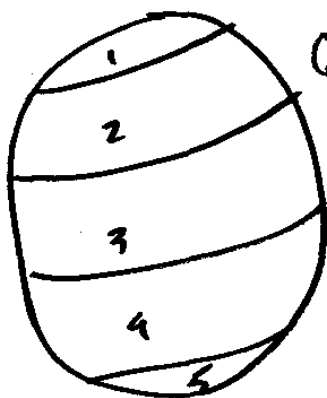


Stratified sampling

Stat 576

10-17-17

①



Get SRSWOR within each stratum

Optimally, we want homogeneity within,
heterogeneity between

- 4 big questions:
1. How many strata?
 2. What should the boundaries be?
Demarcation
 3. How large should n be?
 4. What portion of the sample should come from each stratum? Allocation

②

Notation:

H = # of strata

y_{hj} = j^{th} item sampled in stratum h

N = population size

N_h = " " of stratum h

n = sample size

n_h = " " from stratum h

\bar{Y} = population mean

\bar{Y}_h = " " of stratum h

S_h^2 = modified population variance for stratum h

(3)

t_h = population total for stratum h
 $= N_h \bar{Y}_h$

\bar{y}_h = sample mean for stratum h

S_h^2 = " variance " " " "

Know: $E[\bar{y}_h] = \bar{Y}_h$, $V[\bar{y}_h] = \frac{S_h^2}{n_h} (1 - \frac{n_h}{N_h})$,

$$\hat{V}[\bar{y}_h] = \frac{s_h^2}{n_h} (1 - \frac{n_h}{N_h})$$

Goal: Estimate \bar{Y}

(4)

$$\text{Write } \bar{Y} = \frac{\sum_h N_h \bar{Y}_h}{N} = \sum_{h=1}^H \frac{N_h}{N} \bar{Y}_h$$

An unbiased estimator of \bar{Y} is

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

$$V[\bar{y}_{\text{str}}] = \frac{1}{N^2} \sum_{h=1}^H N_h^2 V[\bar{y}_h]$$

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

$$\hat{V}[\bar{y}_{str}] = \frac{1}{N^2} \sum_{h=1}^H N_h^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \quad (5)$$

Allocation:

What type of allocation would yield a self-weighting sample? (i.e. $\bar{y}_{str} = \bar{y}$)

$$\begin{aligned} \bar{y}_{str} &= \frac{\sum_{h=1}^H N_h \bar{y}_h}{N} = \frac{1}{N} \sum_{h=1}^H N_h \sum_{j=1}^{n_h} y_{hj} \frac{1}{n_h} \\ &= \sum_{h=1}^H \sum_{j=1}^{n_h} \frac{N_h}{N n_h} y_{hj} \end{aligned}$$

To force $\bar{y}_{str} = \bar{y}$, we need $\frac{N_h}{N n_h} = \frac{1}{n}$ (6)

So $n_h = \frac{N_h}{N} n$, which is proportional allocation

ANOVA decomposition:

$$\begin{aligned} SS_{TOT} &= \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \bar{y})^2 \\ &= \sum_{h=1}^H \sum_{j=1}^{N_h} (\underbrace{y_{hj} - \bar{y}_h}_{\text{within group}} + \underbrace{\bar{y}_h - \bar{y}}_{\text{between groups}})^2 \end{aligned}$$

$$= \sum_h \sum_i \overset{(1)}{(y_{hj} - \bar{y}_h)^2} + \sum_h \sum_j \overset{(2)}{(\bar{y}_h - \bar{y})^2} \quad (7)$$

$$+ 2 \sum_h \sum_j (y_{hj} - \bar{y}_h)(\bar{y}_h - \bar{y}) \quad (3)$$

$$\textcircled{1} = \sum_{h=1}^H (N_h - 1) S_h^2 = \text{"within stratum" sum of squares}$$

$$\textcircled{2} = \sum_{h=1}^H N_h (\bar{y}_h - \bar{y})^2 = \text{"between strata" sum of squares}$$

$$\textcircled{3} = 2 \sum_{h=1}^H \left[(\bar{y}_h - \bar{y}) \underbrace{\sum_{j=1}^{N_h} (y_{hj} - \bar{y}_h)}_{\sum_{j=1}^{N_h} y_{hj} - N_h \bar{y}_h} \right] \quad (8)$$

$$\text{So } SS_{\text{TOT}} = SS_{\text{within}} + SS_{\text{between}}$$

Since $V(\bar{y}_{\text{str}})$ only involves SS_{within} , ideally SS_{within} should be small & SS_{between} large

⑨

Optimal allocation

Assume we have a setup of C_0

& a per-item cost of C_h in stratum h

Then the total cost of the sample is

$$C = C_0 + \sum_{h=1}^H n_h C_h$$

$$\text{Also, } V[\bar{y}_{sr}] = \frac{1}{N^2} \sum_{h=1}^H N_h^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

$$= \frac{1}{N^2} \sum_{h=1}^H N_h^2 \frac{s_h^2}{n_h} - \frac{1}{N^2} \sum_{h=1}^H N_h^2 \frac{s_h^2}{n_h} \frac{n_h}{N_h} \quad (10)$$

We will minimize

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

$$\vec{u} \cdot \vec{v} = |\vec{u}| |\vec{v}| \cos \theta$$

$$(\vec{u} \cdot \vec{v})^2 = |\vec{u}|^2 |\vec{v}|^2 \cos^2 \theta \leq |\vec{u}|^2 |\vec{v}|^2$$

Cauchy-Schwarz Inequality

(11)

$$\text{let } u_h = \frac{N_h}{N} \frac{S_h}{\sqrt{n_h}} \text{ and } v_h = \sqrt{n_h} c_h$$

$$\begin{aligned} \text{Now we are minimizing } & \left(\sum_{h=1}^H v_h^2 \right) \left(\sum_{h=1}^H u_h^2 \right) \\ &= |\vec{v}|^2 |\vec{u}|^2 \end{aligned}$$

C-S says that this is minimized when $\vec{u} = k\vec{v}$

$$\text{That is, } \frac{N_h}{N} \frac{S_h}{\sqrt{n_h}} = k \sqrt{n_h} c_h$$

(12)

$$n_h = \frac{N_h}{N} \frac{S_h}{k \sqrt{c_h}} \quad \text{But } k = ?$$

$$n = \sum_{h=1}^H n_h = \sum_{h=1}^H \frac{N_h}{N} \frac{S_h}{k \sqrt{c_h}}$$

$$\Rightarrow k = \frac{1}{n} \sum_{h=1}^H \frac{N_h}{N} \frac{S_h}{\sqrt{c_h}}$$

$$\therefore n_h = \frac{N_h S_h}{N \sqrt{c_h}} \frac{n}{\sum_{h=1}^H \frac{N_h}{N} \frac{S_h}{\sqrt{c_h}}}$$

(13)

$$n_h = \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^H \frac{N_h S_h}{\sqrt{C_h}}} \cdot n$$

This is called optimal allocation

What if the cost per item is the same in all strata?

$$n_h = \frac{N_h S_h}{\sum_{h=1}^H N_h S_h} \cdot n$$

This is called Neyman allocation

(14)

What if the population variances are the same in all strata?

$$n_h = \frac{N_h}{\sum_{h=1}^H N_h} \cdot n = \frac{N_h}{N} \cdot n, \text{ which is proportional allocation}$$

What if the strata are all the same size?

$$n_h = \frac{\frac{N}{H} \cdot n}{N} = \frac{n}{H}, \text{ which is equal allocation}$$