JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Methods in Survey Sampling
Biostat 140.640

Stratified Sampling

Saifuddin Ahmed, PhD
Dept. of Biostatistics

Stratified Sampling

In stratified sampling the population is partitioned into groups, called *strata*, and sampling is performed separately within each *stratum*.

When?

- Population groups may have different values for the responses of interest.

- If we want to improve our estimation for each group separately.

- To ensure adequate sample size for each group.

In stratified sampling designs:

- stratum variables are mutually exclusive (non-over lapping), e.g., urban/rural areas, economic categories, geographic regions, race, sex, etc.

- the population (elements) should be *homogenous* within-stratum, and

- the population (elements) should be *heterogenous* between the strata.

Advantages

- Provides opportunity to study the stratum variations - estimation could be made for each stratum
- Disproportionate sample may be selected from each stratum
- The precision likely to increase as variance may be smaller than SRS with same sample size
- Field works can be organized using the strata (e.g., by geographical areas or regions)
- Reduce survey costs.

The principal objective of stratification is to reduce sampling errors.

Disadvantages

- Sampling frame is needed for each stratum
- Analysis method is complex
  - Correct variance estimation
- Data analysis should take sampling “weight” into account for disproportionate sampling of strata
- Sample size estimation is difficult in practice

When sample is selected by SRS technique independently within each stratum, the design is called *stratified random sampling.*

Theory of Stratified Sampling

With systematic sampling, the target population is partitioned into H > 1 non-overlapping subpopulations of strata.

If the population size consists of N discrete elements, then under stratified sampling,

$$N = N_1 + N_2 + N_3 + \ . \ . \ . \ + N_H$$

That is,

$$N = \sum_{h=1}^{H} N_h$$

Estimation of Total
for a random variable *y*

Let y_{hi} = value of i_{th} unit in stratum h

Then, population total for stratum h is:

$$t_h = \sum_{i=1}^{N_h} y_{hi}$$

And, population total is:

$$t = \sum_{h=1}^{H} t_h$$

That is, $t = t_1 + t_2 + t_3 + \ . \ . \ . \ + t_H$

(compare to: $N = N_1 + N_2 + N_3 + \ . \ . \ . \ + N_H$)

Strata totals are additive

But, not the strata means

Population mean for strata h is:

$$\bar{y}_h = \frac{t_h}{N_h}$$

However,

$$\bar{y} \neq \bar{y}_1 + \bar{y}_2 + ... + \bar{y}_H$$

Because,

$$\bar{y} = \frac{t}{N} = \frac{t_1 + t_2 + ... + t_H}{N_1 + N_2 + ... + N_H} \neq \frac{t_1}{N_1} + \frac{t_2}{N_2} + ... + \frac{t_H}{N_H}$$

$$\boxed{\text{Strata means are not additive}}$$

However, we can formulate an additive relationship, by “weight” factors:

$$\bar{y} = W_1 \bar{y}_1 + W_2 \bar{y}_2 + ... + W_H \bar{y}_H$$

Where,

$$W_h = \frac{N_h}{N}$$

Note that,

$$\sum_{h=1}^{H} W_h = 1$$

Proof:

$$\begin{aligned}
\bar{y} &= W_1\bar{y}_1 + W_2\bar{y}_2 + ... + W_H\bar{y}_H \\
&= \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 + ... + \frac{N_H}{N}\bar{y}_H \\
&= \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 + ... + \frac{N_H}{N}\bar{y}_H \\
&= \frac{t_1}{N} + \frac{t_2}{N} + ... + \frac{t_H}{N} \\
&= \frac{t}{N}
\end{aligned}$$

An example

Two areas: N_A=10,000 and N_B=20,000;
So, N=30,000

$Mean_A$=(5,000/10,000)=0.5
$Mean_B$= (5,000/20,000)=0.25

Overall mean= (5,000+5,000)/(10,000+20,000)=0.33333

Then, W_A=(10,000/30,000)=1/3 and W_B = (20,000/30,000)=2/3

In STATA calculator: $Y=(W_A*Y_A+W_B*Y_B)$

di “overall mean” = (1/3)*0.5+(2/3)*0.25

. “overall mean” = .33333333

Variance Estimation of Stratified Sampling

1. An unbiased estimator of the population mean, μ of a variable Y is the stratified estimator of μ :

$$\overline{Y}_{str} = W_1\overline{Y}_1 + W_2\overline{Y}_2 + ... + W_H\overline{Y}_H$$

Where,

$$W_h = \frac{N_h}{N}$$

Its variance is:

$$\begin{aligned} Var(\overline{Y}_{str}) &= Var(W_1\overline{Y}_1) + Var(W_2\overline{Y}_2) + ... + Var(W_H\overline{Y}_H) \\ &= W_1^2 Var(\overline{Y}_1) + W_2^2 Var(\overline{Y}_2) + ... + W_H^2 Var(\overline{Y}_H) \\ &= \sum_{h=1}^{H} W_h^2 Var(\overline{Y}_h) \\ &= \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h}, \text{ under SRSWR} \\ &= \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h} \frac{N_h - n_h}{N_h - 1}, \text{ under SRSWOR} \end{aligned}$$

An unbiased estimator of the <u>proportion, P,</u> of population elements from stratified sampling is:

$$P_{str} = W_1 P_1 + W_2 P_2 + ... + W_H P_H$$
$$= \sum_{h=1}^{H} W_h P_h$$

$$Var(P_{str}) = \sum_{h=1}^{H} W_h^2 \frac{P_h (1 - P_h)}{n_h} \text{, } under\ SRSWR$$
$$= \sum_{h=1}^{H} W_h^2 \frac{P_h (1 - P_h)}{n_h} \frac{N_h - n_h}{N_h - 1} \text{, } under\ SRSWOR$$

An unbiased estimator of the total, t, of population elements from stratified sampling is:

$$Var(\hat{t}_{str}) = \sum_{h=1} Var(\hat{t}_h)$$

$$= \sum_{h=1} Var(N_h \bar{y}_h)$$

$$= \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}, \ under\ SRSWR$$

$$= \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h} \frac{N_h - n_h}{N_h}, \ under\ SRSWOR$$

Another method of estimating var(y_mean):

$$V\hat{a}r(\bar{y}_{str}) = Var\left(\frac{\hat{t}_{str}}{N}\right)$$

$$= \frac{1}{N^2}\sum_{h=1} Var(\hat{t}_h)$$

$$= \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}, \ under\ SRSWOR$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

$$= \sum_{h=1}^{H} W^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

Variance estimated under stratified sampling is always lower than the variance estimated under SRS.

This is best illustrated by considering that,

variance (total) = variance(within) + variance (between)

In case of stratified sampling, variance (between) = 0, i.e, all variance is due to variability within the strata.

And, because variance (between) < variance (total), stratified sampling variance is lower than that of SRS.

An example
4 groups (strata)

```
. ta group

      group |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        250       25.00       25.00
          1 |        250       25.00       50.00
          2 |        250       25.00       75.00
          3 |        250       25.00      100.00
------------+-----------------------------------
      Total |       1000      100.00
```

An example

```
bysort group: sum x

-------------------------------------------------------------------------------
-> group = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |       250    48.93032    28.93071   .0354402    99.5811

-------------------------------------------------------------------------------
-> group = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |       250    94.12133    59.57098   1.846363   199.6159

-------------------------------------------------------------------------------
-> group = 2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |       250    150.7658    85.04665   .1417242   299.6221

-------------------------------------------------------------------------------
-> group = 3

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |       250    192.7725    118.5134   3.255986   398.5283
```

Under SRS:

```
. *stddev of x: sqrt{variance(x)}

. sum x

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
       x |    1000    121.6475   96.89047   .0354402   398.5283

. *stderr of x: sqrt{variance(x)/n}

. ci x

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
       x |    1000     121.6475    3.063946         115.635      127.66
```

Under Stratified Sampling:

```
. *stderr of x under STRATIFIED SAMPLING
. Svymean x, str(group)

Survey mean estimation
pweight:  <none>                                  Number of obs    =      1000
Strata:   group                                   Number of strata =         4
PSU:      <observations>                          Number of PSUs   =      1000
                                                  Population size  =      1000
------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
---------+--------------------------------------------------------------------
       x |   121.6475    2.532984    116.6769    126.6181    .6834439
------------------------------------------------------------------------------
```

Why the variance/StdErr estimated under stratified sampling is lower than SRS?

```
loneway x group

                      One-way Analysis of Variance for x:
                                             Number of obs =      1000
                                                 R-squared =    0.3186
    Source                SS          df       MS             F     Prob > F
-------------------------------------------------------------------------
Between group          2988029.6       3    996009.86    155.24     0.0000
Within group           6390344.9     996     6416.009
-------------------------------------------------------------------------
Total                  9378374.5     999    9387.7623
```

Why the variance/StdErr estimated under stratified sampling is lower than SRS?

```
loneway x group

                      One-way Analysis of Variance for x:
                                              Number of obs =       1000
                                                  R-squared =     0.3186
    Source                    SS          df       MS              F      Prob > F
-------------------------------------------------------------------------
Between group           2988029.6       3     996009.86    155.24      0.0000
Within group            6390344.9     996      6416.009
-------------------------------------------------------------------------
Total                   9378374.5     999     9387.7623


* {var(between)+var(within)/n-1}/n
. disp ((2988029.57+6390344.95)/999)/1000
  9.3877623

. *stderr estimation
. disp sqrt(9.3877623)
  3.0639455
```

Variance under SRS

Standard error under SRS

stderr estimation under STRATIFIED SAMPLING

```
 One-way Analysis of Variance for x:
                                              Number of obs =      1000
                                                  R-squared =    0.3186
    Source                SS         df      MS            F     Prob > F
-------------------------------------------------------------------------
Between group          2988029.6      3    996009.86    155.24     0.0000
Within group           6390344.9    996     6416.009
-------------------------------------------------------------------------
Total
```

SE under stratified design:

```
. *dis sqrt(6416.00898/1000)
  2.5329842
```

Under Stratified Sampling

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
x	121.6475	2.532984	116.6769	126.6181	.6834439

Under SRS

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
x	1000	121.6475	3.063946	115.635	127.66

Design effect:

Deff = (variance under stratified sampling)/ (variance under SRS)

= 2.5329842^2/3.0639455^2 = .68344393

In stratified sampling it is assumed that “between variance”=0. Total variance under stratified sampling equals to “within variance” only.

Hence, variance from stratified sampling is always lower than under SRS.

Two Basic Rules of Stratified Sampling

- A minimum of two-elements must be chosen from each stratum so that sampling errors can be estimated for all strata independently.
- The population (elements) should be *homogenous* within stratum, and the population (elements) should be *heterogenous* between the strata.

First Rule: Minimum 2 elements in each stratum

- In Stata, the svy commands will not work if less than 2 elements are available in any strata.
- This is often a problem for sub-group analyses. A solution is to combine adjacent strata (you must have information about strata labels).

Second rule: population (elements) should be *homogenous* within stratum

- Suggests that "the gains in variance precision is greatest when the strata are maximally heterogenous *between*, but homogenous *within*.

- **variance(total) = variance(within) + variance(between)**

  fixed

variance(total) = variance(within) + variance (between)

- When the elements are homogenous (quite similar), there is less variance [*variance(within)* is smaller]
- Because *variance(between* strata)=0 in stratified sampling design, smaller the *variance(within)*, smaller the *total(variance)*.

variance(total) = variance(within) + 0

So, the objective is to increase *variance(between)* and decrease *variance(within)*.

Sample Size Estimation for Stratified Sampling Design

- Sample size estimation for stratified sampling is difficult in practice, not for the complexity of sample size formula.

- Sample size estimation depends on variance estimation. Consider the variance of a mean for a variable y:

Under SRS[WR]

$$
\begin{aligned}
V\hat{a}r(\bar{y}_{str}) &= Var\left(\frac{\hat{t}_{str}}{N}\right) \\
&= \frac{1}{N^2}\sum_{h=1} Var(\hat{t}_h) \\
&= \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}, \text{ under } SRSWOR \\
&= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h} \\
&= \sum_{h=1}^{H} W^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}
\end{aligned}
$$

Problem

- The variance estimation, even under "with replacement," needs information on additional three factors: *N*, N_h, s^2_h.

- It is very difficult or impossible to get information on s^2_h from each stratum.

$$V\hat{a}r(\bar{y}_{str}) = Var\left(\frac{\hat{t}_{str}}{N}\right)$$

$$= \frac{1}{N^2}\sum_{h=1} Var(\hat{t}_h)$$

$$= \sum_{h=1}^{H} \frac{N_h^2}{N^2}\frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}, \text{ under } SRSWOR$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

$$= \sum_{h=1}^{H} W^2 \frac{s_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

Sample Size Estimation for Stratified Sampling Design

- For those want to try!

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{(n_h / n)}}{N^2 \left(\frac{d^2}{Z^2_{\alpha/2}} \right) + \sum_{h=1}^{H} N_h S_h^2}$$

- Substitute $p_h(1-p_h)$ for binary outcomes (proportions).

- In practice, stratified sampling SS estimation is done under SRS assumption (more conservative) or preferably multi-stage sampling design method is used, and not done as a single stage sampling strategy.

Allocation of Stratified Sampling

The major task of stratified sampling design is the appropriate allocation of samples to different strata.

Types of allocation methods:

- Equal allocation
- Proportional to stratum size
- Allocation based on variance differences among the strata
- Cost based sample allocation

Equal Allocation

- Divide the number of sample units n equally among the K strata.
- Formula: $n_h = n/K$
- Example: n = 100; 4 strata; sample n_h=100/4 = 25 in each stratum.
- May not be equal in each stratum. (what if you have 3 strata?)
- Need “weighted analysis” (disproportionate selection)

Proportional allocation

- Make the proportion of each stratum sampled identical to the proportion of the population.

- Formula: Let the sample fraction f= n/N.

  So, $n_h = fN_h = n(N_h/N) = nW_h$,

  Where $W_h = N_h/N$ is the stratum weight.

- Note, *f* is constant across strata, but W_h varies among strata.
- Self-weighted (equal proportion from each stratum)

Proportional allocation

Example:

- $N = 1000$
- $n = 100$
- $f = n/N=100/100=.1$
- $N_1 = 700$ $n_1=fN_1=0.1*700=70$
- $N_2 = 300$ $n_2=fN_2=0.1*300=30$

Disadvantages

- A major disadvantage of proportional allocation:
  - Sample size in a stratum may be low – provide unreliable stratum-specific results.
- A major disadvantages of equal allocation:
  - May need to use weighting to have unbiased estimates.

Optimal allocation (Neyman Allocation)

Based on the variability of sampling: more variable strata should be sampled more intensely.

Formula:

$$n_h = n\left(\frac{N_h S_h}{\sum_{k=1}^{H} N_h S_h}\right)$$

- Need “weighted analysis” (disproportionate selection)

Drawing stratified random samples

Stata implementation (from a list):

```
.ta area
    type of |
       area |      Freq.     Percent        Cum.
------------+-----------------------------------
major urban |        343        7.11        7.11
other urban |      1,024       21.23       28.34
      rural |      3,457       71.66      100.00
------------+-----------------------------------
      Total |      4,824      100.00
```

Equal allocation

```
. sample 200, count by(area)
(4224 observations deleted)
. ta area
     type of |
        area |      Freq.     Percent        Cum.
-------------+-----------------------------------
major urban |        200       33.33       33.33
other urban |        200       33.33       66.67
       rural |        200       33.33      100.00
-------------+-----------------------------------
       Total |        600      100.00
```

Proportional allocation

```
. sample 20, by(area)
(3859 observations deleted)
. ta area
    type of |
       area |      Freq.     Percent        Cum.
------------+-----------------------------------
major urban |         69        7.15        7.15
other urban |        205       21.24       28.39
      rural |        691       71.61      100.00
------------+-----------------------------------
      Total |        965      100.00
```

Proportional allocation

SS may not be adequate for stratum specific analysis

```
. sample 20, by(area)
(3859 observations deleted)
. ta area
    type of |
       area |      Freq.     Percent        Cum.
------------+-----------------------------------
major urban |         69        7.15        7.15
other urban |        205       21.24       28.39
      rural |        691       71.61      100.00
------------+-----------------------------------
      Total |        965      100.00
```

Probability Proportional to Size (PPS)

- PPS is very common in large surveys.
- In simplistic sense, the selection probability that a particular sampling unit will be selected in the sample is proportional to the size of the variable of interest (e.g., in a population survey, the population *size* of the sampling unit).
- **PPS sampling provides self-weighted samples.**

Sample selection probabilities at area levels

Area	# HH	Probability of any HH selected
1	5,000	1/5000 = 0.0002
2	20,000	1/20000 = 0.00005
3	3,000	1/3,000 = 0.00033333
4	10,000	1/10000 = .0001

Use of PPS

- when the populations of the sampling units vary, and
- to ensure that every element in the target population has an equal chance of being included in the sample (self weighted).

Steps in PPS Sampling:

- Creating a list of clusters with cumulative population size
- Selecting a systematic sample from a random start using a sampling interval,
- Please see the handout for an example

Step #2: *Systematic selection from the list*

Step #1

Area	# women (15-44)	Cumulative number	Range
1	5,000	5,000	0 –5,000
2	20,000	25,000	5,001-25,000
3	3,000	28,000	25,001-28,000
4	10,000	38,000	28,001-38,000
5	18,000	56,000……	38,001-56,000
		……	
10		75,0000	

Some practical considerations

- Conceptually, quite similar to systematic sampling
- PPS is very attractive in practice because no weighting is required
- However, due to other reasons (missing responses), weighting may not be avoided.