Design and Analysis of Sample Surveys

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

Class 6a: Cluster sampling with unequal cluster sizes

Cluster sampling

- Design
  - Cluster of interest or clusters of necessity or clusters of convenience
  - Sample sizes
- Analysis
  - Simple or weighted averages (when clusters aren't related to outcomes of interest)
  - Averages of cluster means (more statistically appropriate but more effort)
  - Regressions and multilevel models

2-stage cluster sampling with unequal cluster sizes

- Stratified cluster sampling
- Equal-probability sampling at both levels
- Difficulties with equal-probability sampling
- Adjusting for unequal probabilities of selection
- Probability-proportional-to-size sampling
- Approximate probability-proportional-to-size sampling

Stratified cluster sampling

- Stratification of clusters
- Why?
- Example: postal survey
  - Stratified cluster sampling of post offices by size
  - Sampling of mail within sampled post offices
- You could also do stratified sampling *within* clusters but that's not interesting

Equal-probability sampling at both levels

- Simple random sample: a out of A clusters
- Within each sampled cluster α, use a sampling fraction f_b (for example, 1/10)
- Pr (you are selected) = Pr (your cluster is selected) $\times$ Pr (you are selected | your cluster is selected) = $\frac{a}{A} f_b$
- Equal-probability sampling
- Analysis
  - Sample mean is a weighted average of cluster means
  - Cluster size = cluster weight

An idea that doesn't quite work

- Simple random sample: a out of A clusters
- Within each sampled cluster α, simple random sample of b units
- Pr (you are selected) = Pr (your cluster is selected) $\times$ Pr (you are selected | your cluster is selected) = $\frac{a}{A}\frac{b}{B_\alpha}$
- You're oversampling units in _____ clusters!

Probability-proportional-to-size sampling

- First step: sample clusters with probability proportional to size: Pr (cluster α is selected) $\propto N_\alpha$
- Second step: sample exactly 10 units within each sampled cluster: Pr (you are selected | your cluster is selected) = _____
- Pr (you are selected) = Pr (your cluster is selected) $\times$ Pr (you are selected | your cluster is selected) = _____
- Advantages of pps sampling
  - Equal-probability sampling
  - Quick and easy to analyze the data
- "Size" can have different meanings

Doing probability-proportional-to-size sampling

- Choose the number of clusters to sample and the number of units per sampled cluster
- Do the first-stage sampling *with replacement*
  - Why?
- Do the second-stage sampling
- Analyze the data
  - Estimating population averages and totals
  - Learning about large subsets of the population
  - Learning about small subsets
  - Fitting regression models

Approximate probability-proportional-to-size sampling

- Sample with probability proportional to a "measure of size," M_α (instead of the actual size N_α)
- Why?
- Sample size of b_α (rather than simply b) within sampled cluster α
- Pr (you are selected) = Pr (your cluster is selected) $\times$ Pr (you are selected | your cluster is selected) $\propto M_\alpha \frac{b_\alpha}{N_\alpha}$
- Equal-probability sampling if $b_\alpha \propto \frac{N_\alpha}{M_\alpha}$
- Otherwise, weights:
  - Unit weight $w_i = \frac{N_\alpha}{M_\alpha b_\alpha}$
  - Cluster weight $W_\alpha = \frac{N_\alpha}{M_\alpha}$
- Or do quick check to see if weighting is worth the trouble

Design effects in cluster sampling

- Once you have a probability-proportional-to-size sample, you can analyze it and think about it *as if* it were a one-stage cluster sample with equal cluster sizes
- Intraclass correlation and design effect
- Cluster-level predictors
- Geographic and demographic predictors as proxies for cluster

Applying the ideas

- Alcoholics Anonymous survey
- Stratified cluster sampling of post offices
- Sampling medical records from five boroughs
- Sampling fish from a lake
- Women in homeless shelters