

Complex Samples

Brady T. West

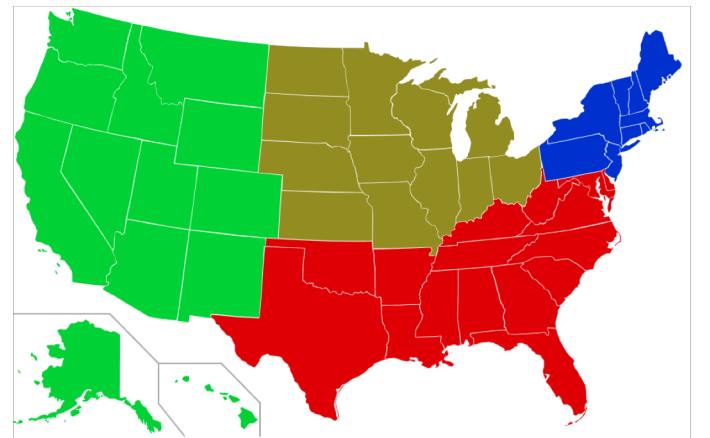
Lecture Overview

**Complex Sample = any probability sample
where design involves more than
Simple Random Sampling (SRS)!**

- More in-depth review of complex samples
- Discuss important considerations for making population inferences based on complex samples

Features of Complex Samples: Stratification

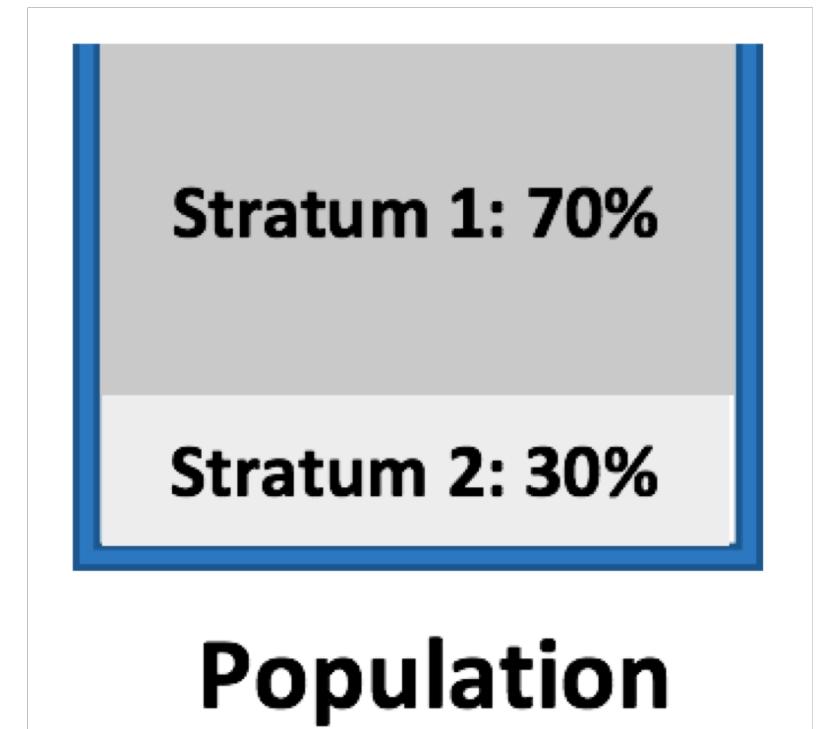
- **Stratification:** Allocation of overall sample to different “strata”, or mutually exclusive divisions of the population (e.g., regions of the United States)
- Several different allocation schemes are possible;
Aim → minimize sampling variance for particular variables given fixed costs



Features of Complex Samples: Stratification

Example: Proportionate Allocation

- If 70% of a population appears in one stratum and 30% in the other;
- Then 70% of the overall sample would be allocated to the first stratum, and 30% to the second



Features of Complex Samples: Stratification

- Stratification will eliminate between-stratum variance in means (or totals) on variable from the sampling variance!
- Important to account for stratification in analysis; else sampling variance may be artificially large → inferences too conservative, confidence intervals too wide!

Features of Complex Samples: Clustering

- **Clustering:** Random sampling of larger clusters of population elements, possibly across multiple stages (e.g., counties, then segments, then households)
- Reduces cost of data collection: expensive \$\$\$ to visit n randomly sampled units from large and widespread population

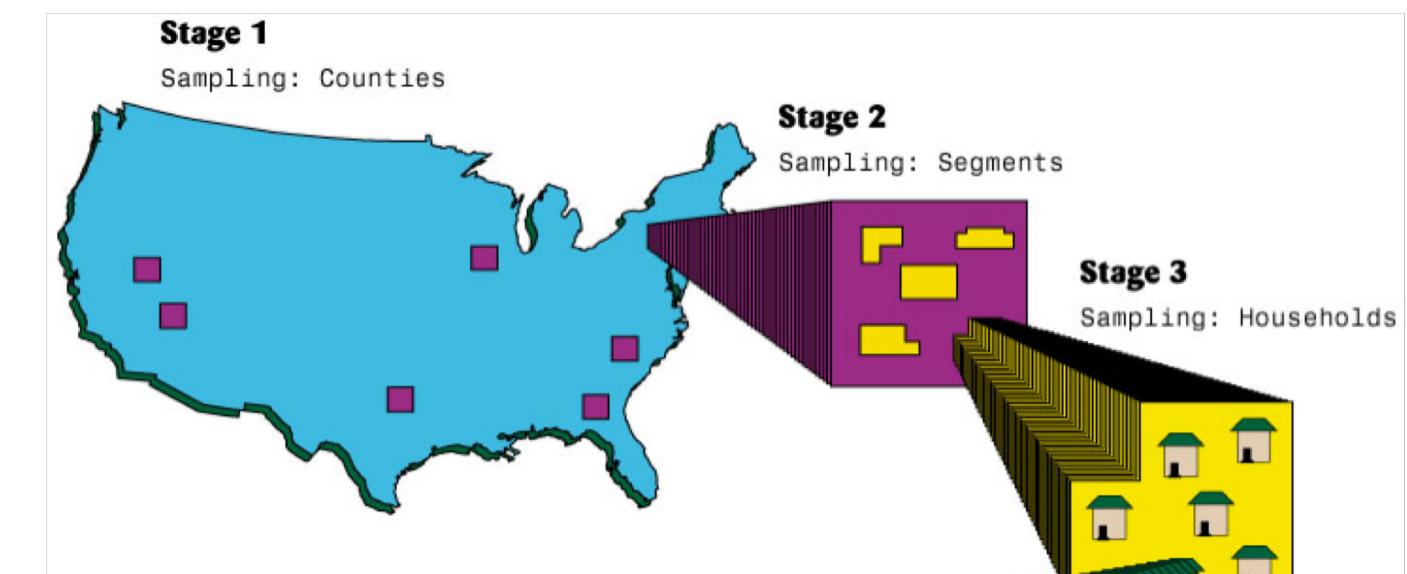


Image Credit: L. Mahadjer, Westat

Features of Complex Samples: Clustering

- Clustering *reduces costs* 😊
BUT tends to *increase sampling variance of estimates* 😞
Why? Units within same cluster have similar (correlated) Values on variables of interest → don't measure unique info!
- **Important** to account for cluster sampling in analysis, else inferences too *liberal*, confidence intervals too *narrow*!

Features of Complex Samples: Weighting

Complex samples are still probability samples, but if ...

- Multiple stages of cluster sampling within strata
 - Or certain subgroups sampled at higher rates
(oversampling)
- **Unequal probabilities of selection** for different units

Need to account for these unequal probabilities
to make **unbiased** population inferences

Features of Complex Samples: Weighting

- **How?** Use of **weights** in analysis ...
(partly) defined by **inverse of probability of selection**

If my probability is $1/100 \rightarrow$ my weight is 100,
I represent **myself** and **99 others** in the population!

Features of Complex Samples: Weighting

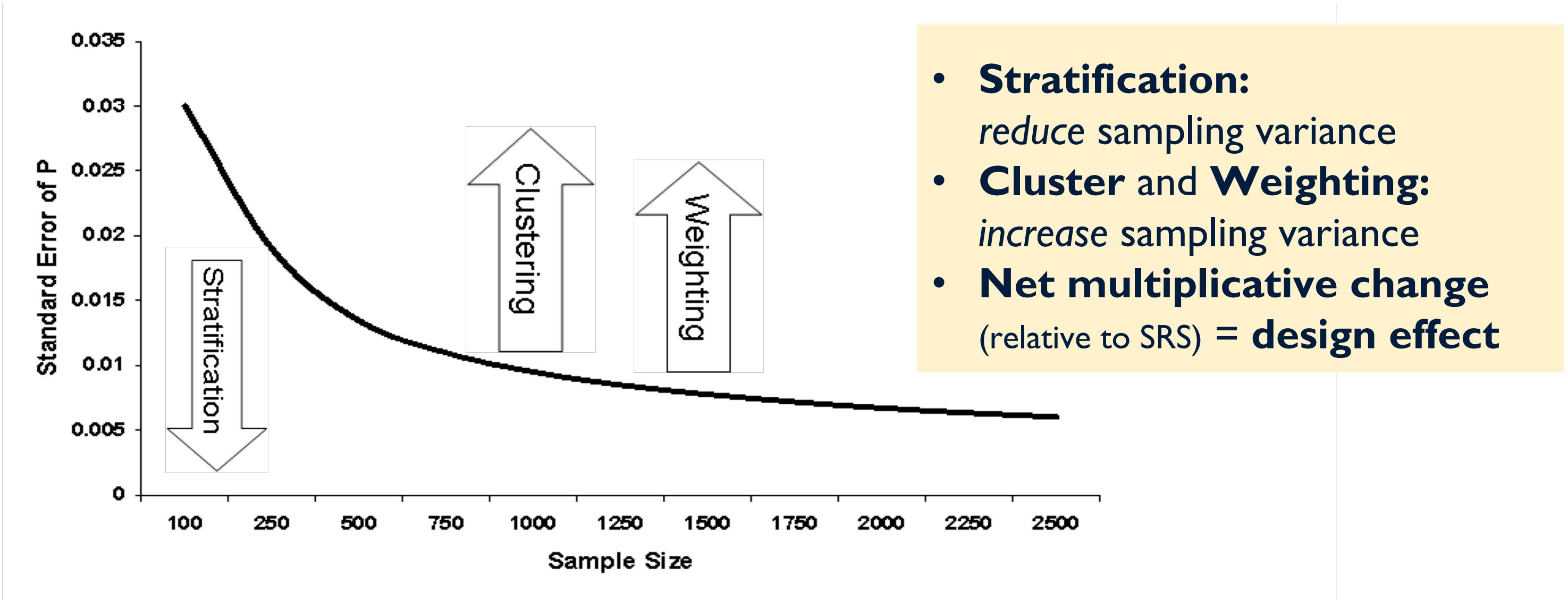
- Weights also **adjusted** for different probabilities of responding in different **subgroups**

If my probability of selection = 1/100
and I belong to subgroup where only 50% responded
→ my adjusted weight = $(1/0.01) \times (1/0.5) = 200$

Features of Complex Sampling: Weighting

- **Important** need to use weights so estimates are unbiased with respect to the sample design; else possible serious bias!
- **Drawback:** like cluster sampling, highly variable adjusted survey weights tend to increase sampling variance of weighted estimates (*even if they produce unbiased estimates!*)

Visualizing Design Effects



Source: Applied Survey Data Analysis (Heeringa et al., 2017)

Complex Samples in Analysis

- Most “survey analysis” procedures in statistical software compute unbiased point estimates (using final survey weights) and unbiased estimates of sampling variance (using stratum and cluster information, or *replicate sampling weights*)
- **Important** need to use appropriate software procedures, and identify all of these features to the software!

Analytic Error...

- Many secondary analysis of survey data collected from complex samples don't do this
→ can lead to biased inferences based on survey data
- Deeper Dive References:
 - <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0158120>
 - https://www.cdc.gov/pcd/issues/2018/17_0426.htm

Important: Look at Documentation!

- Focus = **looking at data** and understanding where data come from
- *Survey data*: Look at the documentation **before** the data!
- Documentation = what complex sampling performed, and what variables capture complex sampling features (weights, stratum codes, cluster codes)

Keywords indicating need to account for complex sampling:
multistage sampling, weights, stratification, cluster sampling, design effects

What's Next?

- **Later courses:** Analyses of survey data from complex samples, and methods in Python for computing unbiased (weighted) estimates and unbiased estimates of sampling variance
- **Deeper Dive Reference**
Applied Survey Data Analysis: <http://isr.umich.edu/src/smp/asda/>