Methods in Sample Surveys

140.640

Cluster Sampling

Saifuddin Ahmed
Dept. of Biostatistics
School of Hygiene and Public Health
Johns Hopkins University
Cluster Sampling

Consider that we want to estimate health insurance coverage in Baltimore city. We could take a random sample of 100 households (HH). In that case, we need a sampling list of Baltimore HHs. If the list is not available, we need to conduct a census of HHs. The complete coverage of Baltimore city is required so that all HHs are listed, which could be expensive. Furthermore, since our sample size is small compared to the numbers of total HHs, we need to sample only few, say one or two, in each block (subdivisions). Alternatively, we could select 5 blocks (say the city is divided into 200 blocks), and in each block interview 20 HHs. We need to construct HH listing frame only for 5 blocks (less time and costs needed). Furthermore, by limiting the survey to a smaller area, additional costs will be saved during the execution of interviews.

Such sampling strategy is known as “cluster sampling.”

The blocks are “Primary Sampling Units” (PSU) – the clusters. The households are “Secondary Sampling Units” (SSU).

Definition:

In cluster sampling, cluster, i.e., a group of population elements, constitutes the sampling unit, instead of a single element of the population.

The main reason for cluster sampling is “cost efficiency” (economy and feasibility), but we compromise with variance estimation efficiency.

Advantages:
- Generating sampling frame for clusters is economical, and sampling frame is often readily available at cluster level
- Most economical form of sampling
- Larger sample for a similar fixed cost
- Less time for listing and implementation
- Also suitable for survey of institutions

Disadvantages:
- May not reflect the diversity of the community.
- Other elements in the same cluster may share similar characteristics.
- Provides less information per observation than an SRS of the same size (redundant information: similar information from the others in the cluster).
- Standard errors of the estimates are high, compared to other sampling designs with same sample size
Need to consider the sampling order:

- Primary sampling units (PSU): clusters
- Secondary sampling units (SSU): households/individual elements

1. We may select the PSU’s by using a specific element sampling techniques, such as simple random sampling, systematic sampling or by PPS sampling.

2. We may select all SSU’s for convenience or few by using a specific element sampling techniques (such as simple random sampling, systematic sampling or by PPS sampling).

Simple one-stage cluster sample:

List all the clusters in the population, and from the list, select the clusters – usually with simple random sampling (SRS) strategy. **All units** (elements) in the sampled clusters are selected for the survey.

Simple two-stage cluster sample:

List all the clusters in the population. First, select the clusters, usually by simple random sampling (SRS). The units (elements) in the selected clusters of the first-stage are then sampled in the second-stage, usually by simple random sampling (or often by systematic sampling).

Multi-stage sampling:

- when sampling is done in more than one stage.
- In practice, clusters are also stratified.

**Question**: Is sampling with probability proportional to size (PPS) a variant of cluster sampling?

**Theory**:

1. It is assumed that population elements are clustered into N groups, i.e., in N clusters (PSUs).

2. Let the size of cluster is $M_i$, for the $i$-th cluster, i.e., the number of elements (SSUs) of the $i$-th cluster is $M_i$.

3. The corresponding number of PSUs (clusters) in sample = n, and the number of elements from the $i$-th PSU = $m_i$. 
Estimation for cluster sampling

Let $y_{ij}$ = measurement for $j$-th element (SSU) in $i$-th cluster (PSU).

In the simple case of equal-sized clusters (although may be unrealistic), the total number of elements in the population,

$$K = N^*M$$

where $M_i = M$ (constant for all the clusters)

If the clusters are of unequal sizes, the total number of elements in the population:

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

Total in the $i$-th population:

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

Estimated sample total for the $i$th PSU:

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

Population total:

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$

Estimated sample total for population:

$$\hat{t} = \sum_{j \in S_i} \hat{t}_i$$

Estimated (unbiased) total for population:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{j \in S_i} t_i$$

Population mean in the $i$-th cluster:

$$\bar{y}_{i,clu} = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i} = \frac{t_i}{M_i}$$

Sample mean for the $i$-th PSU:

$$\bar{y}_{i,clu} = \sum_{j \in S_i} \frac{y_{ij}}{m_i} = \frac{\hat{t}_i}{m_i}$$

Population mean:

$$\bar{y}_{clu} = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$

Sample mean (unbiased):

$$\bar{\hat{y}}_{clu} = \frac{\hat{t}}{\sum_{i \in S} m_i}$$
Variance estimation:

\[ \hat{t}_{unb} = \frac{N}{n} \sum_{j=1}^{s} t_j = N \frac{\sum_{j=1}^{s} t_j}{n} = N\bar{y}_{total}, \text{ where } \bar{y} \text{ is the mean "total" for the clusters} \]

Then, variance:

\[ \text{var}(\hat{t}_{unb}) = N^2 \frac{S_i^2}{n} \left(1 - \frac{n}{N}\right) \]

where,

\[ S_i^2 = \frac{\sum_{i=1}^{N} \left( \frac{t_i - \bar{t}}{N} \right)^2}{N - 1} \]

Note: Variance of total is likely to be larger with unequal cluster sizes.

The mean (with clusters of equal sizes):

\[ \hat{y}_{cha} = \frac{\hat{t}}{NM}, \text{ (because of the equal size } M_i = m_j = M ) \]

The variance of mean is then:

\[ \text{var}(\hat{y}) = \frac{1}{N^2 M^2} \text{ var}(\hat{t}) = \frac{N^2}{N^2} \frac{S_i^2}{nM^2} \left(1 - \frac{n}{N}\right) = \frac{S_i^2}{nM^2} \left(1 - \frac{n}{N}\right) \]

**Intra-class Correlation**

Intra-class correlation reflects the homogeneity of sample.

We may decompose the variance into:
\[ \sigma^2 = \sigma^2_w + \sigma^2_b, \]

that is,

Total variance = variance within + variance between

Intra-class correlation is defined as:

\[ \rho = 1 - \frac{\sigma^2_w}{\sigma^2} = \frac{\sigma^2_b}{\sigma^2 + \sigma^2_w} \]

More specifically:

\[ \rho = 1 - \frac{n}{n - 1} \frac{\sigma^2_w}{\sigma^2} \]

Minimum: When \( \sigma^2_b = 0 \), \( \rho = -1/(n-1) \)

Maximum: When \( \sigma^2_w = 0 \), \( \rho = 1 \)

Derivation of Variance for Cluster Sampling

\[ \rho = 1 - \frac{n}{n - 1} \frac{\sigma^2_w}{\sigma^2} \]

\[ \rho = \frac{(n - 1)\sigma^2 - n\sigma^2_w}{(n - 1)\sigma^2} \]

\[ \Rightarrow n\sigma^2 - \sigma^2 - n(\sigma^2 - \sigma^2_b) = \sigma^2 (n - 1)\rho \]

\[ \Rightarrow n\sigma^2_b = \sigma^2 + \sigma^2 (n - 1)\rho \]

\[ \Rightarrow \sigma^2_b = \frac{\sigma^2}{n} [1 + (n - 1)\rho] \]

\[ \text{var}(\bar{x}) = \frac{\sigma^2}{n} [1 + (n - 1)\rho] \]
Let consider a single-stage cluster sampling, where n units of sample is selected from N clusters, and the (average) size of cluster is M, then the variance of y is:

$$Var_{clu} (y) = \left( \frac{\sigma^2}{nM} \right) [1 + (M-1) \rho]$$

and,

$$Deff = 1 + (M - 1) \rho$$

In cluster sampling, the size of \( \rho \) could be quite large, that may seriously affect the precision of estimates.

In general, as cluster size increases \( \rho \) decreases, but deff depends on both M and \( \rho \), so in cluster sampling, increase in cluster size make sampling more inefficient.

As an example, for a size of cluster 20, if \( \rho = 0.1 \), the \( deff = 1 + (20-1) * 0.1 \) = 2.9 suggesting that the actual variance is 2.9 times above what it would have been with variance from SRS with same sample size. However, if the size of cluster is large, say \( m=200 \), \( deff = 1 + (200-1) * 0.1 = 20.9! \)

When \( \rho = 0.0 \), \( deff = 1 \).

This relationship has important implications for cluster sampling strategies.

Consider a sampling scenario: we need to draw 300 samples. We may draw 10 clusters with 30 elements, or draw 3 clusters with 100 elements. We have said earlier, the principal reason of conducting cluster sampling is to reduce costs. Obviously, the 2\textsuperscript{nd} option is cheaper as we need to go to only 3 clusters. However, as we have shown above, larger the \( m \) size (cluster size), larger the \( deff \). As a result, the first option should be implemented (take more clusters with fewer elements) as a balance between “cost efficiency” and “variance efficiency.”

<table>
<thead>
<tr>
<th>Lessons for Cluster Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use as many clusters as feasible.</td>
</tr>
<tr>
<td>Use smaller cluster size in terms of number of households/individuals selected in each cluster.</td>
</tr>
<tr>
<td>Use a constant “take size” rather than a variable one (say 30 households from each cluster).</td>
</tr>
</tbody>
</table>
Example:

Let us see an example.

```
list area age, clean
```

<table>
<thead>
<tr>
<th>area</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
</tr>
</tbody>
</table>

```

```
. sum age
Variable | Obs | Mean | Std. Dev. | Min | Max
---------+-----+------|-----------|------|-----
age      | 22  | 25   | 6.055301  | 15  | 35
```

```
. ci age
Variable | Obs | Mean | Std. Err. | [95% Conf. Interval]
---------+-----+------|-----------|------------------
age      | 22  | 25   | 1.290994    | 22.31523        27.68477
```

```
. oneway age area
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>550</td>
<td>1</td>
<td>550</td>
<td>50.00</td>
<td>0.0000</td>
</tr>
<tr>
<td>Within groups</td>
<td>220</td>
<td>20</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>770</td>
<td>21</td>
<td>36.6666667</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
*SE under SRS

. disp sqrt((770/21)/22)
1.2909944

UNDER CLUSTER SAMPLING:

svyset, psu(area)
psu is area

. svymean age

Survey mean estimation

pweight: <none>                     Number of obs  =   22
Strata: <one>                      Number of strata =   1
PSU:  area                         Number of PSUs =   2
Population size = 22

---------------------------------------------------------
                      Std. Err.  [95% Conf. Interval]  Deff
---------------------------------------------------------
    age |   25   5    -38.53102    88.53102         15

*Direct estimation of SE under cluster sampling design

. disp sqrt((550/1)/22)
5

*Estimation of deff:
. di 5^2/1.290994^2
15.00001

Use of STATA to estimate intra-class correlation

1. loneway

. loneway age area

One-way Analysis of Variance for age:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between area</td>
<td>550</td>
<td>1</td>
<td>550</td>
<td>50.00</td>
<td>0.0000</td>
</tr>
<tr>
<td>Within area</td>
<td>220</td>
<td>20</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>770</td>
<td>21</td>
<td>36.666667</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Intraclass Correlation and Variance Components

<table>
<thead>
<tr>
<th></th>
<th>Asy. S.E.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intraclass correlation</strong></td>
<td>0.81667</td>
<td>0.22140</td>
</tr>
<tr>
<td><strong>[95% Conf. Interval]</strong></td>
<td>0.38274</td>
<td>1.25059</td>
</tr>
</tbody>
</table>

**Estimated SD of area effect** 7
**Estimated SD within area** 3.316625
**Est. reliability of a area mean** 0.98000
(evaluated at n=11.00)

---

In longway command, *icc(\(\rho\))* is estimated by:

Rho = \(\frac{\text{MSB} - \text{MSW}}{\text{MSB} + (m-1)\text{MSW}}\)

MSB = Mean square between
MSW = Mean square within
M = (average) size of the cluster

\[ \text{di } (550-11)/(550+(11-1)*11) \]
\[ .81666667 \]

2. *xt – command:*

\texttt{xtreg age, i(area)}

---

### Random-effects GLS regression

|                  | Coef. | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|------------------|-------|-----------|-------|------|----------------------|
| _cons_           | 25    | 5         | 5.00  | 0.000| 15.20018             |
|                  |       |           |       |      | 34.79982             |

\texttt{sigma_u} | 7
\texttt{sigma_e} | 3.3166248
\texttt{rho}    | .81666667  (fraction of variance due to u_i)

---

*How icc (rho) is measured:
\[ \text{di } 7^2/(3.3166248^2+7^2) \]
\[ .81666667 \]
However, estimating ICC from binary outcome is done differently:

```
. ta area adversehealth, row

<table>
<thead>
<tr>
<th>adversehealth</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>27.27</td>
<td>72.73</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>72.73</td>
<td>27.27</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>50.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
```

```
. xtlogit adverse, i(area)
```

Fitting comparison model:
Iteration 0:  log likelihood = -15.249238

Fitting full model:

```
Random-effects logistic regression  Number of obs      =        22
Group variable (i): area            Number of groups   =         2
Random effects u_i ~ Gaussian       Obs per group: min =        11
                                    avg =      11.0
                                 max =        11
Log likelihood  = -14.730665        Wald chi2(0)       =      0.00
                                    Prob > chi2        =         .
```

```
                               | Coef.  Std. Err.   z   P>|z|   [95% Conf. Interval]
adversehealth                |       
_cons                         | -1.12e-15  .7128713 -0.00  1.000   -1.397202    1.397202
```

```
                               |        
/insig2u                       | -.5081339  1.802657 -0.00  1.000   -4.041277    3.025009
```

```
                               |        
sigma_u                        | .7756399  .6991063 .1325708    .8622567
rho                            | .1545983  .2356031 .0053138    .4.538082
```

Likelihood-ratio test of rho=0: chi2(01) = 1.04 Prob >= chi2 = 0.154

If the error term is considered to have standard logistic distribution, the variance of error term is $\pi^2/3$

So, $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}}$

```
di .7756399^2/(.7756399^2 + pi^2/3) .15459836
```
SAMPLE SIZE ESTIMATION under CLUSTER SAMPLING:

The major issue: DEFF > 1.0

Solutions:

1. Increase the sample size estimated under SRS by multiplying with an estimated DEFF (from published source, or estimate from the formula as stated below):

\[ \text{deff} = 1 + (m-1)\rho \]

Consider the comparison between:

\[ \frac{\sigma^2}{n} \text{ ... variance under SRS} \]

vs.

\[ \frac{\sigma^2}{nm} [1 + (m-1)\rho] \text{ ... variance under cluster sampling} \]

So, transform sample size estimation formula,

\[ n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(d)^2} \]

to:

\[ nm = \frac{2(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(d)^2} [1 + (m-1)\rho] \text{....total....sample....of ....individuals (n clusters of m size)} \]

In practice, m ~30 and, \( \rho \) is kept very (very) small. The deff values are available from published reports (e.g., Demographic and Health Survey reports). Usually a value of 1.5 to 2.0 for deff is considered for sample size estimation.
Note that DHS (as shown above) reports “deft” which is the “squared of deff”, i.e.,

deft = \frac{\text{std.error(cluster)}}{\text{std.error(srs)}}.

2. You may also calculate the number of clusters required for the study utilizing the
above formulas.

Source: Bangladesh DHS
Essentially, you need the same sample size formula for “randomized community trial.” However, *deff* is called “variance inflation factor” in the randomized community trial (essentially borrowed from survey statistics!).

3. Other methods:

Direct estimation of the number of clusters needed for a survey:

**Exact:**

\[
m = \frac{Z_{1-\alpha/2}^2 MV^2}{Z_{1-\alpha/2}^2 V^2 + (M - 1)d^2}
\]

Example: M= 100 clusters in the population

Need to know (research question): average number of children in the population, based on 100 clusters, for designing a health care facility needs study with following info:

\[
\sigma^2 = 0.5 \\
Y(\text{mean}): 5.6
\]

\[
V^2 = \frac{\sigma^2}{Y} = \frac{0.5}{(5.6)^2} = .01594388
\]

STATA command:

```
. di "m = " (9*100*.01594388)/(9*.01594388+99*(.10^2))
m = 12.659512 ~ 13 clusters
```

Note : (1.96+ 0.84)^2 ~ 9 {for faster calculation}

**Approximate method:**

\[
m = Z_{1-\alpha/2}^2 V^2 \\
\text{STATA Command:}
\text{di " m = " (9*.01594388)/(.10^2)}
m = 14.349492 ~ 15 clusters
\]