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 ith 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:

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

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:

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

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):

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

Variance estimation:

$$\hat{t}_{unb} = \frac{N}{n}\sum_{j\in S_i} t_i = N\,\frac{\sum_{j\in S_i} t_i}{n} = N\bar{y}_{total} \quad , where\ \bar{y}\ is\ the\ mean\ "total"\ for\ the\ clusters$$

Then, variance:

$$var(\hat{t}_{unb}) = N^2\,\frac{S_t^{\,2}}{n}\left(1-\frac{n}{N}\right)$$

$$where,$$

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

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

The mean (with clusters of equal sizes):

$$\hat{\bar{y}}_{clu} = \frac{\hat{t}}{NM} \quad , (because\ of\ the\ equal\ size\ M_i = m_i = M\ )$$

The variance of mean is then:

$$var(\hat{\bar{y}}) = \frac{1}{N^2M^2}\ var(\hat{t}) = \frac{N^2}{N^2}\frac{S_t^{\,2}}{nM^2}\left(1-\frac{n}{N}\right) = \frac{S_t^{\,2}}{nM^2}\left(1-\frac{n}{N}\right)$$

Intra-class Correlation

Intra-class correlation reflects the homogeneity of sample.

We may decompose the variance into:

$$\sigma^2 = \sigma_w^2 + \sigma_b^2,$$
$$\textit{that is},$$
$$Total \ \ \mathrm{var}\,iance = \mathrm{var}\,iance\_within + \mathrm{var}\,iance\_between$$

Intra-class correlation is defined as:

$$\rho = 1 - \frac{\sigma_w^2}{\sigma^2} = \frac{\sigma_b^2}{\sigma^2} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

More specifically:

$$\rho = 1 - \frac{n}{n-1}\frac{\sigma_w^2}{\sigma^2}$$

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

$Maximum: When \ \ \sigma_w^2 = 0, \ \rho = 1$

Derivation of Variance for Cluster Sampling

$$\rho = 1 - \frac{n}{n-1}\frac{\sigma_w^2}{\sigma^2}$$

$$\rho = \frac{(n-1)\sigma^2 - n\sigma_w^2}{(n-1)\sigma^2}$$

$$=> n\sigma^2 - \sigma^2 - n(\sigma^2 - \sigma_b^2) = \sigma^2\,(n-1)\rho$$
$$=> n\sigma_b^2 = \sigma^2 + \sigma^2\,(n-1)\rho$$

$$=> \sigma_b^2 = \frac{\sigma^2}{n}[1 + (n-1)\,\rho]$$
$$\mathrm{var}(\bar{x}\ \ ) = \frac{\sigma^2}{n}[1 + (n-1)\,\rho]$$

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:

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

and,

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

In cluster sampling, the size of ρ could be quite large, that may seriously affect the precision of estimates.

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

As an example, for a size of cluster 20, if ρ = 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 ρ = 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 2nd 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."

Lessons for Cluster Sampling

- **Use as many clusters as feasible.**
- **Use smaller cluster size in terms of number of households/individuals selected in each cluster.**
- **Use a constant "take size" rather than a variable one (say 30 households from each cluster).**

Example:

Let us see an example.

```
list area age, clean

       area   age
  1.      1    15
  2.      1    16
  3.      1    17
  4.      1    18
  5.      1    19
  6.      1    20
  7.      1    21
  8.      1    22
  9.      1    23
 10.      1    24
 11.      1    25
 12.      2    25
 13.      2    26
 14.      2    27
 15.      2    28
 16.      2    29
 17.      2    30
 18.      2    31
 19.      2    32
 20.      2    33
 21.      2    34
 22.      2    35

. 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

                        Analysis of Variance
    Source              SS         df       MS            F     Prob > F
------------------------------------------------------------------------
Between groups             550      1          550     50.00      0.0000
 Within groups             220     20           11
------------------------------------------------------------------------
    Total                  770     21   36.6666667
```

```
*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

------------------------------------------------------------------------------
    Mean |   Estimate    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:

                                                 Number of obs =        22
                                                     R-squared =    0.7143

    Source                SS          df      MS            F     Prob > F
-------------------------------------------------------------------------
Between area                 550       1          550     50.00     0.0000
Within area                  220      20           11
-------------------------------------------------------------------------
Total                        770      21    36.666667
```

```
Intraclass       Asy.
correlation      S.E.       [95% Conf. Interval]
------------------------------------------------
   0.81667     0.22140       0.38274     1.25059

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 loneway command, *icc(ρ)* is estimated by:

Rho= (MSB-MSW)/(MSB+(m-1)MSW)

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

```
. di (550-11)/(550+(11-1)*11)
.81666667
```

2. xt – command:

```
xtreg age, i(area)

Random-effects GLS regression                   Number of obs      =        22
Group variable (i): area                        Number of groups   =         2

R-sq:  within  =      .                         Obs per group: min =        11
       between =      .                                        avg =      11.0
       overall = 0.0000                                        max =        11

Random effects u_i ~ Gaussian                   Wald chi2(0)       =      0.00
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =         .

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |         25          5     5.00   0.000     15.20018    34.79982
-------------+----------------------------------------------------------------
     sigma_u |          7
     sigma_e |  3.3166248
         rho |  .81666667   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

***How icc (rho) is measured:**

```
di 7^2/(3.3166248^2+7^2)
.81666667
```

However, estimating ICC from binary outcome is done differently:

```
.  ta area adversehealth, row
           |     adversehealth
      area |         0          1 |     Total
-----------+----------------------+----------
         1 |         3          8 |        11
           |     27.27      72.73 |    100.00
-----------+----------------------+----------
         2 |         8          3 |        11
           |     72.73      27.27 |    100.00
-----------+----------------------+----------
     Total |        11         11 |        22

          |     50.00      50.00 |    100.00
```

```
. 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

                                                Wald chi2(0)       =      0.00
Log likelihood  = -14.730665                    Prob > chi2        =         .

------------------------------------------------------------------------------
adversehea~h |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -1.12e-15   .7128713    -0.00   1.000    -1.397202    1.397202
-------------+----------------------------------------------------------------
    /lnsig2u |  -.5081339   1.802657                     -4.041277    3.025009
-------------+----------------------------------------------------------------
     sigma_u |   .7756399   .6991063                      .1325708    4.538082
         rho |   .1545983   .2356031                      .0053138    .8622567
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =     1.04 Prob >= chibar2 = 0.154
```

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

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

```
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:

$$\frac{\sigma^2}{n} \text{ ... } \mathrm{var}\,iance\ under\ SRS$$
$$vs.$$
$$\frac{\sigma^2}{nm}[1+(m-1)\rho.... \mathrm{var}\,iance\ under\ clster\ sampling$$

So, transform sample size estimation formula,

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

to:

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

In practice, m ~30 and, ρ 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.

Table B.2 Sampling errors: Total sample, Bangladesh 2004

			Number of cases				Confidence limits	
Variable	Value (R)	Standard error (SE)	Un-weighted (N)	Weight-ed (WN)	Design effect (DEFT)	Relative error (SE/R)	R-2SE	R+2SE
		WOMEN						
Urban residence	0.226	0.006	11,440	11,440	1.557	0.027	0.214	0.238
No education	0.412	0.008	11,440	11,440	1.780	0.020	0.396	0.428
With secondary education or higher	0.294	0.008	11,440	11,440	1.787	0.026	0.279	0.310
Currently married	0.925	0.003	11,440	11,440	1.102	0.003	0.920	0.930
Currently pregnant	0.051	0.002	13,543	13,542	1.122	0.041	0.047	0.056
Children ever born	2.998	0.028	10,417	10,436	1.300	0.009	2.941	3.054
Children surviving	2.591	0.022	10,417	10,436	1.240	0.009	2.547	2.635
Children ever born to women 40-49	5.118	0.072	2,263	2,230	1.415	0.014	4.974	5.261
Ever used any contraceptive method	0.828	0.006	10,553	10,582	1.705	0.008	0.815	0.840
Currently using any contraceptive method	0.581	0.007	10,553	10,582	1.433	0.012	0.567	0.594
Currently using a modern method	0.473	0.007	10,553	10,582	1.451	0.015	0.459	0.487
Currently using pill	0.262	0.006	10,553	10,582	1.352	0.022	0.251	0.274
Currently using IUD	0.006	0.001	10,553	10,582	1.154	0.143	0.004	0.008
Currently using condom	0.042	0.003	10,553	10,582	1.346	0.063	0.037	0.047
Currently using injectables	0.097	0.005	10,553	10,582	1.819	0.054	0.086	0.107
Currently using female sterilization	0.052	0.004	10,553	10,582	1.741	0.072	0.044	0.060
Currently using periodic abstinence	0.065	0.003	10,553	10,582	1.290	0.048	0.059	0.071
Currently using withdrawal	0.036	0.002	10,553	10,582	1.178	0.059	0.032	0.040
Currently using Norplant	0.008	0.001	10,553	10,582	1.251	0.137	0.006	0.010
Obtained method from public sector source	0.573	0.011	4,994	5,053	1.602	0.020	0.550	0.595
Want no more children	0.628	0.006	10,553	10,582	1.188	0.009	0.617	0.640
Want to delay birth at least 2 years	0.212	0.004	10,553	10,582	1.096	0.021	0.203	0.220
Ideal number of children	2.420	0.013	11,012	11,017	1.840	0.006	2.393	2.446
Mothers received ANC (trained provider)	0.487	0.13	5,366	5,416	1.936	0.027	0.460	0.513
Mothers received tetanus injection (last birth)	0.848	0.009	5,366	5,416	1.837	0.011	0.830	0.866
Mothers received medical care at delivery	0.132	0.006	6,908	7,002	1.447	0.049	0.119	0.145
Child had diarrhea in the last 2 weeks	0.075	0.004	6,424	6,498	1.064	0.048	0.068	0.082
Treated with ORS packets	0.672	0.026	485	486	1.193	0.039	0.619	0.724
Sought medical treatment	0.157	0.018	485	486	1.085	0.117	0.120	0.193
Child having health card, seen	0.494	0.017	1,247	1,265	1.199	0.034	0.460	0.528
Child received BCG vaccination	0.934	0.012	1,247	1,265	1.671	0.013	0.911	0.958
Child received DPT vaccination (3 doses)	0.810	0.017	1,247	1,265	1.576	0.022	0.775	0.845
Child received polio vaccination (3 doses)	0.823	0.017	1,247	1,265	1.554	0.020	0.789	0.856
Child received measles vaccination	0.757	0.019	1,247	1,265	1.600	0.026	0.718	0.795
Child fully immunized	0.731	0.020	1,247	1,265	1.563	0.027	0.692	0.770
Height-for-age (-2SD)	0.430	0.009	6,012	6,005	1.421	0.022	0.411	0.449
Weight-for-height (-2SD)	0.128	0.005	6,012	6,005	1.105	0.038	0.119	0.138
Weight-for-age (-2SD)	0.475	0.010	6,012	6,005	1.534	0.022	0.454	0.496
BMI < 18.5	0.343	0.006	10,448	10,431	1.373	0.019	0.330	0.356
Has heard of HIV/AIDS	0.600	0.011	11,440	11,440	2.407	0.018	0.578	0.622
Knows about condoms	0.219	0.008	11,440	11,440	1.987	0.035	0.203	0.234
Knows about limiting partners	0.181	0.007	11,440	11,440	1.953	0.039	0.167	0.195
Total fertility rate (last 3 years)	3.028	0.067	na	38,850	1.497	0.022	2.894	3.161
Neonatal mortality (last 5 years)	41.373	2.861	6,967	7,056	1.149	0.069	35.652	47.095
Post-neonatal mortality (last 5 years)	23.822	2.048	6,978	7,065	1.133	0.086	19.725	27.918
Infant mortality (last 5 years)	65.195	3.604	6,980	7,068	1.181	0.055	57.986	72.403
Child mortality (last 5 years)	23.936	2.434	7,038	7,133	1.282	0.102	19.068	28.805
Under-five mortality (last 5 years)	87.571	4.327	7,053	7,148	1.239	0.049	78.917	96.224

Source: Bangladesh DHS

Note that DHS (as shown above) reports “deft” which is the “squared of deff”, ie., deft=std.error(cluster)/std.error(srs).

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

$$n = \frac{2*(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{m(d)^2}[1 + (m-1)p].....no\ \ of\ \ clusters$$

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:

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

   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(mean): 5.6

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

   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:

$$m = \frac{Z_{1-\alpha/2}^2 V^2}{d^2}$$

   STATA Command:

```
di “ m = “ (9*.01594388)/(.10^2)
      m =  14.349492 ~ 15 clusters
```