Stratified Sampling

Lecture 6

Lecture 6: Stratified Sampling

Reading: Lohr Chapter 3, sections 1-5

- Definitions and Notation
- Why stratify?
- Bias and Variance
- Sample allocation

Motivating Example

Goal: Estimate the average income of OSU graduate students one year past graduation.
How? SRS of graduated students. Ask each alum their salary.

- **target population:**
- **observation unit:**
- **sampling frame:**
- **sampling unit:**
- **sampled population:**

Motivating Example, cont.

$\bar{y}$ = sample average

$$V[\bar{y}] = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$$

Is there a sampling plan that would get us a better estimate without increasing the sample size?

Suggest some strata?

- Major
- Masters/ Doctorate
- Gender
- Current Geography.

Stratified Random Sampling

To estimate a population average:

1. Divide the sampling frame into groups ($\underbrace{\text{strata}}_{\text{plural}}$)
2. Conduct a SRS within each group
3. Estimate the average for each group ($\underbrace{\text{stratum}}_{\text{singular}}$)
4. Take a weighted average of the group averages

$\hookrightarrow$ estimate population mean

SRS

Strata

Strata:

1. are a **partition** of the population
   - Non-overlapping
   - Constitute the whole population
2. Must be defined (partitioned) **before** sampling

Stratification variable **should** be related to the variable(s) of interest.

Population Notation

There are H strata, indexed $h = 1, \ldots, H$.
N_h is the population size of stratum h.

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

Because Strata are a partition.

y_{hj} is the value associated with unit j of stratum h.
$j = 1, \ldots, N_h$

Within each stratum, we have the same estimands as before:

$$p_h = \frac{\#\text{ units w/ characteristic in stratum } h}{N_h}$$

Stratum (pop.) mean

$$\bar{y}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$$

total

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

Estimands

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

$$\bar{y}_{\mathcal{U}} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

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

Within-Stratum Estimates

Let n_h denote the sample size for stratum h.

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

Estimates within each stratum are the same as for SRS:

$$\bar{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} \qquad \hat{p}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} x_{hj} \qquad \hat{t}_h = N_h \bar{y}_h$$

Population Estimates

$$\hat{\bar{y}}_h = \frac{1}{n_h} \sum_{j \in S_h} y_{hj} \qquad \hat{p}_h = \frac{1}{n_h} \sum_{j \in S_h} x_{hj} \qquad \hat{t}_h = N_h \bar{y}_h$$

estimates of pop values.

$$\bar{y}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h \qquad \hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h \qquad \hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h$$

Weights

What is the individual weight for each unit of the population?

Recall $w_{hj} = \dfrac{1}{\text{Probability person } hj \text{ is selected}} = \dfrac{1}{\pi_{hj}}$

$$\pi_{hj} = \frac{n_h}{N_h} \Rightarrow w_{hj} = \frac{N_h}{n_h}.$$

$$\bar{y}_{HT} = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i} = \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj}}$$

$$= \frac{\sum_h \sum_j \frac{N_h}{n_h} y_{hj}}{\sum_h \sum_j \frac{N_h}{n_h}} = \frac{\sum_h N_h \underbrace{\frac{1}{n_h} \sum_j y_{hj}}_{\hat{\bar{y}}_h}}{\sum_h N_h \frac{1}{\cancel{n_h}} \underbrace{\sum_j 1}_{= \cancel{n_h}}} = \frac{\sum_h N_h \hat{\bar{y}}_h}{N} = \bar{y}_{str}.$$

(w/o loss of generality, reorder the population s.t. 1st n_h units in each stratum h are sampled)

Bias?

$$E\left[\bar{y}_{str}\right] = E\left[\sum_{h=1}^{H} \frac{N_h}{N} \hat{\bar{y}}_h\right] = \sum_{h=1}^{H} \frac{N_h}{N} E\left[\hat{\bar{y}}_h\right]$$

$$= \sum_{h=1}^{H} \frac{N_h}{N} \underbrace{\bar{y}_{h\mathcal{U}}}_{}.$$

population average for stratum h.

$$= \bar{y}_{\mathcal{U}}.$$

Variance

$$V\left[\bar{y}_{str}\right] = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}$$

$$V\left[\sum_{h=1}^{H} \frac{N_h}{N}\hat{\bar{y}}_h\right],$$

is $\hat{\bar{y}}_k$ independent from $\hat{\bar{y}}_h$ for $k \neq h$ Yes!

$$= \sum_{h=1}^{H} V\left[\frac{N_h}{N}\hat{\bar{y}}_h\right]$$

$$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 V\left[\hat{\bar{y}}_h\right]$$

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

Stratified better than SRS?

When is $V(\bar{y}_{str}) < V(\bar{y})$?

$$\sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \overset{?}{<} \left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

assume $\frac{n_h}{N_h} = \frac{n}{N}$ for all h. ←

$$\sum_{h=1}^{H} \frac{N_h}{N} S_h^2 \overset{?}{<} S^2$$

$$S^2 = \frac{1}{N-1}\sum (y_i - \bar{y}_u)^2$$

assume $\frac{N-1}{N} \approx \frac{N_h - 1}{N_h} \approx 1$ ←

$$N = \sum N_h$$

$$N-1 \neq \sum (N_h - 1)$$

$$\cancel{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2} \overset{?}{<} \cancel{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2} + \sum_{h=1}^{H} \frac{N_h}{N}(\bar{y}_h - \bar{y}_u)^2$$

EXAMPLE

Suppose this is the population of interest. We would like to estimate the average perimeter.
Two possible stratification schemes:

- stratify by color
- stratify by size (3 squares, 4 squares, etc.)

Advantages

1. Reduce variability of estimates
2. Convenient to administer
3. Protect yourself against a “bad” sample
4. Obtain estimates to specified precision for subgroups.

$$\sum_h \left(\frac{N_h}{N}\right)^2 \left(1-\frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}$$

Allocation

1. Allocate to give specified margins of error for each stratum

2. Proportional allocation

assume

2

$$n_h = \underbrace{n}_{\text{total sample}} \underbrace{\left(\frac{N_h}{N}\right)}_{\text{Proportion of pop in stratum } h.}$$

plug into variance formula for $V[\bar{y}_{str}]$

$$V\left[\bar{y}_{str}\right] = \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2,$$

Solve $e = z_{\alpha/2}\sqrt{V[\bar{y}_{str}]}$

1

$$\boxed{n = \frac{z_{\alpha/2}^2 S^{*2}}{e^2 + \frac{z_{\alpha/2}^2 S^{*2}}{N}}, \quad S^{*2} = \sum_{h=1}^{H}\frac{N_h}{N}S_h^2}$$

Allocation

$\frac{S_h}{n_h}.$

3. Neyman allocation: Minimize total variability by allocating more sample to strata with larger variances:

2

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

This is the function $g(S_h, N_h)$ that minimizes $V[\bar{y}_{str}]$

plug in

$$V\left[\bar{y}_{str}\right] = \frac{1}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{H} \frac{N_h}{N} S_h^2$$

solve for

$e = z_{\alpha/2}\sqrt{V[\bar{y}_{str}]}$

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

1

Allocation

administrative

$$C = c_0 + \sum_{h=1}^{H} c_h n_h .$$

$\Longrightarrow$ Minimize $V[\bar{y}_{str}]$ subject to C using Lagrange Multipliers

4. Optimal allocation

produces this allocation scheme.

$$n_h^* = n\left(\frac{N_h S_h/\sqrt{c_h}}{\sum_{\ell=1}^{H} N_\ell S_\ell/\sqrt{c_\ell}}\right)$$

actually solve this equation with the optimal allocation formula * substituted for n_h.

plug into $V[\bar{y}_{str}]$ and solve for $e = z_{\alpha/2}\sqrt{V(\bar{y}_{str})}$

$$n = \frac{z_{\alpha/2}^2\left(\sum_{h=1}^{H}\frac{N_h}{N}S_h/\sqrt{c_h}\right)^2}{e^2 + \frac{z_{\alpha/2}^2}{N}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_h^2\right)} = \text{sample size required to get MOE } e \text{ at minimum cost.}$$

$$V\left[\bar{y}_{str}\right] = \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_h/\sqrt{c_h}\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2$$

Practical Issues.

1. Must estimate variance/cost to use Neyman/optimal
2. Use SRS to estimate n, use allocation formulas from there. (Generally, this is conservative).
3. Use SRS to estimate n. Discount this value some. Use allocation formulas.

design effects = deff = $\dfrac{V[\bar{y}_{str}]}{V[\bar{y}_{SRS}]}$

$$e/\sqrt{deff} = z_{\alpha/2}\sqrt{V[\bar{y}_{SRS}]}$$

Gender

Undergrad/Grad