Statistics 480: Survey Sampling Techniques

Walter R. Mebane, Jr.

University of Michigan

GSIs: Adam Hall and Fabricio Vasselai
B844 East Hall, 05 March 2018

Outline

cluster sampling

- cluster sample: probability sample in which each sampling unit is a set of elements
- sets or collections of elements are called clusters
- examples of clusters:
  - city blocks (ties to U.S. Census data)
  - households (when individual people are the elements)
  - phone numbers (when individual people are the elements)
  - the set of elements at an iterated common offset in a list of elements: all the elements whose numbering i in the list satisfies $(i - s)/k = 0$, where $k = N/n$ and s is a random start selected from $\{1, 2, \ldots, k\}$ (systematic sampling)
  - others?

cluster sampling: general rationale

- elements that are physically near one another may tend to be similar to one another
- characteristics are often positively correlated among physically proximate elements
- so the amount of information gained about the population from elements in a cluster may not be as much as that gained from physically separated elements
- the properties of a cluster sample depend on how similar to one another the units in each cluster are: the more internally heterogeneous each cluster is, the better
  - opposite from stratified random sampling, in which internally homogeneous strata are desirable
  - ideally, different clusters are exactly similar to one another

cluster sampling

- Example 8.1: sociologist estimating per-capita income in a city has no list of resident adults but does have a map
  - the map shows $N = 413$ rectangular city blocks, two industrial parks and three parks
  - the budget can support data collection in $n = 25$ clusters
  - discuss the sample survey design

cluster sampling: notation

- N: the number of clusters in the population
- n: number of clusters selected by simple random sampling
- m_i: number of elements in cluster i, $i = 1, \ldots, N$
- $\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$: average cluster size in the sample
- $M = \sum_{i=1}^{N} m_i$: number of elements in the population
- $\bar{M} = M/N$: average cluster size in the population
- y_i: total of all observations in cluster i
  - notice: all the elements in each sampled cluster are measured

cluster sampling: estimators

- estimators resemble those for simple random sampling (ratio estimation), except using cluster totals not element values
- estimator for the population mean μ

$$\bar{y} = \sum_{i=1}^{n} y_i \Bigg/ \sum_{i=1}^{n} m_i$$

- estimated variance of $\bar{y}$

$$\hat{V}(\bar{y}) = \frac{1}{\bar{M}^2} \left(\frac{N-n}{N} \right) \frac{s_r^2}{n}$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y} m_i)^2}{n-1}$$

- estimate $\bar{M}$ by $\bar{m}$ if M is unknown

cluster sampling: estimators

- in the estimated variance of $\bar{y}$, notice that

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y} m_i)^2}{n-1}$$

  is the formula used in ratio estimation (with i indexing elements, not clusters):

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - r x_i)^2}{n-1}$$

- what's the intuitive justification for treating cluster total y_i as proportional to m_i?
- what if it's not?

cluster sampling: estimators

- Example 8.2: sociologist (from Example 8.1) studying income in a city uses sample of $n = 25$ clusters from population of $N = 415$; given data, estimate per-capita income and bound the error of estimation
- see file `example8.2.R`

cluster sampling: estimators for the total

- estimator that depends on M: $M\bar{y} = M \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} m_i$
- estimated variance of $M\bar{y}$

$$\hat{V}(M\bar{y}) = M^2 \hat{V}(\bar{y}) = N^2 \left(\frac{N-n}{N}\right) \frac{s_r^2}{n}$$

- estimator that does not depend on M: $N\bar{y}_{\mathrm{t}} = N \left(\frac{1}{n} \sum_{i=1}^{n} y_i\right)$
- estimated variance of $N\bar{y}_{\mathrm{t}}$

$$\hat{V}(N\bar{y}_{\mathrm{t}}) = N^2 \hat{V}(\bar{y}_{\mathrm{t}}) = N^2 \left(\frac{N-n}{N}\right) \frac{s_{\mathrm{t}}^2}{n}$$

$$s_{\mathrm{t}}^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

cluster sampling: estimators

- Example 8.3: sociologist (from Example 8.1) studying income in a city uses sample of $n = 25$ clusters from population of $N = 415$; given data, estimate total income given $M = 2500$ and bound the error of estimation
- see file `example8.34.R`
- Example 8.4: sociologist (from Example 8.1) studying income in a city uses sample of $n = 25$ clusters from population of $N = 415$; given data, estimate total income assuming M is not known and bound the error of estimation
- see file `example8.34.R`

cluster sampling: relative efficiency

- comparison to SRS when $m_i = m$ (all clusters are the same size)
- $m_i = m$ for all i
  - implies $M = Nm$
  - implies

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

    is an unbiased estimator for μ
  - implies $\hat{V}(\bar{y})$ in an unbiased estimator for $V(\bar{y})$
  - implies $M\bar{y} = N\bar{y}_{\mathrm{t}}$

cluster sampling: relative efficiency

- if $m_i = m$ for all i, then, using y_{ij} to denote observation j from cluster i,

$$\bar{y} \equiv \bar{\bar{y}}_{\mathrm{c}} = \frac{1}{m}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

$$\hat{V}(\bar{\bar{y}}_{\mathrm{c}}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{nm^2}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(y_i - \bar{y}_{\mathrm{t}})^2$$

  where $\bar{y}_{\mathrm{t}} = m\bar{\bar{y}}_{\mathrm{c}}$

- if $m_i = m$ for all i, then using $\bar{y}_i = y_i/m$,

$$\frac{1}{m^2 n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y}_{\mathrm{t}})^2 = \frac{1}{m^2 n(n-1)}\sum_{i=1}^{n}(m\bar{y}_i - m\bar{\bar{y}}_{\mathrm{c}})^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}}_{\mathrm{c}})^2$$

cluster sampling: relative efficiency

- if $m_i = m$ for all i, then

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{\bar{y}}_c)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2 + m\sum_{i=1}^{m}(\bar{y}_i - \bar{\bar{y}}_c)^2$$

$$\mathrm{SST} = \mathrm{SSW} + \mathrm{SSB}$$

where SST denotes the total sum of squares, SSW denotes the within-cluster sum of squares, and SSB denotes the between-cluster sum of squares

cluster sampling: relative efficiency (with $m_i = m$ for all i)

- define the between-cluster mean square (MSB) and the within-cluster mean square (MSW) as

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{n-1} = \frac{m}{n-1}\sum_{i=1}^{m}(\bar{y}_i - \bar{\bar{y}}_c)^2$$

$$\mathrm{MSW} = \frac{\mathrm{SSW}}{n(m-1)} = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2$$

- we have

$$\hat{V}(\bar{\bar{y}}_c) = \left(\frac{N-n}{N}\right)\frac{1}{nm}\mathrm{MSB}$$

cluster sampling: relative efficiency (with $m_i = m$ for all i)

- the variance with $m_i = m$ for all i is

$$\hat{V}(\bar{\bar{y}}_c) = \left(\frac{N-n}{N}\right)\frac{1}{nm}\mathrm{MSB}$$

- affecting the design effect is the intraclass correlation

$$\rho = \left(\frac{N-n}{n}\right)\left(\frac{1}{n-1}\right)\frac{n\mathrm{MSB}-\mathrm{SST}}{\mathrm{SST}} \approx \frac{\mathrm{MSB}-\mathrm{MST}}{(n-1)\mathrm{MST}}$$

where $\mathrm{MST} = \mathrm{SST}/(N-1)$ and $-1/(n-1) \le \rho \le 1$

cluster sampling: relative efficiency (with $m_i = m$ for all i)

- a simple random sample of nm elements would have given, for the simple random sample estimator $\bar{y}_{\mathrm{SRS}}$

$$\hat{V}(\bar{y}_{\mathrm{SRS}}) = \left(\frac{Nm - nm}{Nm}\right)\frac{s^2}{nm} = \left(\frac{N-n}{N}\right)\frac{s^2}{nm}$$

- having only a cluster sample, s^2 can be approximated by

$$\hat{s}^2 = \frac{N(m-1)\mathrm{MSW} + (N-1)\mathrm{MSB}}{Nm - 1}$$

- the estimated relative efficiency is

$$\widehat{\mathrm{RE}}(\bar{\bar{y}}_{\mathrm{c}}/\bar{y}_{\mathrm{SRS}}) = \frac{\hat{V}(\bar{y}_{\mathrm{SRS}})}{\hat{V}(\bar{\bar{y}}_{\mathrm{c}})} = \frac{\hat{s}^2}{\mathrm{MSB}}$$

cluster sampling: relative efficiency (with $m_i = m$ for all i)

- even with $m_i = m$ for all i, the true variance $V(\bar{\bar{y}}_{\mathrm{c}})$ is complicated, but we can characterize the design effect using the intraclass correlation

$$\begin{aligned}
\mathrm{deff}(\bar{\bar{y}}_{\mathrm{c}}) = \frac{V(\bar{\bar{y}}_{\mathrm{c}})}{V(\bar{y}_{\mathrm{SRS}})} &= 1 + \frac{N(m-1)}{N-1}\rho \\
&\approx 1 + (m-1)\rho \\
&\approx 1 + (m-1)\frac{\mathrm{MSB} - \mathrm{MST}}{(n-1)\mathrm{MST}} \\
&\approx 1 + (m-1)\frac{\mathrm{SSB}\left(\frac{N-1}{n-1} - 1\right) - \mathrm{SSW}}{(n-1)\mathrm{SST}}
\end{aligned}$$

- so if $m_i = m$ for all i, then $V(\bar{\bar{y}}_{\mathrm{c}}) < V(\bar{y}_{\mathrm{SRS}})$ if and only if $\rho < 0$: if and only if there is sufficiently large within-cluster variation

cluster sampling: relative efficiency (with $m_i = m$ for all i)

- Example 8.5: estimate average number of newspapers purchased per household in a given community; there are $M = 4000$ households grouped into $N = 400$ clusters of size $m = m_i = 10$ each; sample size is $n = 4$; using data supplied, estimate average and give a bound on the error of estimation
- see file `example8.5.R`

cluster sampling: estimators, proportion

- a_i: number of elements in cluster i that have the attribute of interest
- estimator for the population proportion p

$$\hat{p} = \sum_{i=1}^{n} a_i \Bigg/ \sum_{i=1}^{n} m_i$$

- estimated variance of $\hat{p}$

$$\hat{V}(\hat{p}) = \frac{1}{\bar{M}^2} \left(\frac{N-n}{N} \right) \frac{s_{\mathrm{p}}^2}{n}$$

where

$$s_{\mathrm{p}}^2 = \frac{\sum_{i=1}^{n} (a_i - \hat{p} m_i)^2}{n-1}$$

cluster sampling: estimators, proportion

- estimator for the population proportion p

$$\hat{p} = \sum_{i=1}^{n} a_i \Big/ \sum_{i=1}^{n} m_i$$

- estimated variance of $\hat{p}$

$$\hat{V}(\hat{p}) = \frac{1}{\bar{M}^2} \left(\frac{N-n}{N} \right) \frac{s_{\mathrm{p}}^2}{n}$$

- the estimator $\hat{V}(\hat{p})$ is good only when the sample is large ($n \geq 20$)
- if $m = m_i$ for all $i = 1, \ldots, N$ (all clusters are the same size), then $\hat{p}$ and $\hat{V}(\hat{p})$ are unbiased estimators

cluster sampling

- Example 8.9: as part of the data collection done to find out about newspaper consumption, residents were asked whether they rent or own their homes; using data supplied, estimate the proportion of residents who rent and give a bound on the error of estimation
- see file `example8.9.R`

cluster sampling within strata

- this means taking a cluster sample within each of L strata
- presuming the number of elements in each stratum is unknown, use the combined ratio estimator instead of the separate ratio estimator to form the estimate of the population mean across strata
- recall the combined ratio estimator for μ_y (element notation):

$$\hat{\mu}_{y\text{RC}} = \left(\frac{\bar{y}_{\text{st}}}{\bar{x}_{\text{st}}} \right) \mu_x \,, \quad \bar{y}_{\text{st}} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_i$$

$$\hat{V}(\hat{\mu}_{y\text{RC}}) = \sum_{i=1}^{L} \left(\frac{N_i}{N} \right)^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_{ri}^2}{n_i}$$

$$s_{ri}^2 = \frac{\sum_{k=1}^{n_i} (y_k - r_C x_k)^2}{n_k - 1} \,, \quad r_C = \frac{\bar{y}_{\text{st}}}{\bar{x}_{\text{st}}}$$

cluster sampling within strata

- applying the combined ratio estimator to clusters

$$\bar{y}_{\mathrm{c}} = \left(\frac{\bar{y}_{\mathrm{st}}}{\bar{m}_{\mathrm{st}}} \right)$$

where, with y_{ik} denoting the value for cluster k in stratum i,

$$\bar{y}_{\mathrm{st}} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_{\mathrm{t}i} \,, \quad \bar{y}_{\mathrm{t}i} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$$

$$\bar{m}_{\mathrm{st}} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{m}_i \,, \quad \bar{m}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} m_{ik}$$

and

$$\hat{V}(\bar{y}_{\mathrm{c}}) = \frac{1}{M^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_{ci}^2}{n_i} \,, \quad s_{ci}^2 = \frac{\sum_{k=1}^{n_i} (y_{ik} - \bar{y}_{\mathrm{c}} m_{ik})^2}{n_i - 1}$$

cluster sampling within strata

- applying the combined ratio estimator to clusters

$$\bar{y}_{\mathrm{c}} = \left(\frac{\bar{y}_{\mathrm{st}}}{\bar{m}_{\mathrm{st}}} \right)$$

and

$$\hat{V}(\bar{y}_{\mathrm{c}}) = \frac{1}{M^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \frac{s_{ci}^2}{n_i}, \quad s_{ci}^2 = \frac{\sum_{k=1}^{n_i} (y_{ik} - \bar{y}_{\mathrm{c}} m_{ik})^2}{n_i - 1}$$

- if M is unknown, in its place use $\hat{M} = \sum_{i=1}^{L} N_i \bar{m}_i$ in $\hat{V}(\bar{y}_{\mathrm{c}})$

cluster sampling within strata

- Example 8.11: estimate per-capita income in two cities combined; sample from first city with $N = 415$ clusters, $n = 25$; sample from first city with $N = 168$ clusters, $n = 10$; given data, estimate per-capita income and bound the error of estimation
- see file `example8.11.R`

cluster sampling with pps

- estimates better than SRS can often be produced by doing sampling with replacement while defining the selection probabilities to be proportional to a measure of the size of y, the variable of main interest
- a smaller sampling variance occurs when larger clusters have a higher probability of being included in the sample
- such sampling is known as sampling with probabilities proportional to a measure of size, or pps sampling

cluster sampling with pps: estimators

- pps sampling is based on the generally unbiased estimator for the total given sampling with replacement:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i}$$

where δ_i is the probability of selecting unit i

  - simple random sampling with replacement has $\delta = 1/N$
  - in pps sampling δ_i is proportional to y_i; in practice δ_i is at best approximately proportional to y_i
  - what happens when δ_i is exactly proportional to y_i?

cluster sampling with pps: estimators

- what happens when δ_i is exactly proportional to y_i?
- recall: if $\delta_i = y_i/\tau$

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\delta_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{y_i/\tau} = \tau$$

  so that $\hat{\tau}$ estimates τ exactly
- the intuition is that the closer selection probabilities are to a measure of size, the better the unbiased estimate will be

cluster sampling with pps: estimators

- use the cluster sizes to determine the selection probabilities

$$\delta_i = \frac{m_i}{M}$$

- the estimator of the total is

$$\hat{\tau}_{\text{pps}} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\delta_i} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{m_i/M}$$
$$= \frac{M}{n}\sum_{i=1}^{n}\frac{y_i}{m_i} = \frac{M}{n}\sum_{i=1}^{n}\bar{y}_i$$

- the estimator for the mean is

$$\hat{\mu}_{\text{pps}} = \frac{1}{M}\hat{\tau}_{\text{pps}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

cluster sampling with pps

- estimator for μ and estimated variance of $\hat{\mu}_{\text{pps}}$

$$\hat{\mu}_{\text{pps}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

$$\hat{V}(\hat{\mu}_{\text{pps}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{\text{pps}})^2$$

- estimator for τ and estimated variance of $\hat{\tau}_{\text{pps}}$

$$\hat{\tau}_{\text{pps}} = \frac{M}{n}\sum_{i=1}^{n} \bar{y}_i$$

$$\hat{V}(\hat{\tau}_{\text{pps}}) = \frac{M^2}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{\text{pps}})^2$$

cluster sampling with pps

- Example 8.12: estimate average number of days of sick leave in firm that has eight divisions; select $n = 3$ divisions using pps given the following division sizes: 1200, 450, 2100, 860, 2840, 1910, 390, 3200
- solution in **R**: `sample(m, n, replace=TRUE, prob=m/M)`
- see file `example8.12.R`

cluster sampling with pps

- Example 8.12: estimate average number of days of sick leave in firm that has eight divisions; use data to estimate the average
- see file `example8.13.R`
- see also file `example8.1213.R`

cluster sampling: which estimator for the total?

- how to choose among the ratio estimator ($M\bar{y}$), the unbiased estimator ($N\bar{y}_{\mathrm{t}}$) and the pps estimator ($\hat{\tau}_{\mathrm{pps}}$)?
  - use the unbiased estimator if y_i is uncorrelated with m_i
  - if y_i and m_i are positively correlated, either the ratio estimator or the pps estimator gives better estimates than the unbiased estimator does
  - if the within-cluster variation does not change as m_i changes, then the pps estimator is better than the ratio estimator
  - if the within-cluster variation increases as m_i increases, then the ratio estimator is better than the pps estimator

two-stage cluster sampling

- two-stage cluster sample: a probability sample of clusters with a probability sample of elements from each selected cluster
- examples
  - Gallup: $n = 300$ election districts with $m_i \approx 5$ households in each district
  - many: U.S. Census blocks are clusters and households are elements
  - telephone surveys: exchanges (or sets of exchanges) are blocks and complete phone numbers are elements
  - industrial processes: "cartons" are clusters and "packages" are elements

two-stage cluster sampling

- more examples
  - blocks of time (say hours) are clusters and smaller units of time (say two-minute intervals) are clusters
  - locations (say street corners) are clusters and activities, incidents, cars or pedestrians are elements
  - places (say land areas) are clusters and environmental domains (say soil samples) are elements
- multistage example
  - American National Election Study: MSAs are clusters, then census blocks are clusters within MSAs, then households are clusters within blocks, then individual adults are elements (often, in effect, households are treated as elements)

two-stage cluster sampling

- the 2012 Wisconsin recall effort
  - see `http://elections.wi.gov/node/2100` especially `Certificate of Sufficiency-Governor.pdf` and `Committee to Recall Walker Petition Receipt.pdf` (which are posted on the Canvass site in the `WI petitions` folder); also see `http://elections.wi.gov/elections-voting/recall/petitions`
  - see an example of a set ("bundle") of petitions in `GOV 251-300.pdf` (in the `Examples` folder)
- two-stage: "bundles" (pdf files) are clusters and signature sheets are elements
- multistage: "bundles" (pdf files) are first-stage clusters and signature sheets are second-stage clusters, and elements are the spaces where a signature can go

two-stage cluster sampling

- cluster sampling versus stratified random sampling (again)
  - cluster sampling works better when sets of elements are similar to one another: data will be collected only from some sets
  - stratified random sampling works better when sets of elements differ from one another: data will be collected from all sets
- example: consider the petition signatures case—strata or clusters?
  - task: estimate the total number of signatures
  - there are 3,045 pdf file "bundles" of 50 sheets each
- usually the decision to use cluster sampling is driven by costs

two-stage cluster sampling

- conceptually simplest case: simple random sample of clusters followed by simple random samples of elements from the selected clusters
  - sampling at both levels is done without replacement
  - to begin, one needs a sampling frame that lists all the clusters
  - sampling frames of the elements in the selected clusters can be constructed after the cluster sample is selected, and then only for the selected clusters

two-stage cluster sampling

- slightly more complicated: pps with replacement cluster sampling followed by simple random samples of elements
  - sampling frames of elements can again be constructed only for the selected clusters
- slightly more complicated:
  - SRS without replacement cluster sample followed by stratified random samples of elements
  - pps with replacement cluster sample followed by stratified random samples of elements
  - when would such designs be useful?

two-stage cluster sampling

- realistic case: pps WITHOUT replacement cluster sampling followed by simple random samples of elements
  - pps without replacement is beyond the scope of this course, though often used in practice
  - the **R** package `survey` can do analysis that takes these designs into account (for example, `svydesign()` and `svymean()`, see `?election` for examples)
  - the **R** package `sampling` can be used to select such samples (for example, `UPsampford()` and `UPsampfordpi2()`)

two-stage cluster sampling: notation

- N: the number of clusters in the population
- n: number of clusters selected by simple random sampling
- M_i: number of elements in cluster i, $i = 1, \ldots, N$
- m_i: number of elements in the simple random sample selected from cluster i
- $M = \sum_{i=1}^{N} M_i$: number of elements in the population
- $\bar{M} = M/N$: average cluster size in the population
- y_{ij}: observation j in the sample from cluster i
- $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i$: sample mean for cluster i

two-stage cluster sampling: motivating estimators

- with a single-stage cluster sample, unbiased estimators for the total τ and for the mean μ are (using cluster totals y_i)

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

$$\hat{\mu} = \frac{N}{Mn} \sum_{i=1}^{n} y_i$$

- with a two-stage cluster sample, y_i is unknown
- but $M_i \bar{y}_i$ is an unbiased estimator for y_i, so substituting into the above formulas gives unbiased, feasible estimators for τ and μ
  - why does such substitution preserve unbiasedness here?

two-stage cluster sampling

- unbiased estimator of the mean μ:

$$\hat{\mu} = \left(\frac{N}{Mn}\right) \sum_{i=1}^{n} M_i \bar{y}_i$$

- estimated variance of $\hat{\mu}$

$$\hat{V}(\hat{\mu}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right) s_{\mathrm{b}}^2 + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 \left(\frac{M_i - m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$$

where

$$s_{\mathrm{b}}^2 = \frac{\sum_{i=1}^{n} (M_i \bar{y}_i - \bar{M}\hat{\mu})^2}{n-1}$$

and s_i^2 is the sample variance for the sample from cluster i

two-stage cluster sampling

- Example 9.1: $N = 90$ plants have, in all, $M = 4500$ sewing machines; with a sample of $n = 10$ plants and roughly $m_i \approx 20\%$ of the machines in each selected plant, estimated the average number of hours of downtime per machine; use the given data to compute estimate and bound
- see file `example9.1.R`

two-stage cluster sampling

- unbiased estimator of the total τ:

$$\hat{\tau} = M\hat{\mu} = \left(\frac{N}{n}\right)\sum_{i=1}^{n} M_i \bar{y}_i$$

- estimated variance of $\hat{\mu}$

$$\begin{aligned}\hat{V}(\hat{\tau}) &= M^2 \hat{V}(\hat{\mu}) \\ &= \left(\frac{N-n}{N}\right)\left(\frac{N^2}{n}\right) s_{\mathrm{b}}^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2 \left(\frac{M_i - m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)\end{aligned}$$

two-stage cluster sampling

- Example 9.2: $N = 90$ plants have, in all, $M = 4500$ sewing machines; with a sample of $n = 10$ plants and roughly $m_i \approx 20\%$ of the machines in each selected plant, estimated the total number of hours of downtime per machine; use the given data to compute estimate and bound
- see file `example9.2.R`

two-stage cluster sampling

- if M is unknown, then estimating it from the sample estimator leads to a ratio estimator for the mean μ

$$\hat{\mu}_{\mathrm{r}} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

- estimated variance of $\hat{\mu}_{\mathrm{r}}$

$$\hat{V}(\hat{\mu}_{\mathrm{r}}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\hat{\bar{M}}^2}\right)s_{\mathrm{r}}^2 + \frac{1}{nN\hat{\bar{M}}^2}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$$

where $\hat{\bar{M}} = \sum_{i=1}^{n} M_i/n$ estimates $\bar{M}$, and

$$s_{\mathrm{r}}^2 = \frac{\sum_{i=1}^{n} M_i^2(\bar{y}_i - \hat{\mu}_{\mathrm{r}})^2}{n-1}$$

two-stage cluster sampling

- Example 9.3: $N = 90$ plants have an unknown number of sewing machines in all; with a sample of $n = 10$ plants and roughly $m_i \approx 20\%$ of the machines in each selected plant, estimated the average number of hours of downtime per machine; use the given data to compute estimate and bound
- see file `example9.3.R`

two-stage cluster sampling

- estimator for a population proportion p

$$\hat{p} = \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}$$

- estimated variance of $\hat{p}$

$$\hat{V}(\hat{p}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n\hat{\bar{M}}^2}\right)s_{\mathrm{r}}^2 + \frac{1}{nN\hat{\bar{M}}^2}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\left(\frac{\hat{p}_i(1-\hat{p}_i)}{m_i - 1}\right)$$

where

$$s_{\mathrm{r}}^2 = \frac{\sum_{i=1}^{n} M_i^2(\hat{p}_i - \hat{p})^2}{n-1}$$

two-stage cluster sampling

- Example 9.4: $N = 90$ plants have sewing machines (same as Example 9.1); with a sample of $n = 10$ plants, estimate the proportion of machines that have been shut down for repairs; use the given data to compute estimate and bound
- see file `example9.4.R`

two-stage cluster sampling: equal-sized clusters

- equal-size clusters: suppose $M_1 = M_2 = \cdots = M_N = \bar{M}$
- equal-size samples: further suppose $m_1 = m_2 = \cdots = m_n = m$
- in this special case

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

  where y_{ij} measures element j in cluster i
- $\hat{\mu}$ is the simple sample average

two-stage cluster sampling: equal-sized clusters

- let $M_i = \bar{M}$ and $m_i = m$ for all i
- the variance of $\hat{\mu}$ can be written in a special form

$$\hat{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{\text{MSB}}{nm} + \left(1 - \frac{m}{\bar{M}}\right) \left(\frac{1}{N}\right) \frac{\text{MSW}}{m}$$

where

$$\text{MSB} = \frac{m}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \hat{\mu})^2$$

$$\text{MSW} = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} s_i^2$$

two-stage cluster sampling: equal-sized clusters

- let $M_i = \bar{M}$ and $m_i = m$ for all i
- large N implies that $\hat{V}(\hat{\mu}) \approx \dfrac{\mathrm{MSB}}{nm}$ depends only on the cluster means, $\bar{y}_i$
  - so to estimate $V(\hat{\mu})$ well, good estimates of s_i^2 are not needed
- $n = N$ implies $\hat{V}(\hat{\mu}) = \left(1 - \dfrac{m}{\bar{M}}\right) \dfrac{\mathrm{MSW}}{nm}$
  - same as in stratified random sampling with $n = N = L$ strata and $m = n_i$ observations in each stratum

two-stage cluster sampling: equal-sized clusters

- let $M_i = \bar{M}$ and $m_i = m$ for all i
- the true variance of $\hat{\mu}$ may be written

$$V(\hat{\mu}) = \frac{1}{n}\left(\sigma_{\mathrm{b}}^2 + \frac{\sigma_{\mathrm{w}}^2}{m}\right)$$

where

$$\sigma_{\mathrm{b}}^2 = \text{variance between cluster means}$$
$$\sigma_{\mathrm{w}}^2 = \text{variance among elements in each cluster}$$

- MSB/m estimates $\sigma_{\mathrm{b}}^2 + (\sigma_{\mathrm{w}}^2/m)$, and MSW estimates σ_{w}^2, so

$$\frac{1}{m}(\mathrm{MSB} - \mathrm{MSW}) \text{ estimates } \sigma_{\mathrm{b}}^2$$

two-stage cluster sampling: equal-sized clusters

- let $M_i = \bar{M}$ and $m_i = m$ for all i
- find sample sizes m and n that either minimize $V(\hat{\mu})$ for a fixed cost or minimize the total cost for a fixed $V(\hat{\mu})$
  - let c_1 be the cost for each cluster
  - let c_2 be the cost for each element within a cluster
  - the total cost is $c = nc_1 + nmc_2$
- given σ^2_{b}, σ^2_{w}, c_1 and c_2, the sample size m that achieves the desired minimizations is

$$m = \sqrt{\frac{\sigma^2_{\text{w}}/c_2}{\sigma^2_{\text{b}}/c_1}}$$

- given m, use $B = 2\sqrt{V(\hat{\mu})} = 2\sqrt{\left(\sigma^2_{\text{b}} + \sigma^2_{\text{w}}/m\right)/n}$ to find n if $V(\hat{\mu})$ is fixed
- given m, use $c = nc_1 + nmc_2$ to find n if c is fixed

two-stage cluster sampling: equal-sized clusters

- Example 9.5: create a sampling plan to sample n batteries a day with a sample of m plate thickness measurements per battery so that the variance of the estimate is 0.1 thousands of an inch; costs per battery are six times the cost per plate in each battery
- see file `example9.5.R`
- real data example, the signature problem: create a sampling plan to estimate the number of signatures on the petitions submitted to advance the 2012 governor's recall in Wisconsin, using a bound of $B = 50,000$ with 95% confidence; there are $N = 3045$ clusters ("bundles," or pdf files) of signature sheets, with $m_i = 50$ logical sheets per bundle
- see file `GOV251-300.R`

two-stage cluster sampling: with pps cluster sample

- if the number of elements varies greatly between clusters, then pps sampling of clusters may produce more efficient samples
  - pps sampling is sampling with probabilities proportional to a measure of size
- pps in the first stage is still being combined with srswor at the second stage
  - srswor is simple random sampling without replacement

two-stage cluster sampling: with pps cluster sample

- with one-stage pps cluster sampling, we considered

$$\hat{\mu}_{\text{pps}} = \frac{1}{M}\hat{\tau}_{\text{pps}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

  based on the general formula

$$\hat{\mu} = \frac{1}{M}\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\delta_i}$$

  in the special case of selection probability $\delta_i = m_i/M$
- the measure of size in this case is exactly the size in terms of numbers of elements, m_i
- estimators for μ and τ are the same as with one-stage cluster sampling, except estimates of cluster means are used in place of exact cluster means

two-stage cluster sampling: with pps cluster sample

- using $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i$
- estimator for μ and estimated variance of $\hat{\mu}_{\text{pps}}$

$$\hat{\mu}_{\text{pps}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

$$\hat{V}(\hat{\mu}_{\text{pps}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{\text{pps}})^2$$

- estimator for τ and estimated variance of $\hat{\tau}_{\text{pps}}$

$$\hat{\tau}_{\text{pps}} = \frac{M}{n}\sum_{i=1}^{n} \bar{y}_i$$

$$\hat{V}(\hat{\tau}_{\text{pps}}) = \frac{M^2}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{\text{pps}})^2$$

two-stage cluster sampling: with pps cluster sample

- Example 9.6: from $N = 6$ hospitals, sample $n = 3$ hospitals in order to estimate the proportion of current patients who are in the hospital for more than two consecutive days; 10% of the current patient records will be examined in the sampled hospitals to determine length of stays; given supplied information about hospital sizes, select a sample with probabilities proportional to size
- see file `example9.6.R`
- Example 9.7: using the supplied data from a sample of $n = 3$ hospitals, estimate the proportion of patients staying more than two days for all six hospitals and place a bound on the error of estimation
- see file `example9.7.R`

two-stage cluster sampling: with pps cluster sample

- signature problem: let pdf ("bundle") file size be the size measure

$$\delta_i = \frac{\text{size}_i}{\sum_{i=1}^{N} \text{size}_i}$$

- unbiased estimators

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \bar{y}_i}{\delta_i}$$

$$\hat{V}(\hat{\tau}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{\delta_i} - \hat{\tau} \right)^2$$

- see `GOVpps.R`

two-stage cluster sampling: with pps cluster sample

- ratio estimation in a special case: the sample has been designed so that each element has the same probability of inclusion (see `ppsratio.pdf` for more information)
- estimator for $R = \tau_y/\tau_x$

$$\hat{r}_{\text{pps}} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$$

where, using δ_i for the cluster selection probability,

$$\hat{\tau}_y = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{\delta_i}$$

$$\hat{\tau}_x = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{x}_i}{\delta_i}$$

two-stage cluster sampling: with pps cluster sample

- ratio estimation in the special case where each element has the same probability of inclusion
- estimated variance of $\hat{r}_{\text{pps}}$

$$\hat{V}(\hat{r}_{\text{pps}}) \approx \frac{1}{\hat{\tau}_x^2}\left[\hat{V}(\hat{\tau}_y) + \hat{r}_{\text{pps}}^2 \hat{V}(\hat{\tau}_x) - 2\hat{r}_{\text{pps}} \operatorname{cov}(\hat{\tau}_y\,,\,\hat{\tau}_x)\right]$$

where

$$\hat{V}(\hat{\tau}_y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{\delta_i} - \hat{\tau}_y\right)^2$$

$$\hat{V}(\hat{\tau}_x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \bar{x}_i}{\delta_i} - \hat{\tau}_x\right)^2$$

$$\operatorname{cov}(\hat{\tau}_y\,,\,\hat{\tau}_x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{\delta_i} - \hat{\tau}_y\right)\left(\frac{M_i \bar{x}_i}{\delta_i} - \hat{\tau}_x\right)$$