SAMPLING III
BIOS 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2009-08-11 09:52

Outline

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

Cluster Sampling

- Partition 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 y-values of *every* secondary unit within selected primary units

Cluster Sampling

- Seems similar to stratification with cluster = strata
- However these are different designs
- In stratification, we sample some units from all strata
- 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 are sampled within selected blocks.
- 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 the number of primary units in the population
- n the number of primary units in the sample
- M_i the number of secondary units in primary unit i
- Total number of secondary units in the population

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

Notation and Estimands

- y_{ij} 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 selected, 0 otherwise

Estimators

- Assume SRS of primary units/clusters aka *simple cluster sampling*
- Unbiased estimator of τ

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

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

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

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

  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

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

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

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

- These results follow directly from the SRS derivations, thinking of the clusters as units in a population of size N with variables $y_1, \ldots, y_N$
- Unbiased estimator of μ_p is given by $\bar{y} = \hat{\tau}/N$; unbiased estimator of μ is given by $\hat{\mu} = \hat{\tau}/M$; variances and unbiased variances thereof follow accordingly

Cluster Sampling Principle

- Within cluster variance does not effect 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 as possible
- The ideal primary unit should be "representative", i.e., contain the full diversity of the population (Thompson p 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 of population units are chosen by applying an interval of constant length after a random start

Selecting a Systematic Sampling

- A. The sampling frame consists of a list numbered sequentially from 1 to N
- B. A sampling interval (denoted by k) is chosen. If a sample of about n out of N elements is desired, k is usually the ratio, N/n, rounded to the nearest integer.
- C. A random number between 1 and k is chosen. This number is called the *random start* and will be denoted by the symbol, g.
- D. Elements selected in the sample are those numbered g and every k-th element for the remainder of the list; i.e., 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, since we are sampling only 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 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

- Here we consider a two-stage design with SRS at each stage
- In first stage, SRS of n primary units selected
- In second stage, SRS of m_i secondary units is selected from the ith selected primary unit, for $i = 1, \ldots, n$

- $\mu_i = y_i / M_i$ mean per secondary unit in the ith primary unit
- Z_i as before; $Z_{ij} = 1$ if jth secondary unit of ith primary unit in the sample, 0 otherwise

Two-Stage Cluster Sampling

- If the ith primary unit is selected, an estimator of the total $y-$value (i.e., 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 $\bar{y}_i = \sum_{j=1}^{M_i} y_{ij} Z_{ij} / m_i$

- Since SRS 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 $E(\hat{\tau}) = E\{E(\hat{\tau}|Z_1,\ldots,Z_n)\}$
- 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 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

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

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 first term in $Var(\hat{\tau})$ is the variance that would be obtained if every secondary unit in a selected primary unit were observed (i.e. $M_i = m_i$ for all i); So second term can be viewed penalty for having to estimate y_i
- Similarly, note the second term equals the $Var(\hat{\tau})$ when $N = n$, i.e., every primary unit was 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 $Var(\hat{\tau})$, we will use the fact

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

- For the first term, we have

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

  where the second equality follows from results derived for SRS

- To evaluate the second term, first note

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

Two-Stage Cluster Sampling

- Therefore

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

$$= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N} Var(\hat{y}_i|Z_i=1)Z_i$$

$$= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N} M_i(M_i-m_i)\frac{\sigma_i^2}{m_i}Z_i$$

- Thus

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

Two-Stage Cluster Sampling

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

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

  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 $i = 1, \ldots, N$

- Proof is left for the reader

Two-Stage Cluster Sampling

- Estimators for population means follow immediately

  $\hat{\mu}_p = \hat{\tau}/N$ is unbiased for μ_p

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

- Variance expressions follow from $Var(\hat{\tau})$ 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
- The size of the three selected primary units are 24, 20, and 15
- Y-values for first selected primary unit are 8, 12
- Y-values for second selected primary unit are 0, 0
- Y-values for third selected primary unit are 1, 3

Two-Stage Cluster Sampling: Example

- Estimate of population total is

$$\hat{\tau} = \frac{100}{3}(24 * \frac{8+12}{2} + 20 * \frac{0+0}{2} + 15 * \frac{1+3}{2}) = 9000$$

- Estimate of the mean per primary unit

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

- Sample variance between primary unit totals

$$s_u^2 = \frac{1}{3-1}\{(240-90)^2 + (0-90)^2 + (30-90)^2\} = 17100$$

Two-Stage Cluster Sampling: Example

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

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

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 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 element sampling if cluster sampling is used. The amount of the reduction in costs is directly related to the average size of the clusters that are used.
- Since elements in clusters are usually similar (i.e., clusters are internally homogeneous), the amount of information gathered by the survey may not be increased substantially as new measurements are taken within clusters. This tells us that sample cluster sizes should not be too large. 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.

Comments on Cluster Sampling

- 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.
- Cluster sampling eliminates the need for a sampling frame consisting of a list of all elements in the population. Since 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.

Overview

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

- 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.

Overview

- SRS
- Stratified
  - *Proportionate* - default; always better than SRS
  - *Optimal*, *Disproportionate*, *Balanced*
- Cluster
  - *One stage* - SRS of clusters; sample all w/in cluster
  - *Systematic (list)*
  - *Multistage* - Eg Blocks then dwellings