BIOS 662 Fall 2018

Survey Sampling, Part III

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- One-stage cluster sampling
- Systematic sampling
- Multi-stage cluster sampling
- Comments on cluster sampling
- Sampling overview

Cluster Sampling

- Partition the population into exhaustive and mutually exclusive *primary units* or *clusters*
- Each primary unit is composed of *secondary units*
- Select a sample of primary units using some sampling design (e.g., SRS)
- Record the y values of *every* secondary unit within the selected primary units

Cluster Sampling

- This seems similar to stratification, with cluster = strata
- However, these are different designs
- In stratification, we sample some units from every stratum
- In cluster sampling, we sample *all* secondary units from some clusters
- This is sometimes called *one-stage cluster sampling* or *single-stage cluster sampling*

Cluster Sampling Examples

- In a household survey for a small city, a probability sample of blocks is selected. Each block in this case represents a cluster of households. All households within selected blocks are surveyed.
- In a survey of first graders in the schools of a state, a probability sample of schools is selected. All first graders in a school would represent a cluster in this design.
- In a national sample of inpatient hospital visits for individuals with multiple sclerosis during some calendar year, a probability sample of hospitals is chosen. Each hospital in this instance represents a cluster of visits by patients with multiple sclerosis during that year.

Notation and Estimands

- N is the number of primary units in the population
- n is the number of primary units in the sample
- M_i is the number of secondary units in primary unit i
- The total number of secondary units in the population is

$$M = \sum_{i=1}^{N} M_i$$

Notation and Estimands

- y_{ij} is the value of the variable of interest for secondary unit j of primary unit i
- $y_i = \sum_j y_{ij}$
- Population total: $\tau = \sum_{i=1}^{N} y_i = \sum_i \sum_j y_{ij}$
- Population mean per primary unit: $\mu_p = \frac{\tau}{N}$
- Population mean per secondary unit: $\mu = \frac{\tau}{M}$
- Let $Z_i = 1$ if primary unit i is selected, 0 otherwise

Estimators

- Assume SRS of primary units/clusters, also known as *simple cluster sampling*
- An unbiased estimator of τ is:

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{N} y_i Z_i = N\bar{y}$$

where $\bar{y} = \sum_i y_i Z_i / n$ is the sample mean of the primary unit totals

Estimators

- The variance of $\hat{\tau}$ is

$$\text{Var}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n}$$

where σ_u^2 is the finite population variance of the primary unit totals

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_p)^2$$

Estimators

- An unbiased estimator of the variance of $\hat{\tau}$ is

$$\widehat{\text{Var}}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n}$$

where s_u^2 is the sample variance of the primary unit totals

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 Z_i$$

Estimators

- These results follow directly from the SRS derivations, thinking of the clusters as units in a population of size N with variables $y_1, \ldots, y_N$

- An unbiased estimator of μ_p is given by $\bar{y} = \hat{\tau}/N$
- An unbiased estimator of μ is given by $\hat{\mu} = \hat{\tau}/M$
- Variances and unbiased estimators of variances follow accordingly

Cluster Sampling Principle

- Within-cluster variance does not affect variance of estimators
- Rather, it is only between-cluster variance that has an effect
- Thus, to minimize variance, clusters should be chosen to be as similar to one another as possible
- The ideal primary unit should be "representative", that is, contain the full diversity of the population (Thompson, page 118)
- This often runs counter to the practicalities of cluster sampling, e.g., where clusters are composed of geographically adjacent units

Systematic Sampling

- *Systematic Sampling*: A method of probability sampling in which elements on an ordered list are chosen by selecting elements a fixed distance apart on the list:

1. Number units in sampling frame sequentially from 1 to N
2. Choose a sampling interval k. If a sample of size about n is desired, k is usually the ratio, N/n, rounded to the nearest integer.
3. Choose a random number between 1 and k. This is called the *random start* and will be denoted by g.
4. Elements selected in the sample are those numbered g and every k^{th} element for the remainder of the list; that is, g, $g+k$, $g+2k$, etc.

Systematic Sampling

- Systematic sampling can be viewed as a special form of cluster sampling.
- Specifically, the population can be viewed as consisting of k clusters each of which is a possible systematic sample which can be chosen. By choosing a random start and applying a fixed interval in selecting the sample, we are effectively randomly choosing one of the k possible clusters.
- Thus we can obtain an unbiased estimator of the population total or mean. However, because this sample contains just one cluster, it is not possible to obtain unbiased estimators of the variances.

Multi-Stage Cluster Sampling

- *Multi-Stage Cluster Sampling*: A method of probability sampling in which the sample of elements is chosen in two or more stages. Second stage sampling units are chosen from the sampling units selected in the first stage. Third stage units are chosen from second stage sampling units; and so forth.

- Example: Household sample of the non-institutionalized population in Virginia

  Primary Sampling Units: Minor Civil Divisions

  Secondary Sampling Units: Small Groups of Blocks

  Tertiary Sampling Units: Households

Multi-Stage Cluster Sampling

- Example: National Sample of Hospital Discharges

  PSU: Small Groups of Counties

  SSU: Hospitals

  TSU: Patient Medical Records

- Example (Tate and Hudgens, *AJE*, 2007): Estimating the number of individuals at high risk for HIV in Osh, Kyrgyzstan

  PSU: Public venues within the city where risky sexual and drug-use behaviors occur

  SSU: Individuals socializing at these venues

Two-Stage Cluster Sampling

- We consider a two-stage design with SRS at each stage
- First stage: SRS of n primary units selected
- Second stage: SRS of m_i secondary units selected from the i^{th} selected primary unit, for $i = 1, \ldots, n$

- $\mu_i = y_i/M_i$ is the mean per secondary unit in the i^{th} primary unit
- Z_i is as before
- $Z_{ij} = 1$ if the j^{th} secondary unit of the i^{th} primary unit is in the sample, 0 otherwise

Two-Stage Cluster Sampling

- If the i^{th} primary unit is selected, an estimator of the total y-value for that unit (that is, y_i) is

$$\hat{y}_i = \frac{M_i}{m_i} \sum_{j=1}^{M_i} y_{ij} Z_{ij} = M_i \, \bar{y}_i$$

  where

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{M_i} y_{ij} Z_{ij}$$

- Because SRS is used at the second stage, this estimator is conditionally unbiased

$$E(\hat{y}_i | Z_i = 1) = y_i$$

Two-Stage Cluster Sampling

- An unbiased estimator of the population total is given by

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{N} \hat{y}_i Z_i$$

- To prove this, we use the fact that

$$E(\hat{\tau}) = E\Big(E(\hat{\tau}|Z_1, \ldots, Z_n)\Big)$$

Two-Stage Cluster Sampling

- First evaluate the inner expectation

$$E(\hat{\tau}|Z_1,\ldots,Z_n) = E\left(\frac{N}{n}\sum_{i=1}^{N}\hat{y}_i Z_i \middle| Z_1,\ldots,Z_N\right)$$

$$= \frac{N}{n}\sum_{i=1}^{N} E(\hat{y}_i|Z_i=1)Z_i$$

$$= \frac{N}{n}\sum_{i=1}^{N} y_i Z_i$$

- Then evaluate the outer expectation

$$E(\hat{\tau}) = E\left(\frac{N}{n}\sum_{i=1}^{N} y_i Z_i\right) = \frac{N}{n}\sum_{i=1}^{N} y_i E(Z_i) = \sum_{i=1}^{N} y_i = \tau$$

Two-Stage Cluster Sampling

- The variance of $\hat{\tau}$ is

$$\mathrm{Var}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$$

where

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_p)^2$$

(as before) and

$$\sigma_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \mu_i)^2$$

for $i = 1, \ldots, N$

Two-Stage Cluster Sampling

- Note that the first term in $\text{Var}(\hat{\tau})$ is the variance that would be obtained if every secondary unit in a selected primary unit is observed (that is, $m_i = M_i$ for all i). So the second term can be viewed as a penalty for having to estimate y_i.
- Similarly, note that the second term equals the $\text{Var}(\hat{\tau})$ when $n = N$, that is, every primary unit is selected. In this case, we recover the variance from stratified sampling. So the first term can be viewed as a penalty for using cluster sampling instead of stratified sampling.

Two-Stage Cluster Sampling

- To derive $\mathrm{Var}(\hat{\tau})$, we will use the fact

$$\mathrm{Var}(\hat{\tau}) = \mathrm{Var}\big(E(\hat{\tau}|Z_1,\ldots,Z_N)\big)+E\big(\mathrm{Var}(\hat{\tau}|Z_1,\ldots,Z_N)\big)$$

- For the first term, we have

$$\mathrm{Var}\big(E(\hat{\tau}|Z_1,\ldots,Z_N)\big) = \mathrm{Var}\left(\frac{N}{n}\sum_{i=1}^{N} y_i Z_i\right) = N(N-n)\frac{\sigma_u^2}{n}$$

where the second equality follows from results for SRS

- To evaluate the second term, first note that

$$\mathrm{Var}(\hat{y}_i|Z_i = 1) = M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$$

Two-Stage Cluster Sampling

- Therefore

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}|Z_1,\ldots,Z_n) &= \mathrm{Var}\left(\frac{N}{n}\sum_{i=1}^{N}\hat{y}_i Z_i \middle| Z_1,\ldots,Z_N\right) \\
&= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N}\mathrm{Var}(\hat{y}_i|Z_i=1)Z_i \\
&= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N} M_i(M_i-m_i)\frac{\sigma_i^2}{m_i}Z_i
\end{aligned}$$

- Thus

$$E\big(\mathrm{Var}(\hat{\tau}|Z_1,\ldots,Z_n)\big) = \frac{N}{n}\sum_{i=1}^{N} M_i(M_i-m_i)\frac{\sigma_i^2}{m_i}$$

Two-Stage Cluster Sampling

- An unbiased estimator of the variance of $\hat{\tau}$ is

$$\widehat{\text{Var}}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{s_i^2}{m_i}Z_i$$

where

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{N}(\hat{y}_i - \hat{\mu}_p)^2 Z_i$$

and

$$s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{y}_i)^2 Z_{ij}$$

for $i = 1, \ldots, N$

- The proof is left as an exercise

Two-Stage Cluster Sampling

- Estimators for population means follow immediately:

$$\hat{\mu}_p = \hat{\tau}/N \quad \text{is unbiased for} \quad \mu_p$$

$$\hat{\mu} = \hat{\tau}/M \quad \text{is unbiased for} \quad \mu$$

- Variance expressions follow from $\text{Var}(\hat{\tau})$ divided by the appropriate constant

Two-Stage Cluster Sampling: Example

- SRS of $n = 3$ primary units selected from a population of $N = 100$ primary units
- For each of the selected primary units, SRS of $m_i = 2$ secondary units selected
- Sizes of the three selected primary units: 24, 20, 15
- Y-values for the first selected primary unit: 8, 12
- Y-values for the second selected primary unit: 0, 0
- Y-values for the third selected primary unit: 1, 3

Two-Stage Cluster Sampling: Example

- Estimate of the population total:

$$\hat{\tau} = \frac{100}{3}\left(24 \cdot \frac{8+12}{2} + 20 \cdot \frac{0+0}{2} + 15 \cdot \frac{1+3}{2}\right) = 9{,}000$$

- Estimate of the mean per primary unit:

$$\hat{\mu}_p = \frac{\hat{\tau}}{N} = 90$$

- Sample variance across primary unit totals:

$$s_u^2 = \frac{1}{3-1}\left((240-90)^2 + (0-90)^2 + (30-90)^2\right) = 17{,}100$$

Two-Stage Cluster Sampling: Example

- After computing the sample variances within the selected primary units, we have

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\tau}) &= 100(100-3)\frac{17100}{3} \\
&\quad + \frac{100}{3}\left(24(24-2)\frac{8}{2} + 20(20-2)\frac{0}{2} + 15(15-2)\frac{2}{2}\right) \\
&= 55{,}366{,}900
\end{aligned}$$

Comments on Cluster Sampling

- Simple cluster sampling is epsem
- One-stage cluster sampling generally yields estimates with relatively larger variances (i.e., lower precision) than samples of the same size which are chosen by (individual) element (i.e., non-cluster) sampling. The amount of the increase in variance is directly related to the average sample cluster size.

Comments on Cluster Sampling

- Because units of clusters are often close in geographic proximity, the average cost per sample element can be reduced substantially over individual element sampling if cluster sampling is used
- The size of the cost reduction is directly related to the average size of the clusters that are used
- Elements in a cluster are usually similar (i.e., clusters are internally homogeneous), so the amount of information gathered by the survey may not be increased substantially as additional units are surveyed within a cluster
- So sample cluster sizes should not be too large

Comments on Cluster Sampling

- As a general rule, the number of clusters in the population should be large which means that the average size of clusters should be kept as small as possible.
- The survey statistician frequently has some choice in the size of clusters that are used in a survey
- In making this choice, the cost advantages of large (sample) clusters must be properly weighed against the statistical advantages of smaller (sample) clusters

Comments on Cluster Sampling

- Cluster sampling eliminates the need for a sampling frame consisting of a list of all elements in the population
- Because clusters are the units being sampled, a listing of all clusters in the population constitutes an appropriate frame
- Through multi-stage cluster sampling, most of the cost savings can be retained while gaining back some of the statistical losses (i.e., larger variances) of one-stage cluster sampling

Summary

- Identified several basic sampling designs (on next slide)
- Derived properties (expectations, variances, ...) of various estimators
- Illustrated with real data sets

- 664 [164] SAMPLE SURVEY METHODOLOGY (STAT 358) (3). Prerequisite, BIOS 550 or equivalent or permission of the instructor. Fundamental principles and methods of sampling populations, with primary attention given to simple random sampling, stratified sampling, and cluster sampling. Also, the calculation of sample weights, dealing with sources of nonsampling error, and analysis of data from complex sample designs are covered. Practical experience in sampling is provided by student participation in the design, execution, and analysis of a sampling project. Spring.

Summary

- SRS
- Stratified
  - *Proportionate* – default; always better than SRS
  - *Optimal, disproportionate, balanced*
- Cluster
  - *One stage* – SRS of clusters; sample all within cluster
  - *Systematic (list)*
  - *Multi-stage*, for example, blocks then dwellings