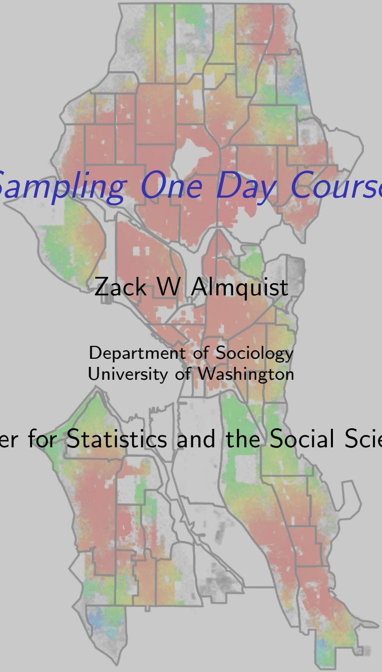


Survey Sampling One Day Course: Part 2

Zack W Almquist

Department of Sociology
University of Washington

Center for Statistics and the Social Sciences

Cluster Sampling

Why Cluster Sampling

Common Examples

- The population may be widely distributed geographically or may occur in natural clusters (e.g., hh or schools).
- It may be much less expensive to take a sample of clusters than a SRS of individuals.
    - Households
    - School districts
    - Schools
    - Classrooms
    - Nursing homes
    - Blocks
    - States
    - Regions
    - Countries

Notation for Cluster Sampling

PSU Level

- y_{ij} = measurement for jth element in ith psu
- N = number of ssus in the population
- M_i = number of ssus in psu i
- $M_0 = \sum_{i=1}^{N} M_i$ = total number of suss in the population
- $t_i = \sum_{j=1}^{M_i} y_{ij}$ = total in psu i
- $t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ = population total
- $S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N}\right)^2$ = population variance of psu totals

Notation for Cluster Sampling

SSU level

- $\bar{y}_U = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0}$ = population mean
- $\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$ = population mean in psu i
- $S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{M_0 - 1}$ = population variance (per ssu)
- $S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1}$ = population variance within psu i

Notation for Cluster Sampling

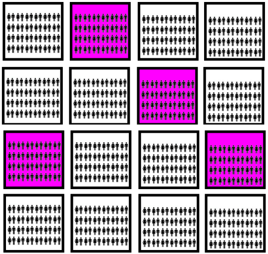
Sample Quantities

- y_{ij} = measurement for jth element in ith psu
- n = number of ssus in the sample
- m_i = number of ssus in the sample from psu i
- $\bar{y}_i = \sum_{j \in S_i} \frac{y_{ij}}{m_i}$ = sample mean (per ssu) for psu i
- $\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij}$ = estimated total in psu i
- $\hat{t}_{unb} = \sum_{i \in S} \frac{N}{n} \hat{t}_i$ = unbiased estimator of the population total
- $s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{\hat{t}_{unb}}{N} \right)^2$ = population variance of psu totals
- $s_i^2 = \sum_{j \in S_i} \frac{(y_{ij} - \bar{y}_i)^2}{m_i - 1}$ = sample variance within psu i
- w_{ij} = sampling weight for ssh j in psu i

One-Stage Cluster Sampling

Clusters of Equal Sizes Estimation

- $M_i = m_i = M$.
- This rarely happens in human systems, but is typical of agricultural or industrial problems.



One-Stage Cluster Sampling

- We have an SRS of n data points $\{t_i, i \in S\}$; t_i is the total for all the elements in psi i.

$$\hat{t}_S = \sum_{i \in S} \frac{t_i}{n}$$

  Estimates the average of the cluster totals.

- To estimate the total income t, we can use the estimator

$$\hat{t} = \frac{N}{n} \sum_{i \in S} t_i$$

- To make this concrete consider estimating the total income of two person households

One-Stage Cluster Sampling

- $V(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}$
- $SE(\hat{t}) = N\sqrt{\left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}}$
- S_t^2 and s_t^2 are the population and sample variance, respectively
- PSU totals:

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(t_i - \frac{t}{N}\right)^2$$

and

$$s_t^2 = \frac{1}{n-1}\sum_{i \in S}\left(t_i - \frac{\hat{t}}{N}\right)^2$$

- $\hat{\bar{y}} = \frac{\hat{t}}{NM}$
- $V(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right) \frac{S_t^2}{nM^2}$
- $SE(\hat{\bar{y}}) = \frac{1}{M}\sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}$

One-Stage Cluster Sampling

- One-stage cluster sampling with an SRS of psus produces a self-weighting sample.
- The weight for each observation unit is:

$$w_{ij} = \frac{1}{\Pr(\text{ssu } j \text{ psu is in sample})} = \frac{N}{n}$$

- $\hat{t} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$
- $\hat{\bar{y}} = \frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}$

Clusters of Equal Sizes: Theory

Clusters of Equal Sizes: Theory

- Goal: Compare cluster sampling to SRS
- Note that cluster sampling almost always provides less precision for the estimators than one would obtain by taking an SRS wight he same number of elements!

Clusters of Equal Sizes: Theory

ANOVA Decomposition

- SST=SSB+SSW
- This corresponds to MST, MSB and MSW
- Variance of estimators
  - Unlike **stratified sampling** where the variances of the estimators depended on the within group variation (MSW), in **cluster sampling** the variances of the estimators depend on the between group variation (MSB).
  - $F = MSB/MSE$
    - If F is large then stratification *decreases* variance relative to an SRS
    - If F is large then clustering *increases* variance relative to an SRS.

Interclass correlation coefficient (ICC)

ICC

- Intraclass correlation coefficient (ICC)
- Intracluster correlation coefficient (ICC)
- Is a measure on how similar elements in the same cluster are.
- It provides a measure of **homogeneity** within the clusters.

Interclass correlation coefficient (ICC)

ICC

- ICC is defined to be the Pearson correlation coefficient for the $NM(M-1)$ pairs (y_{ij}, y_{ik}) for i between 1 and N and $j \neq k$.
- It can be written in terms of the population ANOVA table quantities as:

$$ICC = 1 - \frac{M}{M-1}\frac{SSW}{SST}$$

Interclass correlation coefficient (ICC)

ICC

- B/c $0 \leq SSW/SSTO \leq 1$, it follows that

$$-\frac{1}{M-1} \leq ICC \leq 1$$

- If the clusters are perfectly homogeneous and hence $SSW = 0$ then $ICC = 1$
-

$$MSB = \frac{NM-1}{M((N-1)}S^2[1+(M-1)ICC]$$

- This allows us to say how much precision we lose by taking a cluster sample

$$\frac{V(\hat{t}_{cluster})}{V(\hat{t}_{SRS}} = \frac{MSB}{S^2} = \frac{NM-1}{M(N-1)}[1+(M-1)ICC]$$

Interclass correlation coefficient (ICC)

ICC

- If N, the number of psus is in the population, is large
  - $NM - 1 \approx M(N-1)$ and then the ratio of the variances is approximately $1 + (M-1)ICC$
  - $1 + (M-1)ICC$ suss taken from a one-stage cluster sample, give us approximately the same amount of information as one ssh from an SRS.
  - If $ICC = 0.5$ and $M = 5$ then $1 + (M-1)ICC = 3$, thus we would need 300 elements using cluster sampling to obtain the same precision as an SRS of 100 elements.

Interclass correlation coefficient (ICC)

ICC

- *ICC* is only defined for clusters of equal sizes.
- An alternative measure of homogeneity in general populations is the adjusted R^2, call R_a^2 and defined as:

$$R_a^2 = 1 - \frac{MSW}{S^2}$$

- if all psus are of the same size, then the increase in variance due to cluster sampling is:

$$\frac{V(\hat{t}_{cluster})}{V(\hat{t}_{SRS}} = \frac{MSB}{S^2} = 1 + \frac{N(M-1)}{N-1} R_a^2$$

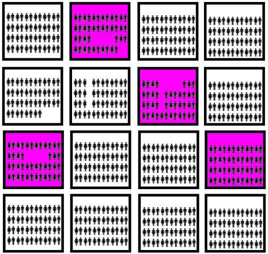
Cluster of Unequal Sizes (One-stage)

Clusters of Unequal Sizes (One-stage)

In a one-stage cluster sample of n of the N psus, we can estimate population totals and means in two ways:

- Unbiased Estimation (SRS Theory and census of the cluster)
- Ratio Estimation

- Unbiased Estimation
  - $\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} t_i$
  - $SE(\hat{t}) = N\sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}$
  - $M_0 = \sum_{i=1}^{N} M_i$
  - $\hat{\bar{y}}_{unb} = \hat{t}_{unb} / M_0$
  - $SE(\hat{\bar{y}}_{unb}) = SE(\hat{t}) / M_0$



Clusters of Unequal Sizes (One-stage)

Sample Weights

- The probability that psu is in the sample is $\frac{n}{N}$ (remember this stage is just SRS).
- B/c this is a one-stage cluster sample, an ssu is included in the sample whenever a psu is included in the sample.
- $w_{ij} = \frac{1}{\Pr(\text{ssu } j \text{ of psu } i \text{ is in the sample})} = \frac{N}{n}$
- Thus One-stage cluster sampling produces a self-weighting sample when psus are selected with equal probabilities.
- $\hat{t}_{unb} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$

Cluster of Unequal Sizes (One-Stage)

Relationship to Ratio Estimation

What do we do if M_0 is not known?

Cluster of Unequal Sizes (One-Stage)

Relationship to Ratio Estimation

One solution is the Ratio Estimator!

Cluster of Unequal Sizes (One-Stage)

Ratio Estimation: Population

$$\bar{y}_U = \frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} M_i} = \frac{t}{M_o}$$

Cluster of Unequal Sizes (One-Stage)

Ratio Estimation: Sample

$$\hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i} = \frac{\sum_{i \in S} M_i \hat{\bar{y}}_i}{\sum_{i \in S} M_i}$$

This can be rewritten using the weights w_{ij}:

$$\hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}}{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}$$

Cluster of Unequal Sizes (One-Stage)

Ratio Estimation: Sample

$$\hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i\in S} t_i}{\sum_{i\in S} M_i} = \frac{\sum_{i\in S} M_i \hat{y}_i}{\sum_{i\in S} M_i}$$

This can be rewritten using the weights w_{ij}:

$$\hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\hat{M}_0} = \frac{\sum_{i\in S}\sum_{j\in S_i} w_{ij} y_{ij}}{\sum_{i\in S}\sum_{j\in S_i} w_{ij}}$$

$$SE(\hat{\bar{y}}_r) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in S}(t_i - \hat{\bar{y}}_r M_i)^2}{n-1}}$$
$$= \sqrt{\left(1 - \frac{n}{N}\right)\frac{1}{n\bar{M}^2}\frac{\sum_{i\in S} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}}$$

Cluster of Unequal Sizes (One-Stage)

Ratio Estimation: Sample

- If we know M_0 then we can use ratio estimation to estimate the population total.
- $\hat{t}_r = M_0\hat{\bar{y}}_r$
- $SE(\hat{t}_r) = M_0 SE(\hat{t}_r)$

Two-Stage Cluster Sampling

Two-Stage Cluster Sampling

Comparison

- In one-stage cluster sampling you sample PSUs and then take a census of all SSUs.
- In two-stage cluster sampling you sample PSUs and then take a sample of SSUs.
- The simplest form is an SRS of PSUs followed by an SRS of SSUs.

Two-Stage Cluster Sampling

Procedure

- Select an SRS S of n psus from the population of N psus.
- Select an SRS of ssus from each selected psu. The SRS of m_i elements from the i psu is denoted S_i and $|S_i| = M_i$.

Two-Stage Cluster Sampling

One-stage cluster sampling $\hat{t}$

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} t_i$$

Two-Stage Cluster Sampling

Two-stage cluster sampling $\hat{t}$

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i$$

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} \hat{t}_i = \sum Nn \sum_{i \in S} M_i \bar{y}_i = \sum_{i \in S} \sum_{j \in S_i} \frac{N}{n} \frac{M_i}{m_i} y_{ij}$$

Two-Stage Cluster Sampling

Two-stage cluster sampling $\hat{t}$

- We can of course rewrite this sum as a weighted sum ...
- $\text{Pr}(\ j\text{th ssu in } i\text{th psu is selected }) =$
  $\text{Pr}(\ i\text{th psu selected}) \times \text{Pr}(\ j\text{th ssu selected — } i\text{th psu selected })$
  $= \frac{n}{N}\frac{m_i}{M_i}$
- Thus $w_{ij} = \frac{NM_i}{nm_i}$ and
- $\hat{t}_{unb} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$

Two-Stage Cluster Sampling

Two-stage cluster sampling $SE(\hat{t})$

$$\hat{V}(\hat{t}_{unb}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$$

- $s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$
- $s_i^2 = \frac{1}{m_i-1} \sum_{j \in S_i} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2$

Two-stage cluster sampling $\hat{\bar{y}}_{unb}$

$$\hat{\bar{y}}_{unb} = \hat{t}_{unb}/M_0$$

$$SE(\hat{\bar{y}}_{unb}) = SE(\hat{t}_{unb})/M_0$$

Two-Stage Cluster Sampling

Two-stage cluster sampling $\hat{\bar{y}}_r$

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i} = \frac{\sum_{i \in S} M_i \bar{y}_i}{\sum_{i \in S} M_i}$$

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i \in S} M_i^2 \left(1 - \frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

where $s_r^2 = \frac{1}{n-1}\sum_{i \in S}(M_i \bar{y}_i - M_i \hat{\bar{y}}_r)^2$

Example: coot

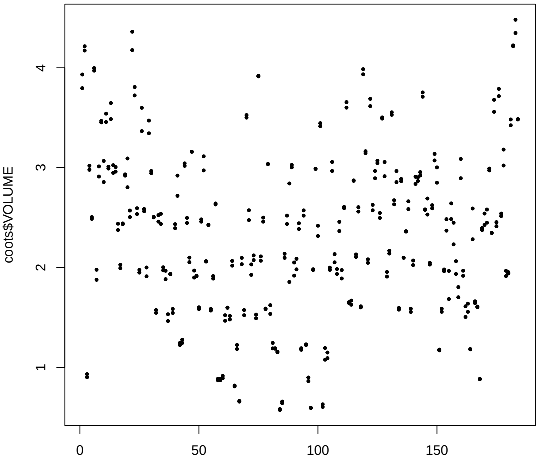
- Arnold's (1991) work on egg size and volume of American Coot eggs in Minnedosa, Manitoba.
- We are looking at the volumes of a subsample of eggs in clutches (nests of eggs) with at least two eggs available for measurement.

```
# devtools::install_github('SSDALab/lohrData')
library(lohrData)
data(coots)
head(coots)
```

```
R >     CLUTCH CSIZE LENGTH BREADTH    VOLUME TMT
R >  1       1    13  44.30   31.10 3.7957569   1
R >  2       1    13  45.90   32.70 3.9328497   1
R >  3       2    13  49.20   34.40 4.2156036   1
R >  4       2    13  48.70   32.70 4.1727621   1
R >  5       3     6  51.05   34.25 0.9317646   0
R >  6       3     6  49.35   34.40 0.9007362   0
```

Example: coot

```r
plot(coots$CLUTCH, coots$VOLUME, pch = 19, cex = 0.5)
```



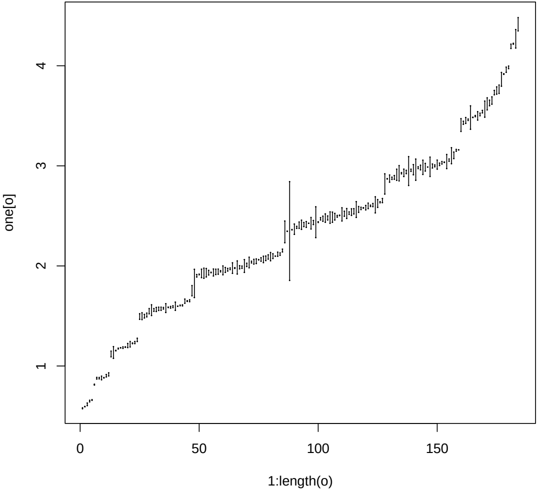
Example: coot

```r
foo <- split(coots$VOLUME, coots$CLUTCH)
# foo[[1]]
ybari <- sapply(foo, mean)

o <- order(ybari)
one <- sapply(foo, "[", 1)
two <- sapply(foo, "[", 2)
plot(1:length(o), one[o], pch = 19, cex = 0.1)
points(1:length(o), two[o], pch = 19, cex = 0.1)

for (i in 1:length(o)) segments(x0 = i, x1 = i, y0 = one[o[i]],
    y1 = two[o[i]], cex = 1)
```

Example: coot



Example: coot

```
foo <- split(coots$VOLUME, coots$CLUTCH)
ybari <- sapply(foo, mean)
si2 <- sapply(foo, var)
Mi <- sapply(split(coots$CSIZE, coots$CLUTCH), mean)
hat.ti <- (Mi/2) * sapply(foo, sum)
ssufpc <- (1 - 2/Mi)
ybar.r <- sum(hat.ti)/sum(Mi)

var1 <- ssufpc * (Mi^2) * (si2/2)
var2 <- (hat.ti - Mi * ybar.r)^2

cluster.table <- data.frame(Clutch = 1:184, Mi, ybari, si2,
    hat.ti, var1, var2)
head(cluster.table)
```

```
 R >    Clutch Mi     ybari          si2    hat.ti
 R >  1      1 13 3.8643033 0.0093972179 50.235943
 R >  2      2 13 4.1941828 0.0009176971 54.524377
 R >  3      3  6 0.9162504 0.0004813808  5.497502
 R >  4      4 11 2.9983346 0.0007950278 32.981681
 R >  5      5 10 2.4957075 0.0001574425 24.957075
 R >  6      6 13 3.9842595 0.0003303709 51.795373
 R >          var1         var2
 R >  1 0.67190108 3.189232e+02
 R >  2 0.06561534 4.904832e+02
 R >  3 0.00577657 8.922633e+01
 R >  4 0.03935387 3.119577e+01
 R >  5 0.00629770 2.630604e-03
 R >  6 0.02362152 3.770530e+02
```

Example: coot

```
ybar.r

 R >  [1] 2.490579

sr2 <- (1/(183)) * sum(var2)
sr2

 R >  [1] 62.51136

### Why no fpc? is this justified?
vhat.ybar.r.nfpc <- (1/(mean(Mi)^2)) * sr2/184
se.ybar.r.nfpc <- sqrt(vhat.ybar.r.nfpc)
se.ybar.r.nfpc

 R >  [1] 0.0610403

CV.hat <- se.ybar.r.nfpc/ybar.r
CV.hat

 R >  [1] 0.02450848
```

Unequal-Probability Sampling

Unequal-Probability Sampling

- Sometimes it is not practical or desirable to sample every cluster or unit with equal probability.
- Theory and estimators in this framework have been developed for **with** and **without** replacement.
- Theory **with** replacement is slightly easier, in the sense that the estimates and sampling procedures are simpler.
- Without replacement is more efficient then with replacement.
- You should read Section 6.0-6.4. Here we will only cover Unequal-probability sampling without replacement.

Probability Proportional to Size (PPS)

Probability Proportional to Size (PPS)

With-replacement Estimation

- $\Pr(\text{unit } i \text{ selected on first draw}) = \psi_i$
- $\Pr(unit i in sample) = \pi_i$
- One-stage Sampling with replacement:
  $\psi = \Pr(\text{select unit } i \text{ on first draw})$
- $\hat{t}_\psi = \frac{1}{n}\sum_{i\in R}\frac{t_i}{\psi_i} = \frac{1}{n}\sum_{i\in R} u_i = \bar{u}$
- $\hat{V}(\hat{t}_\psi) = \frac{s_u^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in R}(u_i - \bar{u})^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i\in R}\left(\frac{t_i}{\psi} - \hat{t}_\psi\right)^2$
- This is known as the Hansen-Hurwitz (1943) Estimator.

Probability Proportional to Size (PPS)

With-replacement Estimation

- We designing selection probabilities, one wants to choose the ψ_i's so that the variance of $\hat{t}_\psi$ is as small as possible.
- Ideally we would choose $\psi_i = t_i/t$ and $\hat{t}_\psi = t$ for all samples and $V(\hat{t}_\psi) = 0$.
- In practice this is not possible, or we are interested in more than a single total from a survey.
- It is common to take ψ to be proportion of the elements in psu i or the relative size of psu i.
- With M_i the number of elements in the ith psu and $M_0 = \sum_{i=1}^{N} M_i$ the number of elements in the population.
- We take $\psi_i = M_i/M_0$.
- This choice of ψ_i is called **probability proportional to size (pps)**

Probability Proportional to Size (PPS)

With-replacement Estimation

- Thus for one-stage ops sampling $t_i/\psi_i = t_i M_o/M_i = M_0 \bar{y}_i$
- $\hat{t}_\psi = \frac{1}{n} \sum_{i \in R} M_0 \bar{y}_i$
- $\hat{\bar{y}}_\psi = \frac{1}{n} \sum_{i \in R} \bar{y}_i$ with $\psi = M_i/M_0$
- $\hat{\bar{y}}_\psi$ is the average of the sampled psu means.
- $\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{n} \frac{1}{n-1} \sum_{i \in R} (\bar{y}_i - \hat{\bar{y}}_\psi)^2$

```r
library(lohrData)
data(statepop)
head(statepop)
```

```
R >   STATE      COUNTY LANDAREA    POPN PHYS FARMPOP
R > 1    AL      Wilcox      889   13672    4     666
R > 2    AZ    Maricopa     9204 2209567 4320    2124
R > 3    AZ    Maricopa     9204 2209567 4320    2124
R > 4    AZ       Pinal     5370  120786   61     881
R > 5    AR     Garland      678   76100  131     524
R > 6    AR Mississippi      898   55060   48     955
R >   NUMFARM FARMACRE VETERANS PERCVIET
R > 1     322   156950      836     20.8
R > 2    2334  1391456   262170     31.5
R > 3    2334  1391456   262170     31.5
R > 4     730  1958489    14858     29.1
R > 5     389    41293    11055     21.3
R > 6     615   488042     5285     33.8
```

```
M0 <- 255077536
totalCounty <- data.frame(state = statepop$STATE, county = statepop$COUNTY,
    popsize = statepop$POPN, psi = statepop$POPN/(M0), numPhys = statepop$PHYS,
    ti_psi = statepop$PHYS/(statepop$POPN/M0))
```

```r
head(totalCounty)
```

```
R >    state      county popsize          psi numPhys
R >  1    AL      Wilcox   13672 5.359939e-05       4
R >  2    AZ    Maricopa 2209567 8.662335e-03    4320
R >  3    AZ    Maricopa 2209567 8.662335e-03    4320
R >  4    AZ       Pinal  120786 4.735266e-04      61
R >  5    AR     Garland   76100 2.983407e-04     131
R >  6    AR Mississippi   55060 2.158559e-04      48
R >       ti_psi
R >  1  74627.72
R >  2 498710.81
R >  3 498710.81
R >  4 128820.64
R >  5 439095.36
R >  6 222370.54
```

```r
### Table Descriptives
sum(totalCounty$ti_psi)
```

```
R >  [1] 57030430
```

```r
sd(totalCounty$ti_psi)/sqrt(100)
```

```
R >  [1] 41401.23
```

```r
## n
nrow(totalCounty)
```

```
R >  [1] 100
```

```r
## Sum of weights
sum(M0/totalCounty$popsize)
```

```
R >  [1] 245072
```

Unequal-Probability Sampling Without Replacement

Unequal-Probability Sampling Without Replacement

The Horvitz-Thompson Estimator for One-stage

- Assume we have a without-replacement sample of n psus, and we know the inclusion probability

$$\pi_i = \mathrm{Pr}(\text{ unit } i \text{ in sample}).$$

- The joint inclusion probability

$$\pi_{ik} = \mathrm{Pr}(\text{ units } i \text{ and } k \text{ are both in the sample}).$$

- The inclusion probability π_i can be calculated as the sum of the probabilities of all sample containing the ith unit and has the property

$$\sum_{i=1}^{N} \pi_i = n.$$

Unequal-Probability Sampling Without Replacement

The Horvitz-Thompson Estimator for One-stage

- For the π_{ik}'s,

$$\sum_{\substack{k=1 \\ k\neq 1}}^{N} \pi_{ik} = (n-1)\pi_i.$$

Unequal-Probability Sampling Without Replacement

The Horvitz-Thompson Estimator for One-stage

- B/c the inclusion probabilities sum to n, we can think of

$$\pi_i/n$$

  as the "average probability" that a unit will be selected on one of the draws.

Unequal-Probability Sampling Without Replacement

The Horvitz-Thompson Estimator for One-stage

- The **Horvitz-Thompson (HT) estimator** of the population total:

$$\hat{t}_{HT} = \sum_{i \in S} \frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i}$$

  where $Z_i = 1$ if psu i is in the sample, and 0 otherwise.
- The HT estimator is unbiased, i.e.,

$$E[\hat{T}_{HT}] = \sum_{i=1}^{N} \pi_i \frac{t_i}{\pi_i} = t.$$

Unequal-Probability Sampling Without Replacement

The Horvitz-Thompson Estimator for One-stage

- The Variance for the HT (One-stage) Cluster Sample is:

$$
\begin{aligned}
V(\hat{t}_{HT}) &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 .
\end{aligned}
$$

- You can see that the variance of the HT estimator is 0 if t_i is proportional to π_i.

Unequal-Probability Sampling Without Replacement

The Horvitz-Thompson Estimator for One-stage

- The Estimated Variance for the HT (One-stage) Cluster Sample is:

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum_{i \in S}(1 - \pi_i)\frac{t_i^2}{\pi_i^2} + \sum_{i \in S}\sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k}.$$

- An alternative estimator proposed by Sen-Yates-Grundy (SYG) for the variance:

$$\hat{V}_{SYG}(\hat{t}_{HT}) = \frac{1}{2}\sum_{i \in S}\sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2$$

Cluster Sampling Unequal-Probability Sampling (Two-Stage)

Horvitz-Thompson for Two Stage

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}$$

Where $Z_i = 1$ if psu i is in the sample, and 0 otherwise.

- The two-stag Horvitz-Thompson estimator is an unbiased estimator of t as long as $E[\hat{t}_i] = t_i$ for each psu i.

Cluster Sampling Unequal-Probability Sampling (Two-Stage)

Horvitz-Thompson for Two Stage

- The variance of the HT Two-Stage estimator:

$$
\begin{aligned}
V(\hat{t}_{HT}) &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i} \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i}
\end{aligned}
$$

Cluster Sampling Unequal-Probability Sampling (Two-Stage)

Horvitz-Thompson for Two Stage

- The estimated variance of the HT Two-Stage estimator:

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum_{i \in S}(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in S}\sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k} + \sum_{i \in S}\frac{V(\hat{t}_i)}{\pi}$$

$$\hat{V}_{SYG}(\hat{t}_{HT}) = \frac{1}{2}\sum_{i \in S}\sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k}\right)^2 + \sum_{i \in S}\frac{V(\hat{t}_i)}{\pi}$$

  - Both estimators are unbiased, however they can be negative in practice.