Probability Sampling: Part 2

Brady T. West

"Complex" Probability Sampling

- SRS rarely conducted in practice; exception = relatively cheap data collection based on well-defined population lists
- With larger populations, **complex samples** often selected, where each sampled unit has known probability of selection

Complex = anything more complicated than SRS!

Complex Samples

- Complex samples have certain key features:
  - Population divided into different **strata**, and part of sample is allocated to each **stratum**; → ensures sample representation from each stratum, and reduces variance of survey estimates (**stratification**)
  - **Clusters** of population units (e.g., counties) are randomly sampled first (with known probability) within strata, to save costs of data collection (collect data from cases close to each other geographically)
  - **Units randomly sampled from within clusters**, according to some probability of selection, and measured

Complex Samples

- **A unit's probability of selection is determined by:**
  - Number of clusters sampled from each stratum
  - Total number of clusters in population in each stratum
  - Number of units ultimately sampled from each cluster
  - Total number of units in population in each cluster

Complex Samples

Example of finding a unit's probability of selection:

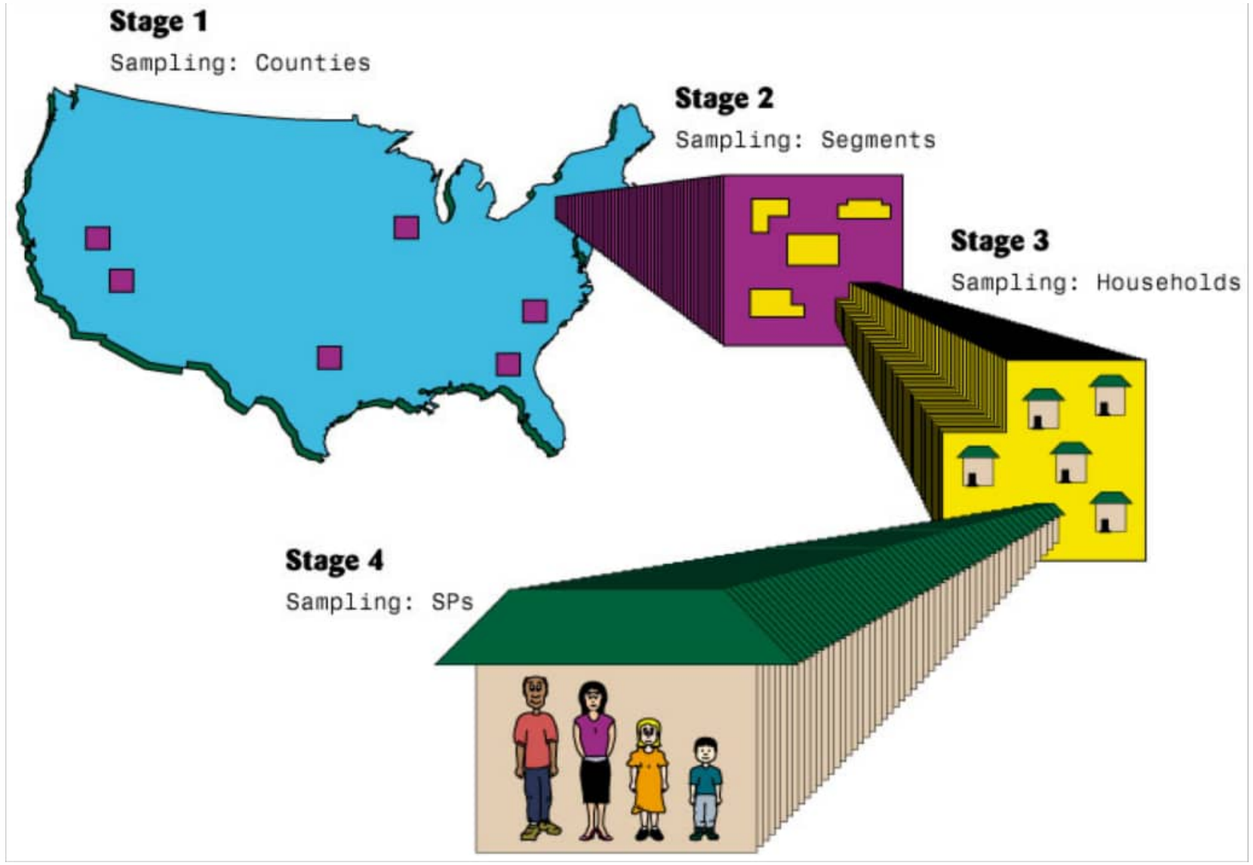
- Select a out of A clusters at random in a given stratum
- then select b out of B units at random from within a selected cluster

Probability of selection: $\left(\frac{a}{A}\right)\left(\frac{b}{B}\right)$

Example: NHANES

- Divide U.S. into different regions based on geography and population density (strata; increase representation!)
- Allocate some number of counties / groups of counties to be sampled from each stratum (clusters; saves costs!)
- Sample certain socio-demographic subgroups at higher rates within counties (oversampling: different probabilities of selection for different people!)

Example: NHANES



- Note multiple stages of random selection: counties (from strata), then area segments, then households, then people
- All random, all with known probabilities of selection

Image Credit: L. Mohadjer, Westat

Example: NHANES

- Drive a huge semi-trailer containing medical equipment and staff to each sampled county, and invite randomly selected people for an interview and a medical exam
- The inverse of a person's probability of selection is then their **sampling weight**
- If my probability is 1/100
  → my weight is 100
  I represent *myself*
  and **99 others** in the population!

Image Credit: Steven Heeringa, Institute for Social Research, University of Michigan

Example: NHANES

- Weights used to compute unbiased estimates of population quantities (e.g., mean BMI), accounting for different probabilities of selection.

- Probabilities of selection play a direct and essential role in computation of unbiased population estimates!

Why Probability Sampling?

- Having known, non-zero probability of selection for each unit in a population and subsequent random sampling ensures all units will have a chance of being sampled

- Probability sampling allows us to compute unbiased estimates, and also estimate features of the sampling distribution of estimates that we would see if many of the same types of probability samples were selected

Why Probability Sampling?

- Most importantly, probability sampling provides a statistical basis for making inferences about certain quantities in larger populations

- Next … learn more about non-probability sampling, and some difficulties with that approach (despite its popularity!)