Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 5b: Cluster sampling with equal cluster sizes

Cluster sampling

- Examples of cluster sampling?
- Why do cluster sampling?
- Units
  - Primary sampling units
  - Measurement units
- One-stage cluster sampling
- Two-stage cluster sampling with equal cluster sizes
- Two-stage cluster sampling with unequal cluster sizes

Cluster sampling in R

- Simulation of cluster sampling
- Analysis using the "survey" package

Cluster sampling: key principles

- Aim for all *measurement units* to be selected with equal probability
  - Probability sampling: all probabilities of selection are _____ and _____
  - Pr (you are selected) = Pr (your cluster is selected) $\times$ Pr (you are selected | your cluster is selected)
  - If you *don't* have equal probability of selection, fix using weights
- Take cluster-level averages and totals, then analyze data at the cluster level
- If you must analyze at the individual level, fit a multilevel model

Cluster sampling: requirements

- You need a sampling frame of clusters
- Clusters must be well-defined, non-overlapping, and exhaustive
- Once you have sampled the clusters, you need sampling frames for each of the *sampled* clusters
- All these rules can be broken, but then it's more complicated, won't discuss further here

Cluster sampling with equal cluster sizes

- First stage sampling:
  - Population of clusters $\alpha = 1, \ldots, A$
  - Sample of clusters $\alpha = 1, \ldots, a$
  - Sampling fraction of clusters $f_a = a/A$
- Second stage sampling:
  - Population: B items within each cluster
  - Sample: b items within each cluster
  - Sampling fraction within clusters $f_b = b/B$
- Population means $\overline{Y}_\alpha$ = average of B units in cluster α
- Sample means $\overline{Y}_\alpha$ = average of b *sampled* units in cluster α

Cluster sampling: inference

- Total sampling fraction $f = f_a f_b = \frac{ab}{AB}$
- For equal cluster sizes, the population mean $\overline{Y}$ is $\frac{1}{A}\sum_{\alpha=1}^{A}\overline{Y}_\alpha$, the mean of the cluster means
- For equal sample sizes within clusters, the sample mean $\bar{y}$ is $\frac{1}{a}\sum_{\alpha=1}^{a}\bar{y}_\alpha$, the mean of the cluster means
- Equal-probability sampling: $\bar{y}$ is an unbiased estimate of $\overline{Y}$
- Std err: $\sqrt{1-f}\,\frac{1}{\sqrt{a}}s_a$
  - f = sampling fraction
  - a = sample size of clusters
  - s_a = std dev of cluster averages

One-stage cluster sampling: efficiency

- Design effect
$$= \frac{\text{variance under cluster sampling}}{\text{variance under simple random sampling}} = \frac{(1-f)\frac{1}{a}S_a^2}{(1-f)\frac{1}{n}S^2} = B\frac{S_a^2}{S^2}$$
  - B = number of units in a cluster
  - S_a = std dev of cluster averages
  - S = std dev of the individual units
- Special cases
  - Clusters are randomly assigned: design effect should be _____
  - Clusters are completely homogeneous: design effect should be _____

Two-stage cluster sampling with equal cluster sizes

- For simplicity, assume: first stage SRS of clusters, second stage SRS of units within each cluster
- Sample a units, sample b units per cluster
- Sample size $n = ab$, sampling fraction $f = \frac{ab}{AB}$
- "Data" $\bar{y}_1, \ldots, \bar{y}_a$
- Estimate is $\bar{y} = \frac{1}{a} \sum_{\alpha=1}^{a} \bar{y}_\alpha$
- To estimate the variance, think of this as a SRS of size a from a population of $A\frac{B}{b}$ batches
- That is, each "batch" is of size b, there are $\frac{B}{b}$ batches per cluster, so there are a total of $A\frac{B}{b}$ batches in the population

Systematic sampling as cluster sampling

- Sample every 10th voter
- Consider this as a cluster sample of size 1 (out of 10 clusters)
- How to estimate the standard error?
  - If you assume simple random sampling, how will you go wrong?
  - Treat is as a stratified sample with 2 units per stratum

Cluster sampling: key principles (again)

- Aim for all *measurement units* to be selected with equal probability
  - Probability sampling: all probabilities of selection are _____ and _____
  - Pr (you are selected) = Pr (your cluster is selected) $\times$ Pr (you are selected | your cluster is selected)
  - If you *don't* have equal probability of selection, fix using weights
- Take cluster-level averages and totals, then analyze data at the cluster level
- If you must analyze at the individual level, fit a multilevel model