STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 4 - Part II: Stratified Random Sampling

Ramya Thinniyam

May 27, 2014

When should you use a SRS?

Simple Random Samples are the easy to design and analyze, but may not be appropriate in some cases.

Use a SRS when:

- Little/no extra information is available about characteristics in the population
- Data users insist on SRS formulas: averaging sample values
- Main interest is multivariate relationships (regression equations) for the population: easier to perform and interpret for SRSs

Do NOT use a SRS when:

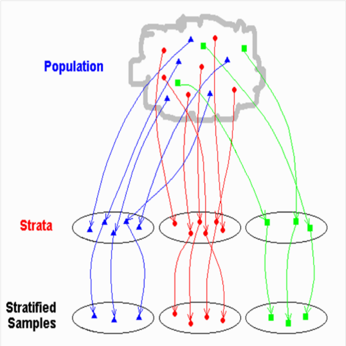
- A controlled experiment is appropriate (not a survey sample)
- List of observation units in population is not available or too expensive/time consuming to take SRS
- You have additional information about population characteristics that can improve survey design / cost effective design

Stratified Random Sampling

Recall that in Stratified Random Sampling (STRS):

1. Population split into L distinct strata / groups:

   Strata should partition the population (should not overlap and should comprise of whole population so that each sampling unit belongs to exactly one stratum).
2. Take independent probability samples (SRS) from each stratum
3. Pool information to get overall population parameters



Why choose a STRS?

- Can obtain more representative sample than SRS
- Elements homogeneous within strata:

  Smaller variances / more precise estimates
  $\longrightarrow$ narrower CIs Think about "age groups"
- Cost most likely lower, more convenient, easier to administer than SRS:

  Can use different sampling procedures for different strata
- May be interested in estimates within subpopulations with known precision:

  Choose subpopulations as strata, sample according to population proportions or depending on precision

Examples: How would you stratify in each case?

(1) A study about blood pressure ages/gender/BMI/...

(2) A study about concentration of plants in an area rainfall/temperature/size

(3) Political Survey minority groups ...

(4) Absences of Primary School Children Example age/grade ...

(5) A study on salaries of university instructors faculty/ ...

Theory and Notation for STRS

- Divide population of size N into L strata with N_i sampling units in stratum i
- $N_1, N_2, \ldots, N_{L-1}, N_L$ population sizes known and $N = \sum_{i=1}^{L} N_i$
- Take SRS of size n_i from each stratum, denoted $\mathcal{S}_i$
- Total sample size: $n = \sum_{i=1}^{L} n_i$
- $i = 1, \ldots, L$: index for strata
- $j = 1, \ldots, N_i$: index for elements within stratum i

Population parameters are:

- y_{ij} : variable/measurement value of jth unit in stratum i
- $\tau_i = \sum_{j=1}^{N_i} y_{ij}$: Population total in stratum i
- $\tau = \sum_{i=1}^{L} \tau_i$: Population total (overall)
- $\bar{y}_{iU} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$: Population mean in stratum i
- $\bar{y}_U = \frac{\tau}{N} = \frac{\sum_{i=1}^{L} \sum_{j=1}^{N_i} y_{ij}}{N}$: Population mean (overall)
- $S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{iU})^2$: Population variance within stratum i
- $S^2 = \frac{1}{N-1} \sum_{i=1}^{L} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_U)^2$: Population variance (overall) - may not be useful!

Sample Quantities / Estimators

Use SRS estimators within each stratum to obtain:

- $\bar{y}_i = \frac{1}{n_i} \sum_{j \in \mathcal{S}_i} y_{ij}$: estimates $\bar{y}_{iU}$
- $\hat{\tau}_i = \frac{N_i}{n_i} \sum_{j \in \mathcal{S}_i} y_{ij} = N_i \bar{y}_i$: estimates τ_i
- $s_i^2 = \frac{1}{n_i - 1} \sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2$: estimates S_i^2
- $\hat{\tau}_{st} = \sum_{i=1}^{L} \hat{\tau}_i = \sum_{i=1}^{L} N_i \bar{y}_i$: estimates τ
- $\bar{y}_{st} = \frac{\hat{\tau}_{st}}{N} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$: estimates $\bar{y}_U$

  $\hookrightarrow$ Weighted average of sample stratum averages, weights are proportions of population units in each stratum.

weight: $\frac{N_i}{N}$

- Must know sizes or relative sizes of strata to use STRS

Properties of Estimators

- Unbiasedness:

  $\bar{y}_{st}$ is unbiased for $\bar{y}_U$ and $\hat{\tau}_{st}$ is unbiased for τ

- Variances:

  $$V(\bar{y}_{st}) = \sum_{i=1}^{L} \left(1 - \frac{n_i}{N_i}\right) \left(\frac{N_i}{N}\right)^2 \frac{S_i^2}{n_i}$$

  $$V(\hat{\tau}_{st}) = \sum_{i=1}^{L} \left(1 - \frac{n_i}{N_i}\right) N_i^2 \frac{S_i^2}{n_i}$$

- Standard Errors:

  In order to estimate variances, we need to sample at least 2 units from each stratum

  O.W. 1 unit $\Rightarrow$ Variance $= 0$

  $(1 - \frac{1}{1}) = 0$ ✓

  $$SE(\bar{y}_{st}) = \sqrt{\sum_{i=1}^{L} \left(1 - \frac{n_i}{N_i}\right) \left(\frac{N_i}{N}\right)^2 \frac{s_i^2}{n_i}}$$

  $$SE(\hat{\tau}_{st}) = \sqrt{\sum_{i=1}^{L} \left(1 - \frac{n_i}{N_i}\right) N_i^2 \frac{s_i^2}{n_i}}$$

Proofs of Properties

Remember:

- Properties of SRS estimators
- Properties of expectations and variances
- Independence when sampling from strata

Recall: $\bar{y}_{st} = \sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i$

since an SRS is taken from each stratum i, $E(\bar{y}_i) = \bar{y}_{iU}$

so, $E(\bar{y}_{st}) = \sum_{i=1}^{L} \frac{N_i}{N} E(\bar{y}_i) = \frac{1}{N}(\sum_{i=1}^{L} N_i \bar{y}_{iU}) = \frac{T}{N} = \bar{y}_U$

since an SRS is taken from each stratum i, $Var(\bar{y}_i) = (1-\frac{n_i}{N_i})\frac{S_i^2}{n_i}$

So $Var(\bar{y}_{st}) = Var(\sum_{i=1}^{L} \frac{N_i}{N} \bar{y}_i) = \sum_{i=1}^{L} Var(\frac{N_i}{N} \bar{y}_i)$ since each stratum indep from each others

$= \sum_{i=1}^{L} \frac{N_i^2}{N^2} Var(\bar{y}_i) = \sum_{i=1}^{L} (\frac{N_i}{N})^2 (1-\frac{n_i}{N_i})\frac{S_i^2}{n_i}$

Since $\hat{T}_{st} = N\bar{y}_{st}$, $E(\hat{T}_{st}) = N\, E(\bar{y}_{st}) = N\bar{y}_U = T$

$Var(\hat{T}_{st}) = N^2 Var(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2 \frac{S_i^2}{n_i}(1-\frac{n_i}{N})$

Confidence Intervals

If either:

(1) Sample sizes within each stratum are large OR

(2) Large number of strata

Then,

An approximate $100(1-\alpha)\%$ CI for the population mean, $\bar{y}_U$ is:

$$\bar{y}_{st} \pm z_{\alpha/2} SE(\bar{y}_{st})$$

An approximate $100(1-\alpha)\%$ CI for the population total, τ is:

$$\hat{\tau}_{st} \pm z_{\alpha/2} SE(\hat{\tau}_{st})$$

* Note: Some software use t_{n-L} critical values rather than standard normal *

Installing Sampling Contributed Package in ‘R’

1. Open R
2. Be sure you are connected to the internet
3. At the top of the R window click on **Packages**
4. A list will open, click on **Install Packages**
5. A list of mirror sites appears. Select **Canada (ON)**, and click **OK**
6. Another list will open, click on **Sampling** and then click **OK**
7. A lot of information will appear on the screen, but at the end you will get the R prompt $>$
8. Again click **Packages**, then click **Load Package**, select **Sampling** and **click OK**

Example: Using R for Stratified Sampling

Groups A,B,C,D and one variable (response)

> attach(strsex)

Read the data into R:

```
> strsex<-read.csv("strsex.csv")
> strsex
   response group
1       8.8     A
2      10.6     A
3      10.6     A
4       7.6     A
5       7.7     A
6      10.0     A
. . .
75      8.3     D
76     12.3     D
77      9.4     D
78      7.9     D
79      6.9     D
80     11.2     D
```

Find population mean and total:

```
> mean(strsex$response)
[1] 9.8325
> sum(strsex$response)
[1] 786.6
> sum(strsex$response)/length(strsex$response)
[1] 9.8325
```

Take a STRS:

```
> strs.sample<-strata(strsex,c("group"),size=c(3,4,5,6),method=c("srswor"))
> strs.sample
   group ID_unit Prob Stratum
1      A      12 0.15       1
2      A      14 0.15       1
3      A      19 0.15       1
4      B      23 0.20       2
5      B      30 0.20       2
6      B      33 0.20       2
7      B      36 0.20       2
8      C      43 0.25       3
9      C      46 0.25       3
10     C      48 0.25       3
11     C      49 0.25       3
12     C      51 0.25       3
13     D      64 0.30       4
14     D      67 0.30       4
15     D      68 0.30       4
16     D      69 0.30       4
17     D      75 0.30       4
18     D      77 0.30       4
```

Look at STRS data:

```
> strs.sample.data<-getdata(strsex,strs.sample)
> strs.sample.data
   response group ID_unit Prob Stratum
1       9.3     A      12 0.15       1
2       9.4     A      14 0.15       1
3      13.2     A      19 0.15       1
4      11.1     B      23 0.20       2
5       8.4     B      30 0.20       2
6      10.2     B      33 0.20       2
7      10.1     B      36 0.20       2
8      10.5     C      43 0.25       3
9       7.7     C      46 0.25       3
10      7.9     C      48 0.25       3
11     10.3     C      49 0.25       3
12      7.5     C      51 0.25       3
13      7.5     D      64 0.30       4
14     11.8     D      67 0.30       4
15      6.1     D      68 0.30       4
16      9.2     D      69 0.30       4
17      8.3     D      75 0.30       4
18      9.4     D      77 0.30       4
```

Calculate N_i , n_i, $\bar{y}_i$, s_i^2 for each stratum:

```
> Ni<-tapply(strsex$response,strsex$group,length)
> ni<-tapply(strs.sample.data$response,strs.sample.data$group,length)
> ssqi<-tapply(strs.sample.data$response,strs.sample.data$group,var)
> ybari<-tapply(strs.sample.data$response,strs.sample.data$group,mean)
> Ni
 A  B  C  D
20 20 20 20
> ni
A B C D
3 4 5 6
> ssqi
       A        B        C        D
4.943333 1.270000 2.212000 3.741667
> ybari
        A         B         C         D
10.633333  9.950000  8.780000  8.716667
```

Population size:

```
> N = length(strsex$response)
> N
[1] 80
```

Calculate $\bar{y}_{st}$:

```
> ybar.st<-sum(Ni*ybari)/N
> ybar.st
[1] 9.52
```

Calculate $\hat{V}(\bar{y}_{st})$:

```
> var.ybar.st<-sum(Ni^2*(1-ni/Ni)*ssqi/ni)/N^2
> var.ybar.st
[1] 0.1514337
```

Calculate $\hat{\tau}_{st}$:

```
> N*ybar.st
[1] 761.6
```

Calculate $\hat{V}(\hat{\tau}_{st})$:

```
> N^2*var.ybar.st
[1] 969.1755
```

Example: Confidence Intervals

a) Use the R output and data to find a 95% CI the population mean and a 95% CI for the population total (assuming the required assumptions are met).

$\bar{y}_{st} \pm 1.96\sqrt{\hat{Var}(\bar{y}_{st})}$

$= 9.52 \pm 1.96\sqrt{0.1514} = (8.7574, 10.2826)$

$N=80$

$\bar{y}_{st} = 9.52$

$\hat{Var}(\bar{y}_{st}) = 0.1514$

true value $\bar{y}_u = 9.8325$

contained in CI good ✓

Since $\hat{t}_{st} = N\bar{y}_{st}$ $N=80$

95% CI for $\hat{t}$ is $(N\times 8.7574, N\times 10.2826)$

$= (700.592, 822.608)$

b) Use the R output to find a 95% CI for the mean of group D (assuming the required assumptions are met).

Actually using SRS: 95% CI for $\bar{y}_{Du}$: $n_D=6$

from group D $N_D=20$

$\bar{y}_D \pm 1.96\sqrt{\left(1-\frac{n_D}{N_D}\right)\frac{s_D^2}{n_D}} = 8.7167 \pm 1.96\sqrt{\left(1-\frac{6}{20}\right)\frac{3.7417}{6}} = (7.4217, 10.0117)$

Stratified Sampling for Proportions

Recall that proportions are simply means of indicator vairables.

Use: $\hat{p}_i = \bar{y}_i$ and $s_i^2 = \frac{n_i}{n_i - 1}\hat{p}_i(1 - \hat{p}_i)$.

$$\boxed{\hat{p}_{st} = \sum_{i=1}^{L} \frac{N_i}{N}\hat{p}_i}$$

$$\boxed{\hat{V}(\hat{p}_{st}) = \sum_{i=1}^{L} \left(1 - \frac{n_i}{N_i}\right)\left(\frac{N_i}{N}\right)^2 \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i - 1}}$$

An approximate $100(1 - \alpha)\%$ CI for the proportion, p is:

$$\hat{p}_{st} \pm z_{\alpha/2} SE(\hat{p}_{st})$$

Estimating Total Number of Population Units with a Characteristic

$\hat{\tau}_{st} = \sum_{i=1}^{L} N_i \hat{p}_i$

i.e. the estimated total number of population units with the characteristic = sum of the estimated totals in each stratum

$\hat{V}(\hat{\tau}_{st}) = N^2 \hat{V}(\hat{p}_{st})$

An approximate $100(1-\alpha)\%$ CI for the population total, τ is:

$$\hat{\tau}_{st} \pm z_{\alpha/2} SE(\hat{\tau}_{st})$$

Example: Television Advertising

An advertising firm is interested in estimating the proportion of households in a certain county that watch TV show ‘X’, in order to target their advertising more efficiently. The county has two towns, A and B, and a rural area - Town A is built around a factory and most households contain factory workers with school-age children, while Town B contains mostly elderly residents with few children at home.

Location	Population Size	Sample Size	# of households viewing show ‘X’
Town A	155	20	16
Town B	62	8	2
Rural	93	12	6

a) Discuss the merits of using STRS in this case.

b) Estimate the proportion of households in this county that view ‘X’ and place a bound on the error of the estimation (based on 95% confidence).

Sampling Weights

$\pi_{ij} = \frac{n_i}{N_i}$, so the sampling weights are:

$$w_{ij} = \frac{1}{\pi_{ij}} = \frac{N_i}{n_i}$$

- sampling weight interpreted as the number of units in the population represented by the sample member y_{ij} : each sampled unit in stratum i represents itself $+ \left(\frac{N_i}{n_i} - 1\right)$ other units in stratum i that were not selected in the sample
- sum of the weights is N
- $$\hat{\tau}_{st} = \sum_{i=1}^{L} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} \quad \text{and} \quad \bar{y}_{st} = \frac{\sum_{i=1}^{L} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i=1}^{L} \sum_{j \in \mathcal{S}_i} w_{ij}}$$
- STRS is self-weighting if the sampling fraction $\frac{n_i}{N_i}$ is the same for each stratum (i.e. sampling weight is $\frac{N}{n}$ like for SRS. But variance depends on stratification - weights do not tell you the stratum membership of observations)