Statistics 480: Survey Sampling Techniques

Walter R. Mebane, Jr.

University of Michigan

GSIs: Adam Hall and Fabricio Vasselai
B844 East Hall, 19 March 2018

Outline

systematic sampling

- motivation for 1-in-k systematic sample
  - (example) task is to choose n names from a long list of names
  - N is the number of names in the list (the elements)
  - method: choose $k = N/n$ and pick a random starting point from among the first k names, then select every kth name
- more formal definition: 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 $\{1, 2, \ldots, k\}$
- complications
  - what if n does not evenly divide N?
  - what if N is unknown?

systematic sampling

- reasons to use systematic sampling
  - easy to implement
  - can be used even when "a good frame is not available" (includes the case of not knowing N)
  - can be more efficient than srs (only when intraclass correlation ρ is negative; systematic sampling is a type of cluster sampling)
- consider situation from text: $N = 1000$ travel vouchers, $n = 200$, $k = 5$ (what if last 500 have errors?)
  - systematic sample better than srs?
- another situation: a sample of $n = 50$ shoppers is desired; use $k = 20$ and sample until $n = 50$ is obtained, then stop
  - pitfalls?

systematic sampling

- reasons to use systematic sampling: considerations in different types of populations and given different types of "lists" (see file `poptypes.R`)
  - "random" populations
  - "ordered" populations
  - "periodic" populations
- a systematic sample from a "random" population can be "effectively a simple random sample"
- a systematic sample from an "ordered" population usually gives more precise estimates than from using simple random sampling
- a systematic sample from a "periodic" population must take care not to align with the periodicity in the data (why?)

systematic sampling: mean estimators

- two estimators for μ, using element values y_i
  - optimistic

$$\hat{\mu} = \bar{y}_{\mathrm{sy}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

    "assuming a randomly ordered population"
  - realistic

$$\hat{\mu} = \frac{1}{n_{\mathrm{s}}}\sum_{i=1}^{n_{\mathrm{s}}} \bar{y}_i$$

    using n_{s} repeated systematic samples of size $k' = kn_{\mathrm{s}}$, $k = N/n$, where $\bar{y}_i$ is the mean of the ith systematic sample

systematic sampling: mean estimators

- estimated variance (optimistic, "assuming a randomly ordered population")

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

- true variances

$$V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \qquad \text{(simple random sampling)}$$

$$V(\bar{y}_{\mathrm{sy}}) = \frac{\sigma^2}{n}[1+(n-1)\rho] \qquad \text{(systematic sampling)}$$

where the intraclass correlation is

$$-\frac{1}{n-1} \leq \rho = \frac{(k-1)n\mathrm{MSB} - \mathrm{SST}}{(n-1)\mathrm{SST}} = \frac{\mathrm{SSB}\left(\frac{nk-1}{n-1} - 1\right) - \mathrm{SSW}}{(n-1)\mathrm{SST}} \leq 1$$

- no unbiased estimate of $V(\bar{y}_{\mathrm{sy}})$ is possible using data from one systematic sample

systematic sampling

- Example 7.1: systematic sample of size $n = 20$ from population of $N = 140$ SIC industry groups; using given data, estimate mean number of employees in 2001 and mean loss in employees between 2000 and 2001
- see file `example7.1.R`

systematic sampling: multiple random starting points

- no unbiased estimate of $V(\bar{y}_{\mathrm{sy}})$ is possible using data from one systematic sample
- one approach is to "change the random starting point several times"

systematic sampling: multiple random starting points

- to estimate $V(\bar{y}_{\mathrm{sy}})$, one approach is to "change the random starting point several times"
  - with population size N and sample size n, $k = N/n$
  - divide the population into T subsets of sizes N_j
  - pick a random starting point s_1 in $\{1, 2, \ldots, k\}$, and select every element i satisfying $(i - s_1)/k = 0$ (every kth element) from population subset 1 so long as $i \leq N_1$
  - select a new starting point s_2 in $\{1, 2, \ldots, k\}$ and then sample i satisfying $(i - (N_1 + s_2))/k = 0$ so long as $i \leq N_2$
  - repeat for each population subset using new starting points and successively the rule $(i - (s_h + \sum_{j=1}^{h-1} N_t))/k = 0$ while $i \leq N_h$, $h \in \{3, \ldots, T\}$ to select elements
  - use srs results to approximate $V(\bar{y}_{\mathrm{sy}})$

repeated systematic sampling

- no unbiased estimate of $V(\bar{y}_{\mathrm{sy}})$ is possible using data from one systematic sample
- another approach is to use repeated systematic sampling
- use n_{s} systematic samples and use the square of the deviations of the n_{s} estimates of the mean around the overall mean to estimate $V(\bar{y}_{\mathrm{sy}})$
  - given N, choose n and n_{s}
  - compute $k = N/n$ and $k' = n_{\mathrm{s}}k$
  - choose n_{s} random starts in $\{1, \ldots, k'\}$
  - add $\{0, k', 2k', \ldots, (n/n_{\mathrm{s}} - 1)k\}$ to the random starts to index the selected elements
- see file `example7.repeated.R`

repeated systematic sampling

- estimator of μ using n_s repeated systematic samples:

$$\hat{\mu} = \sum_{i=1}^{n_\mathrm{s}} \frac{\bar{y}_i}{n_\mathrm{s}}$$

  where $\bar{y}_i$ is the mean from systematic sample i
  - $\hat{\mu}$ here is effectively $\bar{y}_\mathrm{t}$ of equation (8.6) in the textbook
- estimated variance of $\hat{\mu}$:

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

  where

$$s_{\bar{y}}^2 = \frac{1}{n_\mathrm{s}-1}\sum_{i=1}^{n_\mathrm{s}}(\bar{y}_i - \hat{\mu})^2$$

repeated systematic sampling

- Example 7.6: state park, $N \approx 400$, $n = 80$, $n_{\mathrm{s}} = 10$; using given data, estimate average number of people per car and show bound
- see file `example7.6.R`

systematic sampling difference estimator

- let $y_1, y_2, \ldots, y_n$ be a random sample with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$
  - usual estimators for σ^2 are based on $\sum_{i=1}^{n}(y_i - \bar{y})^2$, but if $\mu = 0$ were known $\sum_{i=1}^{n} y_i^2/n$ would be an unbiased estimator for σ^2
- suppose, and in general, $\mu \neq 0$
- difference estimator of variance: for $d_i = y_i - y_j$ for all $i \neq j$, $E(d_i) = 0$ and $V(d_i) = 2\sigma^2$
- using n_d such differences, $\sum_{i=1}^{n_\mathrm{d}} d_i^2/n_\mathrm{d}$ is an estimator of $2\sigma^2$
- for a sample of n from population of size N, an estimator of the variance of $\bar{y}_\mathrm{sy}$ is

$$\hat{V}(\bar{y}_\mathrm{sy}) = \left(\frac{N-n}{Nn}\right)\frac{1}{2n_\mathrm{d}}\sum_{i=1}^{n_\mathrm{d}} d_i^2$$

systematic sampling difference estimator

- using the $n_{\mathrm{d}} = n - 1$ successive differences $d_i = y_{i+1} - y_i$, $i = 1, \ldots, n-1$
- difference estimator of variance of $\bar{y}_{\mathrm{sy}}$:

$$\hat{V}(\bar{y}_{\mathrm{sy}}) = \left(\frac{N-n}{Nn}\right)\frac{1}{2n_{\mathrm{d}}}\sum_{i=1}^{n_{\mathrm{d}}} d_i^2$$

- the preceding estimator seems well motivated if the population is effectively random
  - note the covariance complication if "effectively random" meant SRS: recall that SRS implies $\mathrm{cov}(y_i, y_j) = -\frac{1}{N-1}\sigma^2$ for sample values y_i and y_j
- what if not?
  - in particular, what happens if the population is "ordered"?

systematic sampling difference estimator

- difference estimator of variance of $\bar{y}_{\mathrm{sy}}$:

$$\hat{V}(\bar{y}_{\mathrm{sy}}) = \left(\frac{N-n}{Nn}\right)\frac{1}{2n_{\mathrm{d}}}\sum_{i=1}^{n_{\mathrm{d}}} d_i^2$$

- the difference estimator can work well if the population is "ordered"

systematic sampling variance: randomly ordered population

- the true variance of the mean estimator with systematic sampling (if the population is randomly ordered)

$$V(\bar{y}_{\mathrm{sy}}) = \frac{\sigma^2}{n}[1 + (n-1)\rho] \qquad \text{(systematic sampling)}$$

where the intraclass correlation is

$$-\frac{1}{n-1} \leq \rho = \frac{(k-1)n\mathrm{MSB} - \mathrm{SST}}{(n-1)\mathrm{SST}} \leq 1$$

systematic sampling variance: randomly ordered population

- consider the mean $\bar{y}_i$ for each of the k possible systematic samples (clusters) when $N = nk$
- the overall mean per element is $\bar{\bar{y}} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$

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

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

$$\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2$$

$$\rho = \frac{(k-1)n\mathrm{MSB} - \mathrm{SST}}{(n-1)\mathrm{SST}}$$

systematic sampling variance: randomly ordered population

- intraclass correlation

$$\rho = \frac{(k-1)n\mathrm{MSB} - \mathrm{SST}}{(n-1)\mathrm{SST}}$$

- for large N,

$$\rho \approx \frac{\mathrm{MSB} - \mathrm{MST}}{(n-1)\mathrm{MST}}$$

where $\mathrm{MST} = \mathrm{SST}/(nk-1)$

systematic sampling variance: randomly ordered population

- true variance of the mean estimator with systematic sampling (if the population is randomly ordered)

$$V(\bar{y}_{\mathrm{sy}}) = \frac{\sigma^2}{n}[1 + (n-1)\rho]$$

where for large N,

$$\rho \approx \frac{\mathrm{MSB} - \mathrm{MST}}{(n-1)\mathrm{MST}}$$

- true variance of the mean estimator with simple random sampling

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

systematic sampling variance: randomly ordered population

- the variances

$$V(\bar{y}_{\mathrm{sy}}) = \frac{\sigma^2}{n}[1 + (n-1)\rho] \qquad \text{(systematic sampling)}$$

$$V(\bar{y}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right) \qquad \text{(simple random sampling)}$$

  imply the design effect of systematic sampling:

$$\begin{aligned} \mathrm{deff}(\bar{y}_{\mathrm{sy}}) &= \frac{V(\bar{y}_{\mathrm{sy}})}{V(\bar{y})} = \frac{1 + (n-1)\rho}{1 - n/N} \\ &\approx 1 + \frac{\mathrm{MSB} - \mathrm{MST}}{\mathrm{MST}} = \frac{\mathrm{MSB}}{\mathrm{MST}} \quad \text{(for large } N\text{)} \end{aligned}$$

- $\rho < 0$ with an ordered population, so $\mathrm{deff}(\bar{y}_{\mathrm{sy}}) < 1$ is then likely

systematic sampling examples

- examples
  - random data (see files `table7.5` and `work7.5.R`)
  - ordered data (see files `table7.6` and `work7.6.R`)

systematic sampling

- Example 7.8: use data set `river77-78` to find the mean daily flow rates for a Florida river; estimate the average flow rate for October, November and December 1997 using a systematic sample with $k = 10$
- see file `example7.8.R`