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

Lecture 5: Stratified Random Sampling (cont'd)

Ramya Thinniyam

May 29, 2014

Review: STRS Theory and Notation

- 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!

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.

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

Stratified Sampling for Proportions

Recall that proportions are simply means of indicator vairables.

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

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

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

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).

a) advantages of STRS:
We use geographic location/township as strata
→ each location should have similar patterns among residents for TV viewing.
→ convenience for administering/collecting data.
→ can allow estimation for each location separatedly.

b). Estimate p with a bound on error.
Let p=true proportion of households in the county that view TV show "X".

A.(1) $N_1=155$, $n_1=20$, $\sum y_i=16=n_1\hat{p}_1$
B.(2) $N_2=62$, $n_2=8$, $\sum y_i=2=n_2\hat{p}_2$
R.(3) $N_3=93$, $n_3=12$, $\sum y_i=6=n_3\hat{p}_3$

So $\hat{p}_1=0.8$, $\hat{p}_2=0.25$, $\hat{p}_3=0.5$

$$\hat{p}_{st}=\frac{1}{N}\sum_{i=1}^{3}N_i\hat{p}_i=\frac{1}{310}(155\times0.8+62\times0.25+93\times0.5)=0.6$$

$$\hat{V}(\hat{p}_{st})=\frac{1}{N^2}\sum_{i=1}^{4}\left[\left(1-\frac{n_i}{N_i}\right)N_i^2\frac{\hat{p}_i(1-\hat{p}_i)}{n_i-1}\right]=0.0045$$

95% CI: error $e=1.96\sqrt{\hat{V}(\hat{p}_{st})}=1.96\sqrt{0.0045}=0.1315$

p is estimated as $\hat{p}_{st}\pm e=0.6\pm0.1315$

Sampling Weights

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

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

- 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 (total)
- $$\boxed{\hat{\tau}_{st} = \sum_{i=1}^{L} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} \quad \text{and} \quad \bar{y}_{st} = \frac{\sum_{i=1}^{L} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i=1}^{L} \sum_{j \in \mathcal{S}_i} w_{ij}}}$$
- 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)

ratio same

$\Downarrow$

weight same

SRS is self-weighting

b/c in this case, weights are the same

"efficient" <=> low Variance

Analysis of Variance (ANOVA)

STRS is most efficient when:

- The observations are homogenous within each strata and heterogenous between strata
- ☆ Stratum means differ widely so that the variation amongst strata is high and the variation within each stratum is small.

Regression with response $= y_i$

Factor $=$ Stratum

with L Factor level

$$Y_i = \beta_0 + \beta_1 I(i \in Str_1) + \beta_2 I(i \in Str_2) + \cdots + \beta_{L-1} I(i \in Str_{L-1}) + \underbrace{\varepsilon}_{\text{error term}}$$

$i = 1, \cdots, N$ (every member of population)

ANOVA Table for Population:

Source	df	Sum of Squares
Between Strata	$L-1$	$SSB = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(\bar{y}_{iU} - \bar{y}_U)^2 = \sum_{i=1}^{L} N_i(\bar{y}_{iU} - \bar{y}_U)^2$
Within Strata	$N-L$	$SSW = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_{iU})^2 = \sum_{i=1}^{L}(N_i - 1)S_i^2$
Total (about $\bar{y}_U$)	$N-1$	$SSTO = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_U)^2 = (N-1)S^2$

Good: SSB $\uparrow$

SSW $\downarrow$

Allocating Observations to Strata

- Allocation: How to determine the number of observations to sample in each stratum
- Allocation depends on 3 factors:
  1. The total number of elements in each stratum
  2. The variability of observations within each stratum
  3. The cost of obtaining an observation from each stratum

     money / time
- Allocation Schemes:
  1. Proportional Allocation
  2. Optimal Allocation
  3. Neyman Allocation

Allocation Schemes

1) Proportional Allocation:

- number of sampled units in each stratum is proportional to size of stratum in population
- Ex. Population with 2400 men and 1600 women: a proportional allocation with a 10% sample means you would sample 240 men and 160 women
- proportional allocation ensures that sample reflects population wrt stratification variable and sample is a mini version of population
- $\pi_{ij} = \frac{n}{N}$ for all strata → self-weighting sample b/c $\frac{240}{2400} = \frac{160}{1600}$
- when strata are large enough, $V_{prop}(\bar{y}_{st}) \leq V_{SRS}(\bar{y})$ with same sample size, regardless of stratification scheme

showing the advantage of using STRS (proportional allocation)

proof of the efficiency of STRS comparing with SRS needs ANOVA (HW)

2) Optimal Allocation:

- allocate sampling units to strata so variance of estimator is minimized for a given total cost
- proportional allocation is best for increasing precision if variances S_i^2 are approximately equal across all strata
- optimal allocation will result in smaller costs when S_i^2 differ a lot (b/c you want to sample heavily on those with higher variations.
- expect larger variation among larger strata so sample higher percentage of them
- optimal allocation works well when sampling units vary in size and some strata are more expensive to sample than others
- objective: gain most information with least cost using a "cost function" :

$$C = c_0 + \sum_{i=1}^{L} c_i n_i$$

Minimizing Cost: $C = c_0 + \sum_{i=1}^{L} c_i n_i$

- minimize $V(\bar{y}_{st})$ for a given total cost, C : minimize C for a fixed $V(\bar{y}_{st})$
- $\rightarrow n_i \propto \frac{N_i S_i}{\sqrt{c_i}}$
- $$n_i = \left(\frac{\frac{N_i S_i}{\sqrt{c_i}}}{\sum_{\ell=1}^{L} \frac{N_\ell S_\ell}{\sqrt{c_\ell}}} \right) n$$
- sample heavily from a stratum if:
  - • stratum accounts for large part of population
  - • variance within stratum is large - sample heavily to compensate for heterogeneity
  - • sampling from stratum is inexpensive
- Note: if formula gives an optimal $n_i > N_i$, take sample of size N_i for that stratum and apply formula again for remaining strata

3) Neyman Allocation:

NOT variances

- special case of optimal allocation when costs in each strata (not variances) are approximately equal
- $n_i \propto N_i S_i$
- if variances S_i^2 specified correctly, Neyman allocation gives an estimator with smaller variance than proportional allocation

$V_{Neyman}(\bar{y}_{st}) < V_{prop}(\bar{y}_{st})$

Allocation for Specified Precision within Strata:

- interested in comparing between strata (rather than precision of estimate for entire population)
- determine sample size using sample size calculations from SRS (Lecture 4 - Part I)

Comparing Methods of Allocation

- if all variances and costs are equal, proportional allocation is same as optimal allocation (& Neyman)
- if variances within each stratum are known and differ, optimal allocation gives smaller variance for estimator of $\bar{y}_U$ than proportional allocation
- optimal allocation more complicated
- proportional allocation simpler and has self-weighting property - often worth the extra variance
- optimal allocation will differ for each variable measured whereas proportional allocation will not (depends only on sizes of strata in population)
- proportional allocation almost always has smaller variance than SRS

Determining Sample Size, n

(overall)

allocation is about partially sample size.

Allocation methods determine the relative sample sizes, $\frac{n_i}{n}$

Need to construct strata, allocate observations to strata, then determine sample size necessary for a specified margin of error, e.

Recall:

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

Following sample size formulas from before, we obtain:

$$n = \frac{z_{\alpha/2}^2 v}{e^2}$$

Problems with STRS

- May be hard to get sampling frames within strata: stratum membership of unit may only be available after sampling (post-stratification)
- May select strata that don't have homogeneous populations
- May not always reduce SRS variances by using STRS, since estimated variances are not weighted average of strata variances
- If more than one response variable is measured:
  - What variable do we stratify by?
  - What variance estimate do we use in Neyman allocation?