Statistics 522: Sampling and Survey Techniques

Topic 5

Topic Overview

This topic will cover

- One-stage Cluster Sampling
- Two-stage Cluster Sampling
- Systematic Sampling

Cluster Sampling: Basics

- Consider sampling children in an elementary school.
- We could take an SRS.
- An alternative is to take a random sample of classes and then measure all students in the selected classes.

Terminology

- The classes are the *primary sampling units* (*psu*s) or *clusters*.
- The students within the classes are the *secondary sampling units* (*ssu*s).
- Often the ssus are the elements of the population.

Why use cluster samples?

- Constructing a frame of the observation units may be difficult, expensive, or impossible.
  - Customers of a store
  - Birds in a region
- The population may be widely distributed geographically or may occur in natural clusters.
  - Residents of nursing homes
  - College students in dorms/classes

Comparison with stratification

- With both clusters and stratification we partition the population into subgroups (strata or clusters).
- With stratification, we sample from *each* of the subgroups.
- With cluster sampling, we sample *all* of the units in a *subset* of subgroups.

Precision

- In general, for a given total sample size n,
  - Cluster sampling will produce estimates with the largest variance.
  - SRS will be intermediate.
  - Stratification will give the smallest variance.

Notation

PSU level

- Measurement for jth element in the ith psu is $y_{i,j}$.
- In "design of experiments" we would call this a *nested* design.
- N is the number of psus in the population.
- M_i is the number of ssus in the ith psu.
- K is the number of ssus in the population.
- t_i is the total in the ith psu.
- t is the population total.
- S_t^2 is the population variance of the psu totals (between cluster variation).

SSU level

- $\bar{y}_U$ is the population mean.
- $\bar{y}_{i,U}$ is the population mean in the ith psu.
- S^2 is the population variance (total variation).
- S_i^2 is the population variance within the ith psu.

Sample values

- n is the number of psus in the sample.
- m_i is the number of elements in the sample for the ith psu.
- $\bar{y}_i$ is the sample mean for the ith psu.
- $\hat{t}_i$ is the estimated total for the ith psu.
- $\hat{t}_{unb}$ is the unbiased estimate of t (weighted mean of t's).
- s_t^2 is the estimated variance of psu totals.
- s_i^2 is the sample variance within the ith psu.

Clusters of equal size

- Think about the t_i as the basic observations and use the SRS theory.
- $M_i = M$ for all i.

Estimate of total

$$\begin{aligned}\hat{t} &= \frac{N}{n}\sum t_i \\ \text{Var}(\hat{t}) &= N^2 fpc\frac{S_t^2}{n}\end{aligned}$$

- To get the SE, substitute the sample estimate s_t^2 for S_t^2 and take the square root.
- For 95% the MOE is 1.96 times the SE.

Estimate of mean

- The estimate of $\bar{y}_U$ is $\hat{\bar{y}}$, the estimate of the population total divided by the number of units in the population.
- $\hat{\bar{y}} = \hat{t}/(NM)$
- The SE for this estimate is the SE of $\hat{t}$ divided by NM.
- For 95% the MOE is 1.96 times the SE

Example

- Study Example 5.2 on page 137.
- A dorm has 100 suites, each with four students.
- Select an SRS of 5 suites.
- Ask each student in the selected suites to report their GPA.
- Key is suite-to-suite variation.

Some theory

- Think in terms of an anova decomposition of sums of squares (between and within clusters):

$$SST = SSB + SSW$$

- And the corresponding mean squares: MST, MSB, MSW

Variance of estimators

- For stratified sampling
  - Variances of the estimators depend on the within group variation MSW.
- For cluster sampling
  - 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.
- If $MSB > MST = S^2$ then cluster sampling is less efficient than an SRS.

ICC

- *Intraclass* (or *intracluster*) *correlation coefficient* (ICC) is the common correlation among pairs of observations from the same cluster.

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

- If clusters are perfectly homogeneous, then $ICC = 1$.
- ICC could also be negative.

Design effect

- The design effect is the ratio of the variances for two different designs having the same number of sampled units, usually with the variance of the SRS in the denominator.

- The design effect for cluster sampling relative to simple random sampling is MSB/MST (or MSB/S^2)

$$\frac{NM-1}{M(N-1)}[1-(M-1)ICC].$$

Clusters of unequal size

- No new ideas

- Formulas are messier.

- See text Section 5.2.3 on pages 143-144.

Ratio Estimation

- Use the M_i, the number of *ssu*s in the ith *psu*, as the auxiliary variable (x_i).

- Formulas are in Section 5.2.3.2 on pages 144-145.

Comparison

Need to know K to

- Estimate $\bar{y}$ using unbiased estimation:

$$\begin{aligned}\hat{t}_{unb} &= \frac{N}{n}\sum_{sam} t_i \\ \hat{\bar{y}}_{unb} &= \frac{\hat{t}_{unb}}{K}\end{aligned}$$

- Estimate t using ratio estimation

$$\begin{aligned}\hat{\bar{y}}_r &= \frac{\sum t_i}{\sum M_i} \\ \hat{t}_r &= K\hat{\bar{y}}_r\end{aligned}$$

Two-stage cluster sampling

- If the items within a cluster are very similar, it is wasteful to measure all of them.

- Alternative is to take an SRS of the units in each selected psu (cluster).

First stage

- Population is N psus (clusters).
- Take a SRS of n psus.

Second stage

- M_i is the number of ssus in cluster i.
- For each of the sampled clusters, draw an SRS.
- The sample size for cluster i is m_i.

Estimation of the total

- In one-stage cluster sampling, we use $\hat{t}_{unb} = \frac{N}{n} \sum_{sample} t_i$ as the estimate of the population total.
- Note that the t_i are known without error because we sample all ssus in the sampled psus.
- For two-stage cluster sampling, we need to estimate the t_i.

Estimate of t_i

- Within each cluster, we have an SRS so all that we have learned about estimation with SRSs applies.
- The sample mean for cluster i is

$$\bar{y}_i = \frac{1}{m_i} \sum_{\text{in cluster } i} y_{i,j}$$

- To estimate the total for cluster i we multiply by M_i,

$$\hat{t}_i = M_i \bar{y}_i$$

Estimate of population total

- The estimate of the population total is obtained from the $\hat{t}_i$.
- We first find the average of these (divide by n) and then multiply by the population size (N).

$$\hat{t}_{unb} = \frac{N}{n} \sum_{sample} \hat{t}_i$$

Estimated variance

- The estimated variance for $\hat{t}_{unb}$ is obtained by deriving a formula for the true variance and substituting sample estimates for unknown parameters in this formula.
- The formula contains two terms:
  - A term equal to the expression for one-stage clustering (S_t^2).
  - An additional term to account for the fact that we took an SRS at the second stage (S_i^2's).
- The derivation is given in the text for the general case of unequal probability sampling in Section 6.6.

Between cluster variance

- We estimate the between cluster variance, viewing the $\hat{t}_i$ as an SRS.

$$s_t^2 = \sum_{sample} (\hat{t}_i - \bar{\hat{t}})^2/(n-1)$$

- Note the text uses $\hat{t}_{unb}/N$ for $\bar{\hat{t}}$.
- s_t^2 is an estimate of S_t^2 the true variance of the t_i.

Within cluster variance

- We estimate the within cluster variance, viewing the $y_{i,j}$ as an SRS.
- For cluster i

$$s_i^2 = \frac{1}{m_i - 1} \sum_{sample} (y_{i,j} - \bar{y}_i)^2$$

- There is an *fpc* for each cluster

$$fpc_i = (1 - m_i/M_i)$$

Estimated Variance of $\hat{t}_{unb}$

- First term as in single-stage.

$$N^2 fpc \frac{s_t^2}{n}$$

- Plus the within term

$$\frac{N}{n}\sum_{sample} fpc_i \frac{M_i^2 s_i^2}{m_i}$$

  Take the square root to get the SE

- Multiply the SE by 1.96 for the MOE

- The 95% CI is $\hat{t}_{unb} \pm MOE$

Population mean

- K is the total number of elements in the population (assume this is known).

- The estimate of the population mean is the estimate of the population total divided by K ($\hat{t}/K$).

- The SE for this estimate is the SE for the total divided by K.

Ratio Estimate

- We use the same procedure that we used for one-stage clustering.

- M_i is the auxiliary variable (x_i).

$$\bar{Y}_{ratio} = \frac{\sum_{sample} \hat{t}_i}{\sum_{sample} M_i}$$

- The approximate variance formula is messy.

- See page 148.

Example 5.6, page 148

- File name is `coots.dat`.

- American coot eggs from Minnedosa, Manitoba.

- Clusters (psus) are clutches or nests of eggs.

- Two eggs (ssus) from each clutch were measured.

- We will look at the egg volume.

Some details

- The sample size for clutches is $n = 184$.
- The population size N is unknown.
- The number of eggs in each clutch is M_i and varies.
- We have a sample of $m_i = 2$ eggs from each clutch.
- We will use a ratio estimate.

Import and check the data (`SLL148.sas`)

```
options nocenter;
proc contents data=a1;
proc print data=a1;
run;
```

The data

```
Obs CLUTCH CSIZE        VOLUME
  1      1     13    3.7957569
  2      1     13    3.9328497
  3      2     13    4.2156036
  4      2     13    4.1727621
  5      3      6    0.9317646
  6      3      6    0.9007362
```

Calculate some clutch summaries

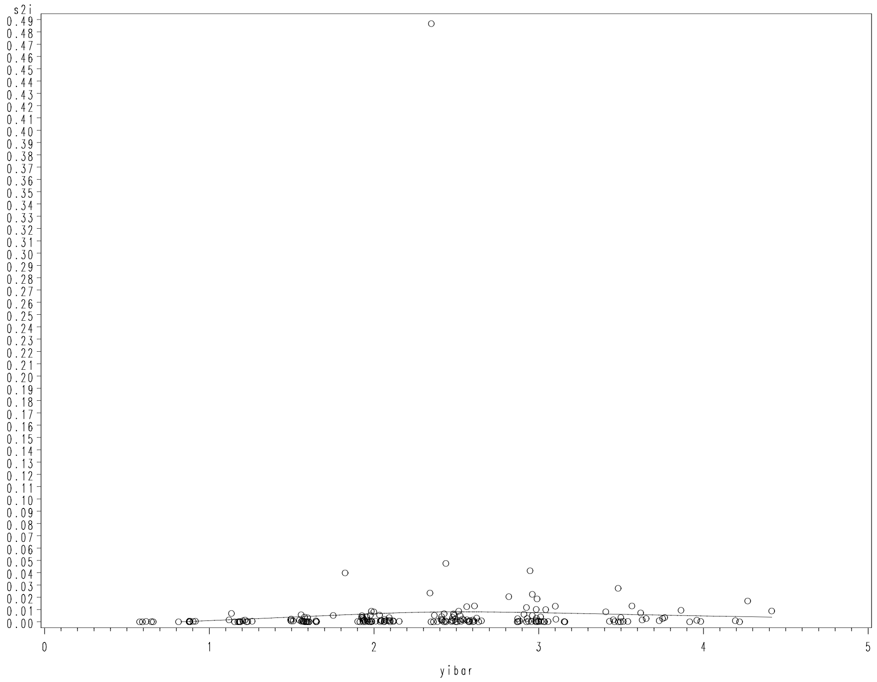
```
proc means data=a1 noprint;
   by clutch;
   var volume csize;
   output out=a2
     mean=yibar Micap var=s2i x1
     sum=tihat x2 n=milow x3;
data a2; set a2;
   keep clutch Micap yibar
        s2i tihat milow;
proc print data=a2;
run;
```

Output

```
Obs     CLUTCH           yibar      Micap       s2i      tihat milow
  1          1       3.8643033       13    0.009397  7.7286066     2
  2          2      4.19418285       13    0.000918  8.3883657     2
  3          3       0.9162504        6    0.000481  1.8325008     2
  4          4      2.99833465       11    0.000795  5.9966693     2
```

A plot

```
symbol1 v=circle i=sm70;
proc sort data=a2; by yibar;
proc gplot data=a2;
   plot s2i*yibar/frame;
run;
```



Find the outlier

```
proc print data=a2;
   where s2i ge .2;
run;
```

Output

```
Obs    clutch     yibar     Micap      s2i      milow     tihat
 89      88      2.34829      10     0.48669      2      23.4829
```

Clutch 88

```
proc print data=a1;
   where clutch eq 88;
run;
```

Output

```
Obs    clutch    csize    length    breadth     volume    tmt
176      88         9      45.17     32.69     1.85500     0
177      88        11      46.32     32.14     2.84159     0
```

Find the estimate

```
proc means data=a2 noprint;
   var tihat Micap;
   output out=a3
      sum=Stihat SMicap;
data a3; set a3;
   ybar_rat=Stihat/SMicap;
proc print data=a3;
run;
```

Output

```
Obs     Stihat    SMicap    ybar_rat
 1     4378.29     1758      2.49050
```

Calculations for the SE

```
data a4; set a2;
   if _n_ eq 1 then set a3;
   Npop=1000000;
   *a very large number;
   fpci=1-milow/micap;
   withini2=fpci*(micap**2)*s2i
              /milow;
   betwi=tihat-Micap*ybar_rat;
   betwi2=betwi*betwi;
proc print data=a4; run;
```

SE

- Finish calculations using outline given for Example 5.6 on page 151.
- SE expressed as relative error is 2.45%.

Final Comment

Unbiased estimation does not work well (ratio estimation works better) when

$$\begin{aligned} \text{Var}(M_i) &= \textit{constant} \\ t_i &\propto M_i \end{aligned}$$

Weights

- In many practical situations, weights are used for estimates with cluster sampling.
- The weight of an element is the reciprocal of its probability of selection.

- Consider ssu corresponding to $y_{i,j}$.
- First, we need to have psu i selected in the first stage
  - The probability is n/N
- Then, ssu j needs to be selected.
  - The probability is m_i/M_i.
- So the probability that $y_{i,j}$ is selected is nm_i/NM_i.
- And the weight is NM_i/nm_i.

Estimates

- For total, multiply by the weights and then sum.
- For mean, divide total by the sum of the weights in the sample.
- This is a ratio estimator.
- If N is unknown, relative weights can be used, but the total cannot be estimated.

Design issues

- Precision needed
- Size of the psu
- How many ssus to sample within each selected psu
- How many psus to select

PSU

- Often this is some natural unit.
  - Clutch of eggs
  - Class of children
- Sometimes we have some control.
  - Area of a forest
  - Time interval for customers
- Principle – more area ⇒ more variability within psu's (ICC smaller)

Subsampling sizes

- The relative sizes of MSB and MSW are relevant.
- $R^2_{adj} = 1 - \frac{MSW}{MST}$ is the adjusted R^2.
- If units within clusters are very similar relative to units from other clusters, we do not need to sample large numbers within each cluster.

Cost

- One approach to determining sample sizes is to consider costs
- c_1 is the cost of obtaining a psu.
- c_2 is the cost of obtaining a ssu.
- C is the total cost

$$C = c_1 n + c_2 nm$$

Minimum cost

- Use calculus to find n and m that minimize the variance of the estimator.

$$n = \frac{C}{c_1 + c_2 m}$$

- Formula for m involves MSW and MSB (or R^2_{adj})
- See page 156.
- We are assuming the cluster sizes are equal (M).

Other issues

- For unequal cluster sizes the same approach is reasonable.
- Use $\bar{M}$ and $\bar{m}$ in place of M and m.
- Then take $m_i = \bar{m}$ or
- if the M_i do not vary very much, we often take m_i proportional to M_i ($\frac{m_i}{M_i} = constant$)

PSU's

- The number of psus to sample (n) can be determined from the desired MOE using some approximations.
- See Section 5.5.3 on pages 158-159.

Systematic sampling

- We mentioned earlier that systematic sampling is a special case of cluster sampling.
- It is a one-stage cluster sample.
- Suppose we take every 10th unit.
- Then the ten clusters are $\{1, 11, \dots\}$, $\{2, 12, \dots\}$, $\{3, 13, \dots\}$, $\dots$ $\{10, 20, \dots\}$.

Variance

- The variance of the estimate of the population mean for systematic sampling is approximately

$$\frac{S^2}{M}(1 + (M - 1)ICC)$$

- Here M is the size of the systematic sample.
- If the ICC is zero this is the variance for an SRS.

ICC

- If the ICC is negative, the systematic sample is better that an SRS.
  - This happens when the within cluster variation is *larger* that the overall variance (clusters are diverse).
  - If the ICC is positive, then SRS is better.

Example

- List in random order.
  - Systematic similar to SRS.
- List is in decreasing or increasing order based on something correlated with y.
  - Systematic better to SRS.
- Periodic pattern in the list
  - Could be a disaster

Advantage

Puts absolute (not probabilistic) bounds on event detection.

Periodicity

- One remedy is to take more than one systematic sample.
- This is called *interpenetrating systematic sampling*.
- Each systematic sample is viewed as a cluster and the methods of this chapter apply.

Models for cluster sampling

- Basic idea is the one-way anova model with random effects

$$Y_{i,j} = A_i + e_{i,j}$$

- (Fixed effects one-way model used for stratified sampling.)
- Where A_i and $e_{i,j}$ are independent with means μ and zero, and variances σ^2 and σ_A^2, respectively.

ICC

- The intraclass correlation coefficient (ICC) is

$$\rho = \frac{\sigma_A^2}{\sigma^2 + \sigma_A^2}$$

- Note that this quantity is always nonnegative (not appropriate if "competing resources").

$$\mathrm{Cov}_{M1}(Y_{i,j}, Y_{k,\ell}) = \sigma_A^2 I(i = k) + \sigma^2 I(j = \ell)$$

- We can use this framework to derive formulas for the SEs.

Properties

- Design-unbiased estimators can be model-biased when error variance assumed constant. (Ratio estimator unbiased.)
- Important diagnostic: does $\mathrm{Var}(\hat{T})$ depend on M_i?
- Different model assumptions can lead to different designs.