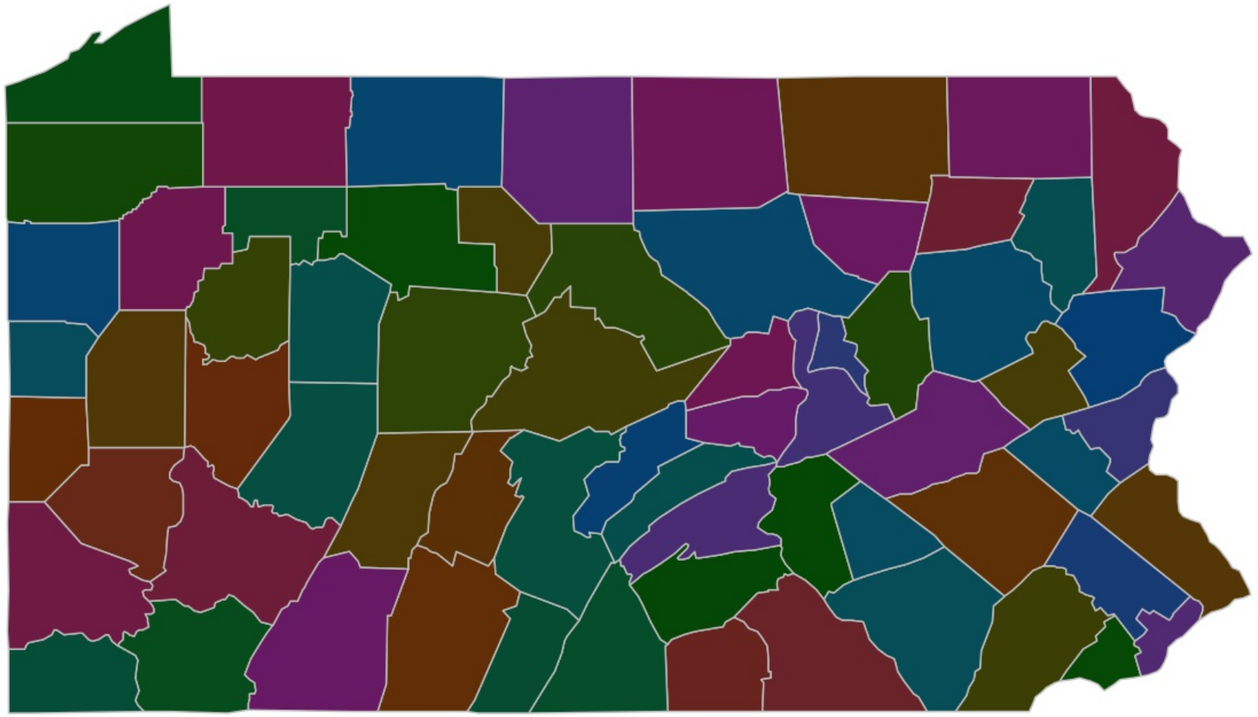


Stratified sampling

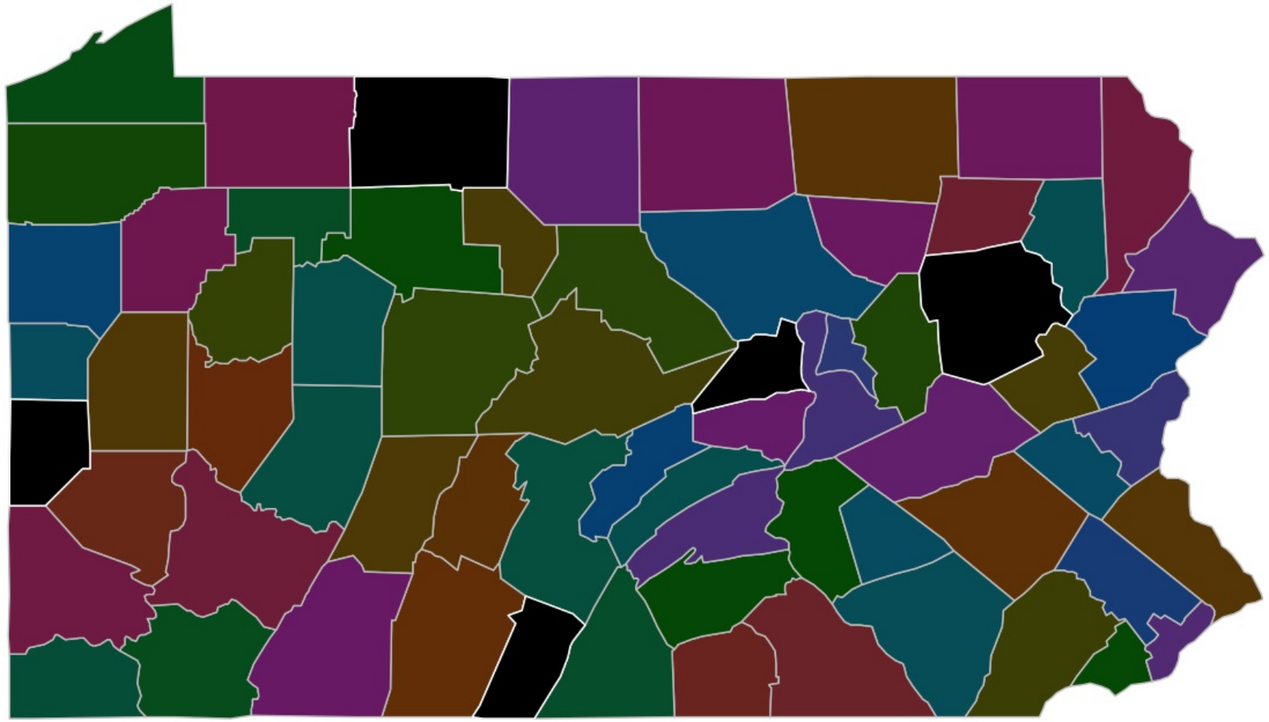


```
library(survey)
stratified_design <- svydesign(data = paSample, id = ~1, weights = ~wts,
                               strata = ~county, fpc = ~N)
```

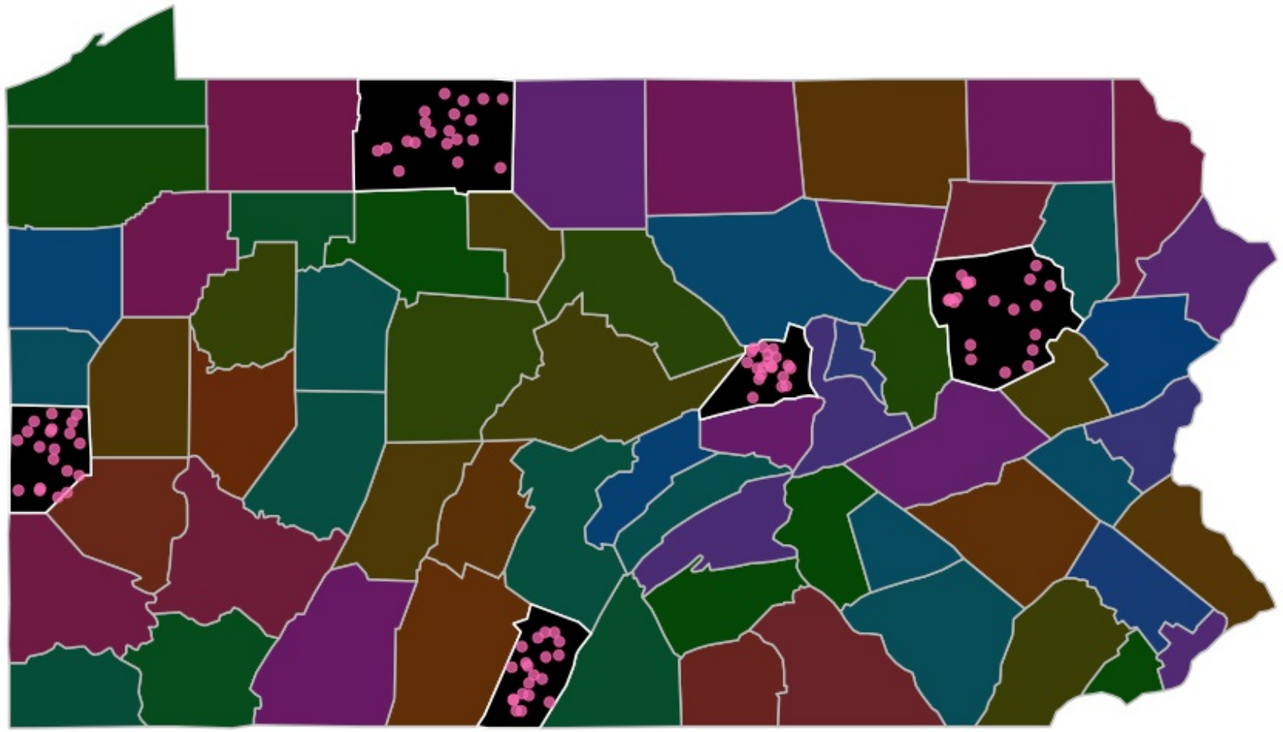
Cluster sampling



Cluster sampling



Cluster sampling



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

ANALYZING SURVEY DATA IN R

Let's practice!

ANALYZING SURVEY DATA IN R

Impact of weights

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

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

```
[1] 20293    78
```

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

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

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

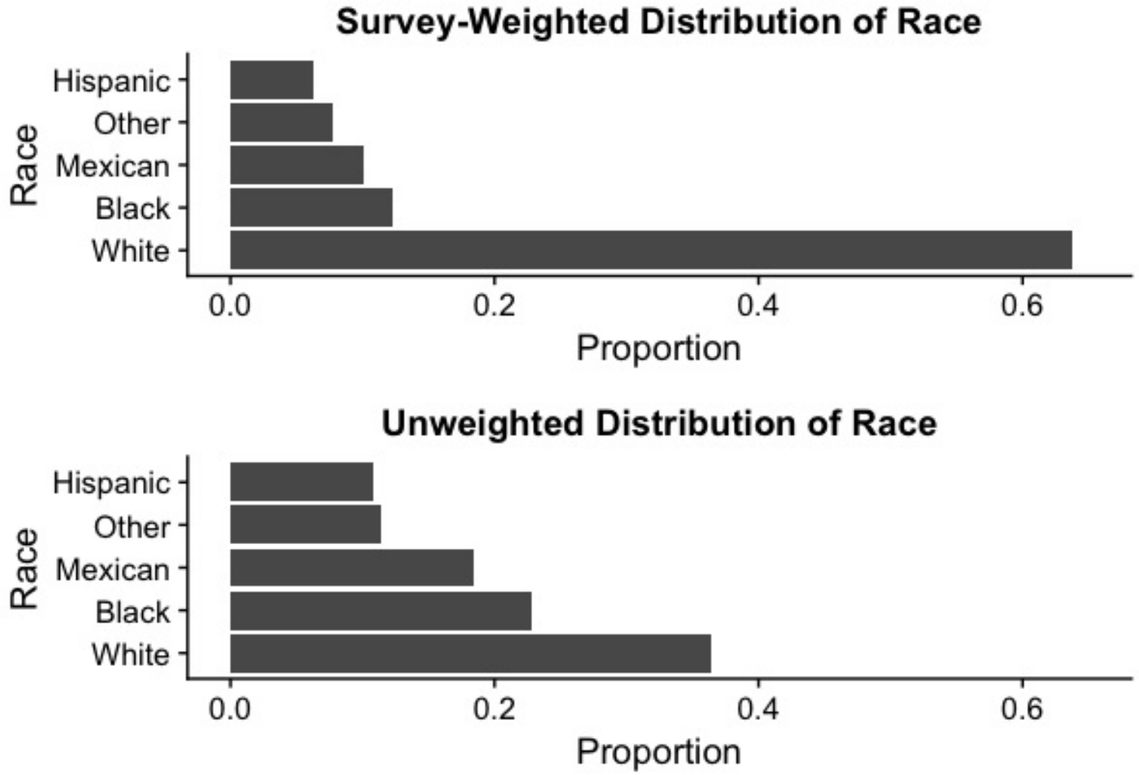
NHANES

```
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



ANALYZING SURVEY DATA IN R

Let's practice!