Washington University in St. Louis
Brown School

Analysis of Public Health Data with a Complex Sample Survey Design

F. Hunter McGuire, MPH, PhD student

February 28, 2022

Learning objectives

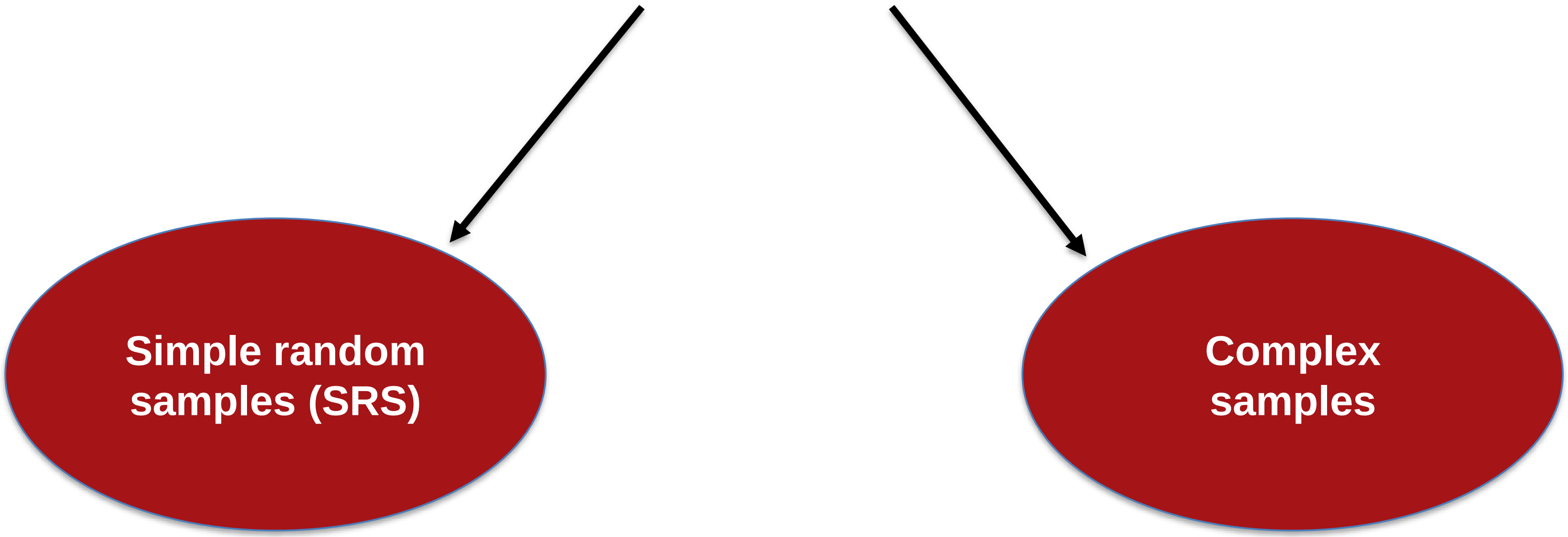
Gain an understanding of:

1. Complex sample designs, weighting, design effects, design-based analysis, and variance estimation issues
2. Analysis best practices in common statistical packages (R, Stata, and SAS)
    - *Guided analysis tutorial in R (time permitting)*

Portions of this presentation were adapted from a short course led by Brady T. West, PhD (University of Michigan)

INTRODUCTION TO COMPLEX SAMPLE DESIGNS

Probability sample designs



Simple random samples (SRS)

Complex samples

Simple random sample (SRS)

Typical features

- **Each sampling unit** *(e.g., potential survey participants)* in the target population has an **equal probability** of being sampled
- Requires an **enumerated list** of all population members *(e.g., contact info for all enrolled students)*
- **Assign random numbers** to each population member + use a **random number generator** to select participants

Benefits & Drawbacks

- Benefits:
  - **Easiest way** to obtain an **unbiased sample**
  - **Relatively simple to infer results** back to the target population (basis of common statistical assumptions)
- Drawbacks:
  - **Difficult to enumerate** the entire population (esp. as populations increase)
  - Sampled participants may span a **wide geographic range** (increasing costs)
  - **Subpopulations**: difficulty recruiting and obtaining valid estimates

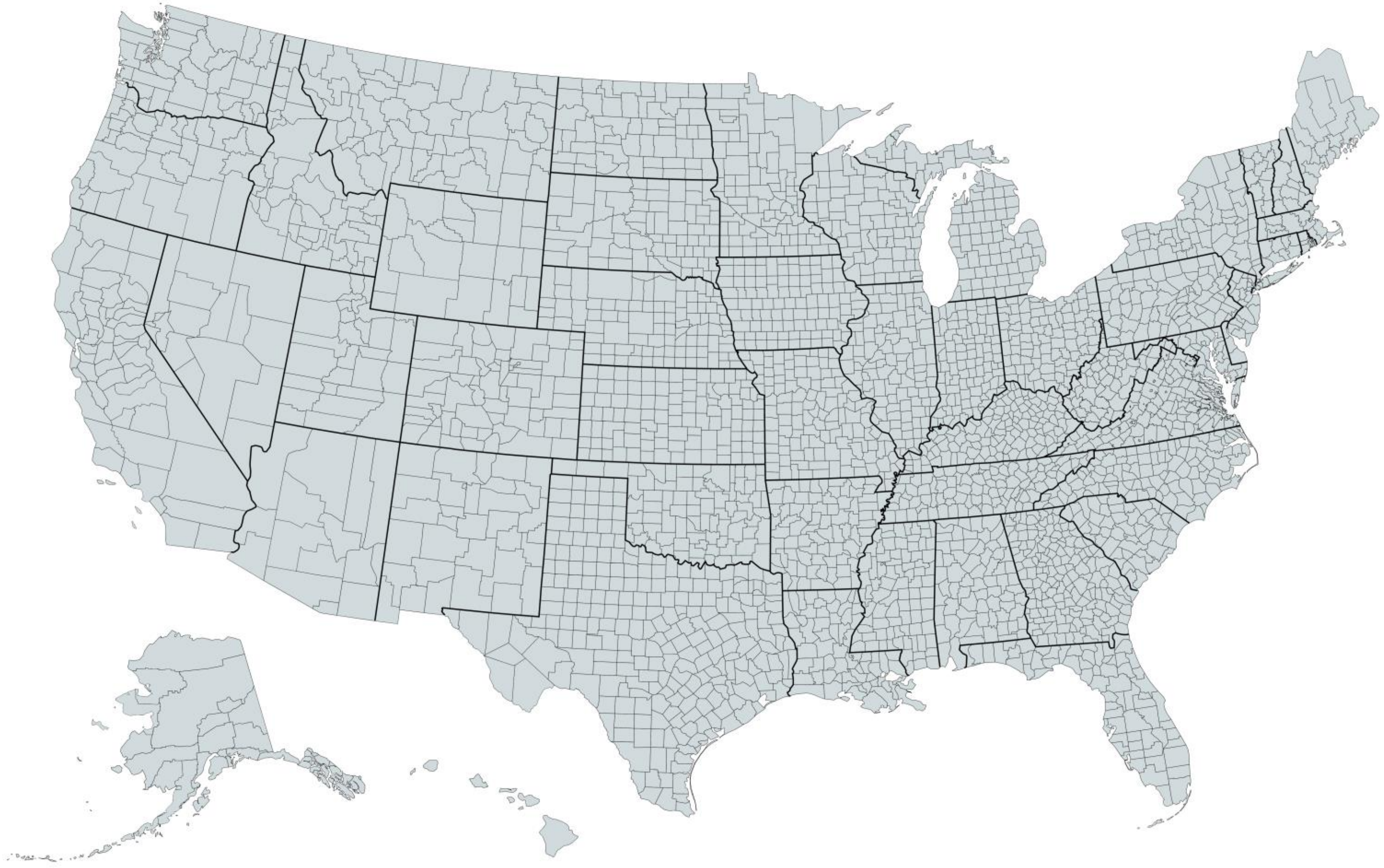
What are complex sample designs?

- Complex = sampling units ***do not*** have equal probability of selection
- Why?
  - **Stratification** = divide population into mutually exclusive strata *(e.g., age groups)*, then sample within strata *(e.g., young people)*
  - **Clustering** (single or multi-stage) = divide the population into clusters *(typically by geography – counties, neighborhoods, etc.)* then randomly select clusters
- These procedures inherently **violate assumptions of SRS**
  - **Independence of observations** *(e.g., **units within clusters can be similar** to one another in terms of demographics, political views, health behaviors and outcomes)*
  - **Equal probability of selection** *(e.g., young people are oversampled and have a much higher likelihood of ending up in the sample compared to older adults)*

Why are complex sample designs used?

- Reduce costs
- Improve precision for subgroup analysis *(e.g., young people, Black people, lower income people)*
- Improve statistical efficiency (variance estimation)

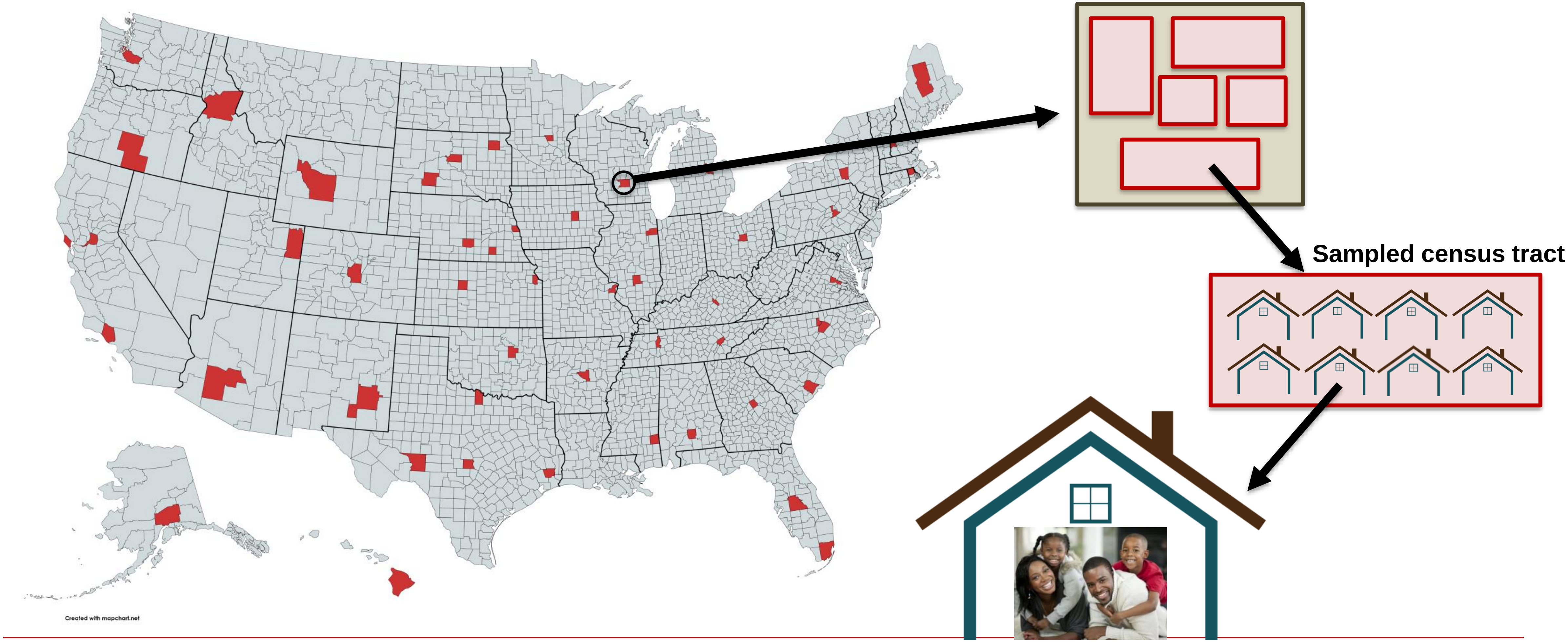
Example: Multi-stage cluster sampling of U.S. households



Created with mapchart.net

- Primary purpose: **Reduce total survey costs**
  - Travel and interviewer time/costs

Example: Multi-stage cluster sampling of U.S. households



Created with mapchart.net

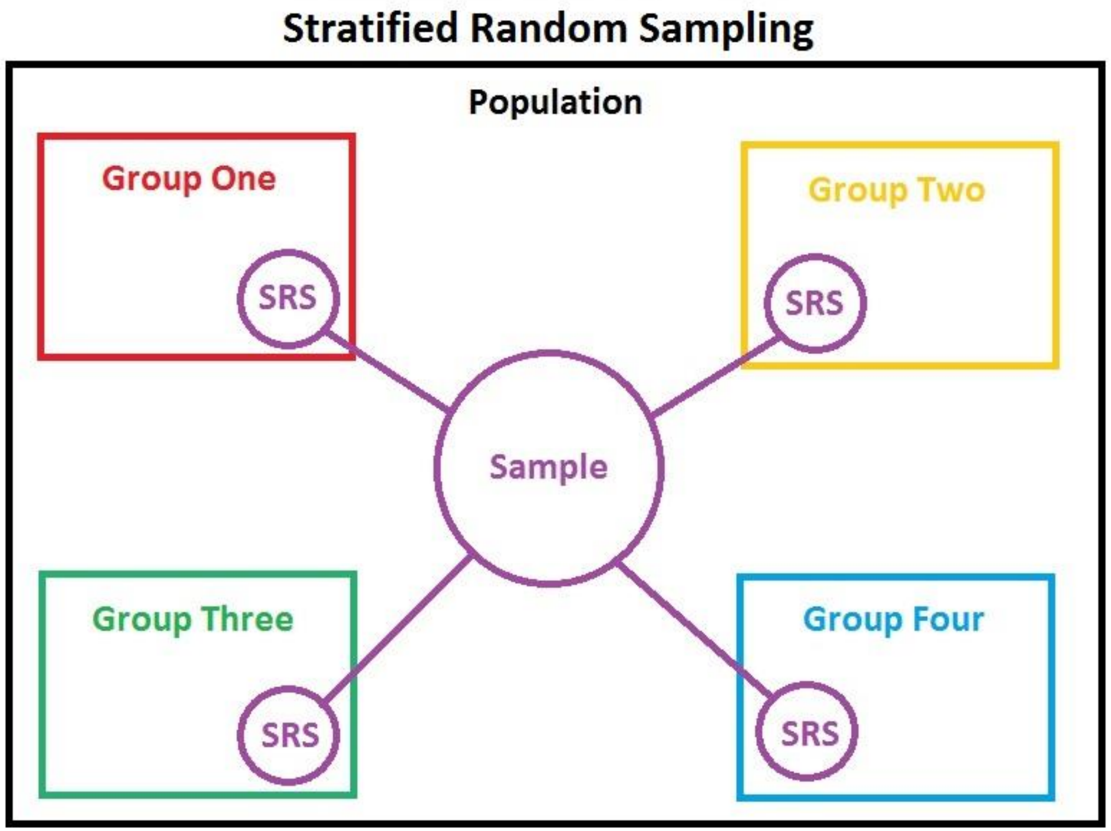
Sampled county

Sampled census tract

Sampled household

Example: Stratified sampling

- Primary purpose: **More efficient variance estimation** *(tighter standard errors / CIs)*
  - Can be combined with **oversampling:** Ensure sufficient sample sizes of population subgroups
  - Reduce necessary sample size (*effective sample size*)



Stratified sampling, design effects, & effective sample size

Population	Stratum 1 (Urban)	Stratum 2 (Rural)
Size N=1,000,000	N_1 = 200,000 (20%)	N_2 = 800,000 (80%)
Variance s^2 = 1,800,000	$s^2{}_1$ = 4,000,000	$s^2{}_2$ = 1,000,000
Mean 1,400	3,000	1,000

Design effect

$$D^2(y) = \frac{V_{st}(y)}{V_{srs}(y)} = \frac{1333}{1500} = 0.89$$

Effective sample size

$$\text{Effective } n = \frac{n}{D^2(y)} = \frac{1200}{0.89} = 1348$$

Simple random sample

variance calculation

n = 1200

$$V_{srs}(y) = \frac{s^2}{n} = \frac{1{,}800{,}000}{1200} = 1500$$

Stratified sampling

variance calculation

n_1 = 240; n_2 = 960 (total n = 1200)

$$V_{st}(y) = \frac{W_1{}^2 s_1{}^2}{n_1} + \frac{W_2{}^2 s_2{}^2}{n_2}$$

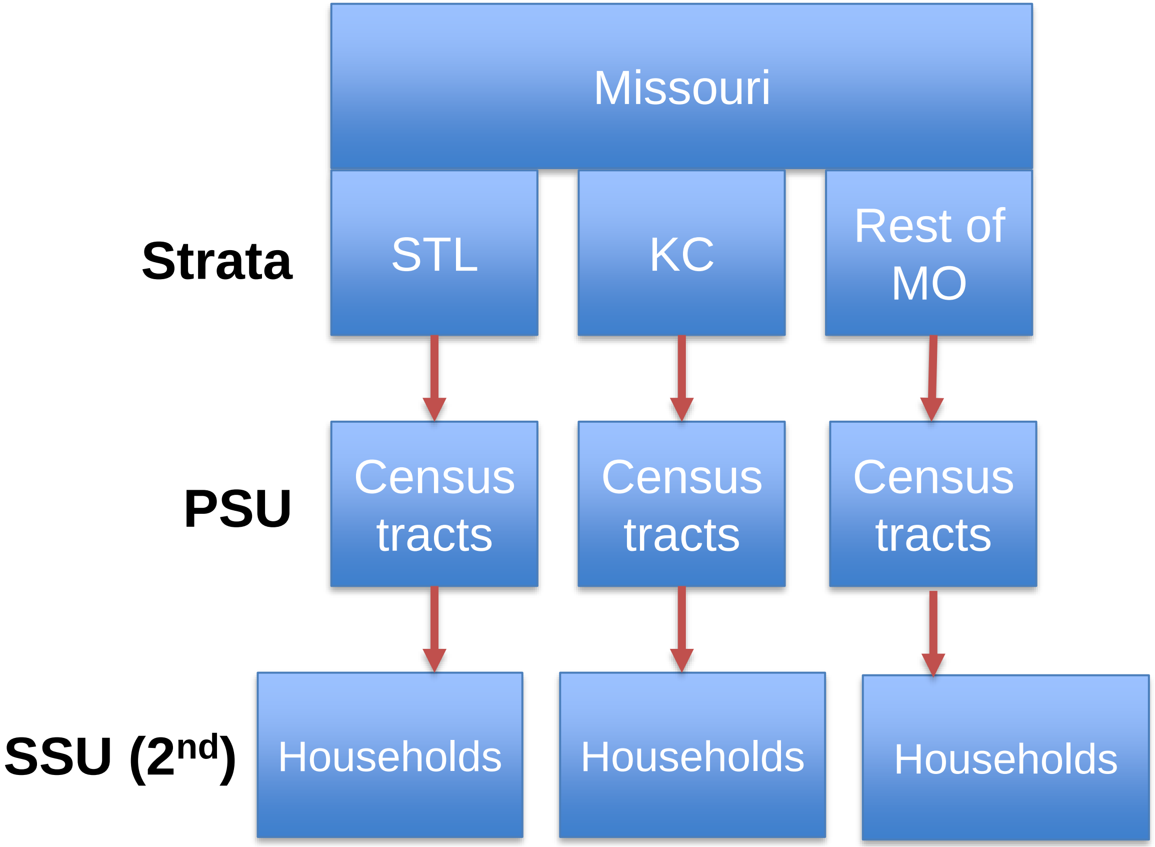
$$V_{st}(y) = \frac{(0.2)^2(4{,}000{,}000)^2}{240} + \frac{(0.8)^2(1{,}000{,}000)^2}{960} = 1333$$

More on stratified sampling: Over/under-sampling (n=25,000)

Strata	% of population	Sample size under SRS	Stratified sample size	% of stratified sample
Asian	5.9%	1,475	2,500	10%
Black	13.4%	3,350	5,000	20%
Hispanic	18.5%	4,625	5,000	20%
Native American or Alaska Native	1.3%	325	2,500	10%
Native Hawaiian	0.2%	50	2,500	10%
Multiracial	2.8%	700	2,500	10%
White	60.1%	15,025	5,000	20%

Common variables used in design-based analysis

- **Strata**
  - Mutually exclusive population subgroups groups (e.g., by geography, demographics, etc.)
- **Primary sampling units (PSUs)**
  - The first unit sampled in the design (typically after stratification)
- **Sampling (final) weights**
  - Inverse of the probability of being included in the sample given the sample design
  - Used to weight the sample back to the population of interest
  - Accounts for unequal probability of selection, unit non-response, and sampling frame errors



Missouri

Strata: STL, KC, Rest of MO

PSU: Census tracts, Census tracts, Census tracts

SSU (2nd): Households, Households, Households

GUIDED TUTORIAL:
NATIONAL SURVEY ON DRUG USE AND HEALTH

National Survey on Drug Use and Health

NSDUH

National Survey on Drug Use and Health

- **Substance Abuse and Mental Health Services Administration** (SAMHSA)
- **Aim:** Describe population trends and correlates of tobacco, alcohol, and drug use, and mental health and other health-related issues in the United States
- Annual, **cross-sectional** data collection

NSDUH Sample Design

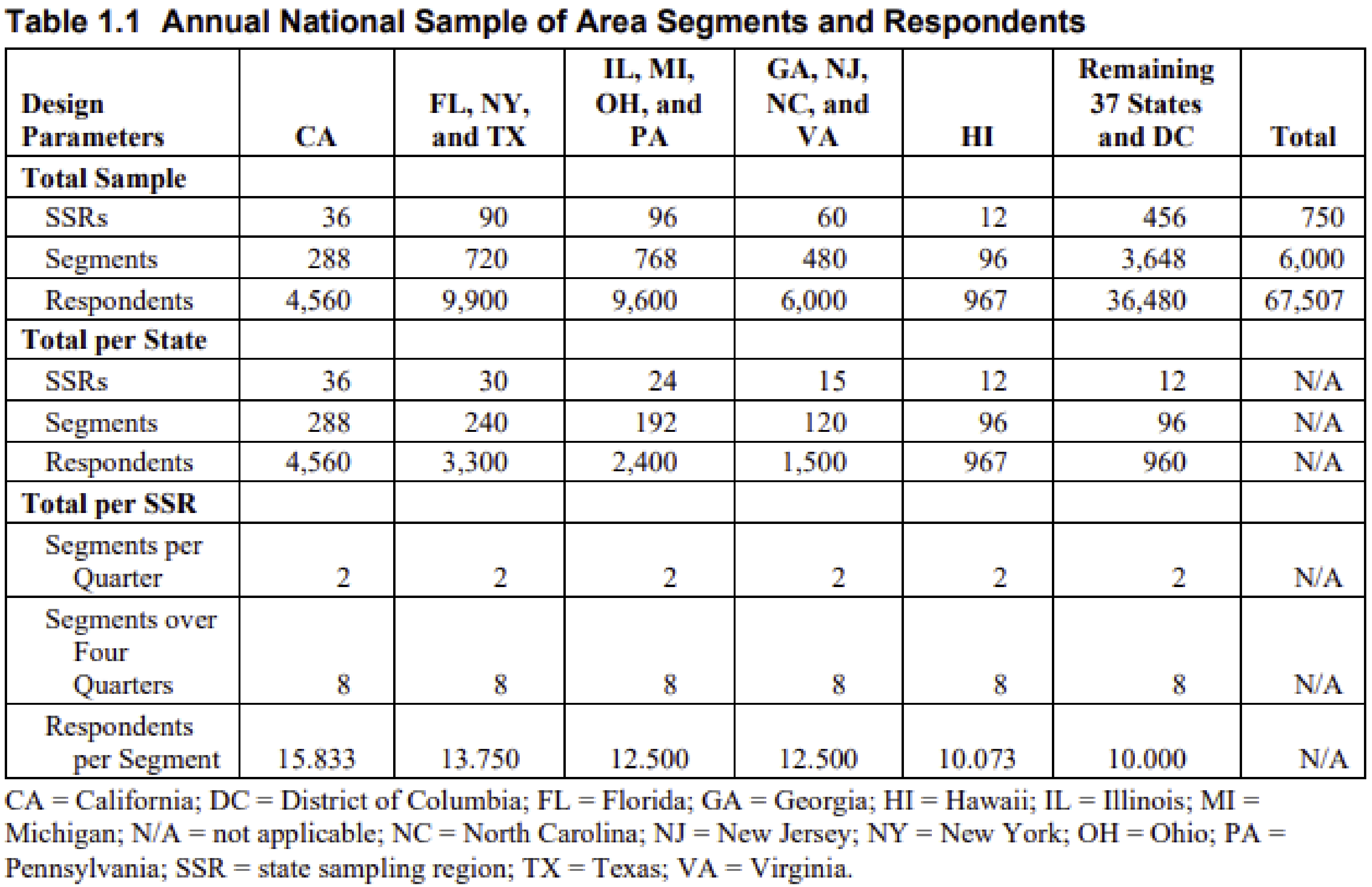


Table 1.1 Annual National Sample of Area Segments and Respondents

Design Parameters	CA	FL, NY, and TX	IL, MI, OH, and PA	GA, NJ, NC, and VA	HI	Remaining 37 States and DC	Total
Total Sample							
SSRs	36	90	96	60	12	456	750
Segments	288	720	768	480	96	3,648	6,000
Respondents	4,560	9,900	9,600	6,000	967	36,480	67,507
Total per State							
SSRs	36	30	24	15	12	12	N/A
Segments	288	240	192	120	96	96	N/A
Respondents	4,560	3,300	2,400	1,500	967	960	N/A
Total per SSR							
Segments per Quarter	2	2	2	2	2	2	N/A
Segments over Four Quarters	8	8	8	8	8	8	N/A
Respondents per Segment	15.833	13.750	12.500	12.500	10.073	10.000	N/A

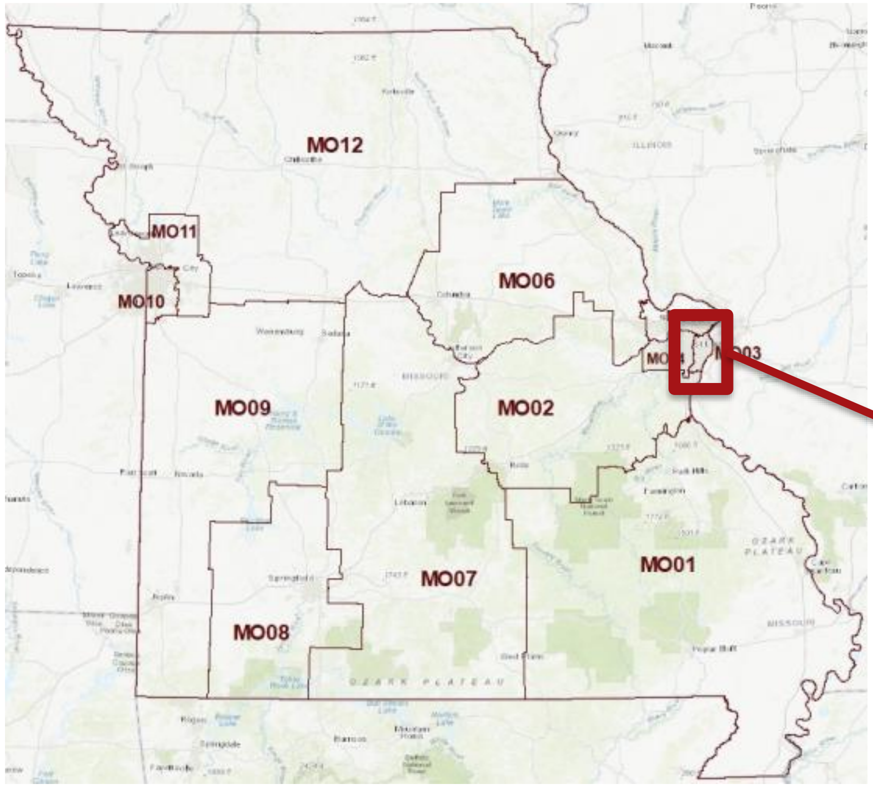
CA = California; DC = District of Columbia; FL = Florida; GA = Georgia; HI = Hawaii; IL = Illinois; MI = Michigan; N/A = not applicable; NC = North Carolina; NJ = New Jersey; NY = New York; OH = Ohio; PA = Pennsylvania; SSR = state sampling region; TX = Texas; VA = Virginia.

- **State-stratified, multi-stage cluster design**
- **SSRs** = equal-sized population regions within each state (made up of census tracts)
- **Segments** = census blocks (smallest units)
- **Respondents** = individuals within households selected to participate
- Larger states get **more SSRs, segments, and respondents**

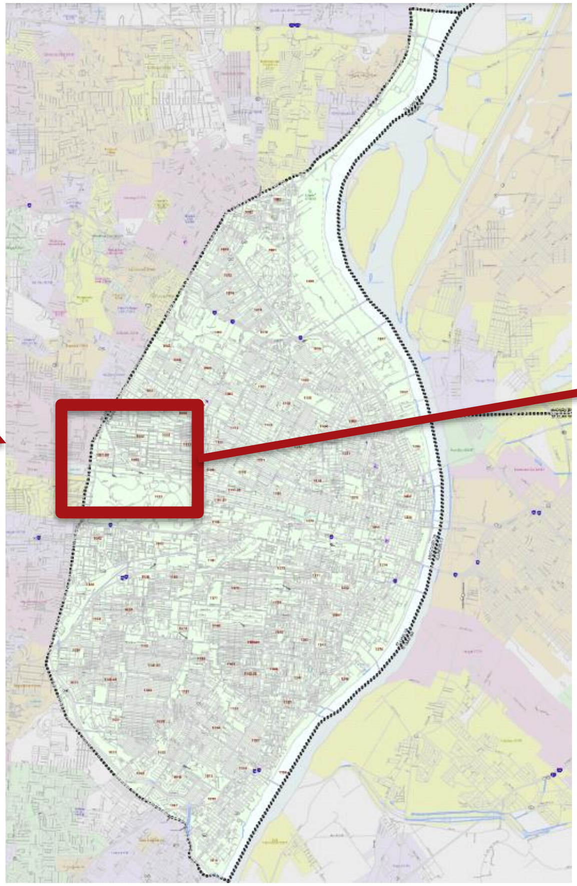
Source: 2019 NSDUH Sample Design Report, https://www.samhsa.gov/data/sites/default/files/reports/rpt34664/NSDUHmrbSampleDesign2019.pdf

NSDUH Sample Design in Missouri

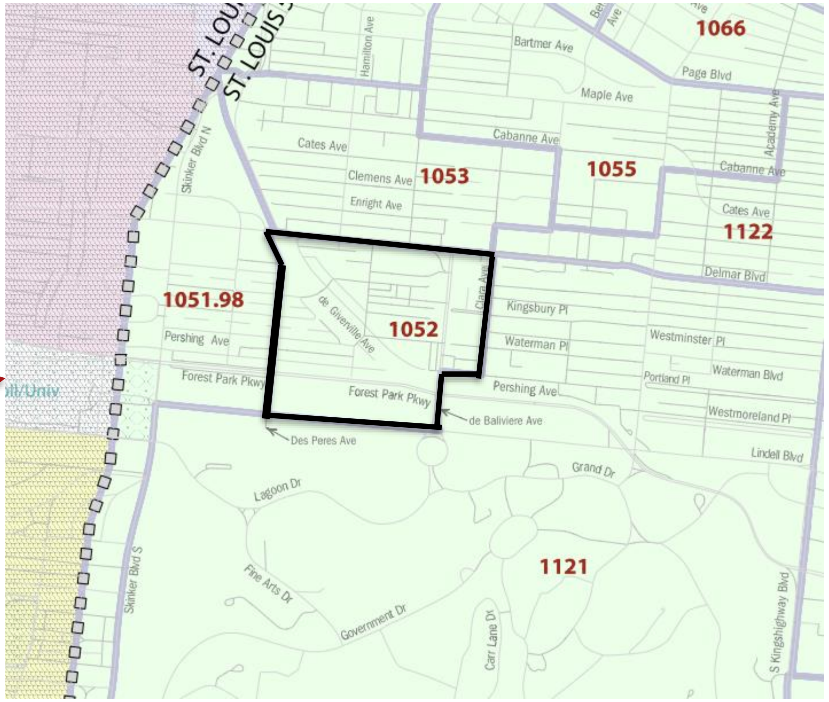
12 SSRs in Missouri



SSR MO03



Within each SSR,
select a census tract (1052)



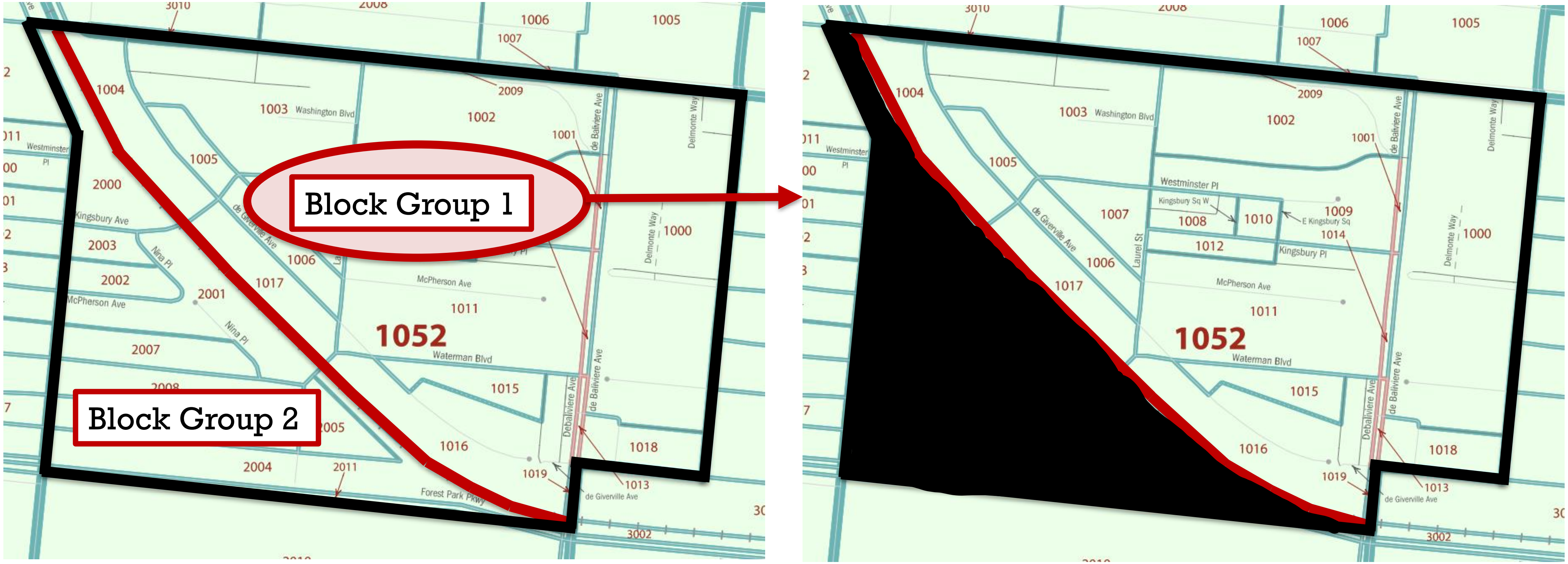
This is the PSU

Source: 2019 NSDUH Sample Design Report, https://www.samhsa.gov/data/sites/default/files/reports/rpt34664/NSDUHmrbSampleDesign2019.pdf

NSDUH Sample Design in Missouri

Within census tracts, select a census block group

Within the census block group, select 8 census blocks



Key steps to analyze NSDUH (and other CSS) data

- **Read the data documentation!!**
- Load/install packages (e.g., `survey` package in R)
  - Survey packages are automatically included in SAS and Stata
- Read the data into your stats program
- *For R and Stata:* Use `svydesign` (R) and `svyset` (Stata) to tell the program about the survey design elements
  - **If necessary: identify a subpopulation** to run unconditional analysis
- Analyze data using appropriate commands. Examples:
  - R: svymean
  - Stata: svy: mean
  - SAS: proc surveymeans

Why identify a subpopulation?

- We often want to make inferences for a specific group *(e.g., adolescents)*
- Item-level missing data is common

- **We should NOT drop participants/observations.** Why?
  - The program **cannot process *all possible*** design elements (strata and PSUs)
  - **Accurate variance estimation** requires full design information
    - Removes data used to calculate variance
    - Standard errors and confidence intervals will be incorrect → **incorrect inferences!**

Key variables to run design-based analysis

- **Strata**
- **Primary sampling unit** (PSU)
- **Sampling weight** (also called “person-level” or “analysis” or “final” weights)

<u>NSDUH:</u>

vestr

verep

analwt_c

Note: Read the data documentation to identify these!

Combining multiple waves of data

- Adjust the sampling/person-level weights prior to analysis
- Simple calculation!
  - **Divide the weight variable by the number of waves included**

- Ex: To analyze five waves of data (e.g., 2015-2019), you divide the "analwt_c" variable by 5:
  - **R:** nsduh$adjwt_5 <- nsduh$analwt_c/5
  - **Stata:** gen adjwt_5 = analwt_c/5
  - **SAS:** data nsduh;

```
          adjwt_5 = analwt_c/5
      run;
```

R: Analyzing NSDUH 2015-2019 data

1. Generate overall survey design object

```
nsduh_design <-
  svydesign(
    id = ~verep,
    strata = ~vestr,
    weights = ~adjwt_5,
    data = nsduh,
    nest = TRUE)
```

2. If needed: identify the subpopulation (e.g., Hispanic females)

```
nsduh <- nsduh %>%
    mutate(
      sp = factor(ifelse(
          hispanic==1 &  female==1), 1, 0))
```

3. Generate subpopulation survey design object

```
nsduh_design_sp <-
   subset(nsduh_design,
          sp==1)
```

Additional R resource: https://stats.oarc.ucla.edu/r/seminars/survey-data-analysis-with-r/

Stata: Analyzing NSDUH 2015-2019 data

Degrees of freedom = # of SSRs = 750

1. Inform Stata of the design elements (*svyset* command)

```
svyset verep [pw=adjwt_5], strata(vestr) dof(750)
```

If combining waves with unequal SSRs, use the lower amount for more conservative analysis

2. Conduct analysis with appropriate *svy* command

```
svy: means var1 var2 … var-n, options
svy: tab var1 var2, options
svy: reg dvar ivar1 ivar2 … IV-n, options
```

3. If needed: Identify the subpopulation with subpop(sp)

```
svy, subpop(sp): logistic dvar ivar1 ivar2… IV-n, options
```

Additional Stata resource: https://stats.oarc.ucla.edu/stata/seminars/svy-stata-8/

SAS: Analyzing NSDUH 2015-2019 data

- **SAS is a little different** – there is no survey object, so you must identify the design elements with each *survey* procedure.

General structure

```
proc surveymeans data = dataName;
  weight weightVar;
  cluster clusterVar;
  strata strataVar;
  domain subpopVar;
  var var1 var2 … var-n;
run;
```

NSDUH

```
proc surveymeans data = nsduh;
  weight adjwt_5;
  cluster verep;
  strata vestr;
  domain sp;
  var depression_score stress_score;
run;
```

Common SAS survey procedures: surveyfreq, surveymeans, surveyreg, surveylogistic

Additional SAS resource: https://stats.oarc.ucla.edu/sas/seminars/sas-survey/

Guided tutorial in R

- Guided R tutorial available on my GitHub account:
- **Zip file:** https://github.com/fhmcguire/NSDUH-complex-sample-analysis-tutorial/archive/refs/heads/main.zip
  - Extract all files to a folder on your computer
  - R Markdown and R dataset must be in the same folder!