Stratified Sampling

Idea: divide heterogeneous population to stratas, which are homogeneous within each strata and heterogeneous between strata

1) $N = \sum_{i=1}^{k} N_i$ divide into k strata

2) n_i sample size in each strata (SRS WOR)

$$n = \sum_{i=1}^{k} n_i$$

⊛ Cluster sampling:

within heterogenous

between homogenous

Stratified (variance reduction technique)

$E(Y)$

n : $\frac{1}{n}\sum y_i = \bar{y}$ $\quad \bar{y}$ – unbiased

$$var(\bar{y}) = \frac{var(y)}{n}$$

K subgroups : w_k – prob. to be in group k

$$n_k = w_k \cdot n$$

$$\hat{y} = \sum w_k \cdot \bar{y}_k$$

$\bar{y}_k$ – s. mean within k^{th} group

$$E(\hat{y}_{strat}) = \sum_{k=1}^{K} w_k \, E(\bar{y}_k) = \sum_{k=1}^{K} w_k \mu_k = \mu$$

$$var(\hat{y}_{strat}) = \sum_{k=1}^{K} w_k^2 \cdot var(\bar{y}_k) =$$

$$= \sum_{k=1}^{K} \frac{n_k^{\cancel{2}}}{n^2} \cdot \frac{1}{\cancel{n_k}} \cdot \sigma_k^2 = \frac{1}{n} \sum_{k=1}^{K} w_k \cdot \sigma_k^2$$

$$E(\bar{y}) = \dots = \mu$$

$$Var(\bar{y}) = \sigma^2 / n$$

$$\underset{srs}{Var(\bar{y})} \overset{?}{=} \underset{strat}{Var(\hat{y})}$$

z - strat. covariate

$$Var(y) = E(Var(y|z)) + Var(E(y|z)) =$$

$$E\left(\sum \sigma_k^2 \cdot I(z=k)\right) + Var\left(\sum \mu \cdot I(z=k)\right) =$$

$$\sum \sigma_k^2 \cdot E(I(z=k)) + E\left(\left(\sum \mu \cdot I(z=k)\right)^2\right) -$$

$$E^2\left(\sum \mu \cdot I(z=k)\right) =$$

$$\sum \sigma_k^2 \cdot w_k + \sum \mu_k^2 w_k - \mu^2$$

$$= \underbrace{\sum \sigma_k^2 w_k}_{\text{within}} + \underbrace{\sum w_k (\mu_k - \mu)^2}_{\text{between}}$$

$$\underset{srs}{Var(\bar{y})} = \frac{1}{n} \sum \sigma_k^2 w_k + \frac{1}{n} \sum w_k (\mu_k - \mu)^2$$

$$\text{''} \quad Var(\hat{y}_{strat})$$

How to chose n_i, $i=1,\dots,k$

(i) min cost of survey

(ii) max precision

1) Equal allocation:

$$n_i = \frac{n}{k}$$

2) Proportional allocation:

$$n_i \propto N$$

$$n_i = \delta N_i$$

$$n = \delta \cdot N \Rightarrow \delta = n/N$$

$$n_i = \left(\frac{n}{N}\right) \cdot N_i$$

Sampling error $(\hat{y}_{prop}) \leq$ s.e. $(\hat{y}_{SRS})$

more diverse $\Rightarrow$ more precise $\hat{y}_{prop}$.

3) Optimal (Neyman) allocation

$n_i \propto N_i S_i$

$n_i = \delta^* N_i S_i$

$n = \delta^* \sum N_i S_i \Rightarrow \delta^* = \frac{n}{\sum N_i S_i}$

$n_i = \frac{n}{\sum N_i S_i} N_i S_i$

$\hat{y}_{opt}$ As precise / more precise than $\hat{y}_{prop}$.

more precise when stratas' se. differ

Cost of survey:

$$C = C_0 + \sum c_i n_i$$

c_i ↑ per unit in ith state

$$Var(\bar{y}) \to \min_{n_i} \quad \text{s.t. } C = c_0 + \sum c_i n_i$$

$$\mathcal{L} = Var(\bar{y}_{strat}) + \lambda^2 (C - C_0) =$$

$$= \sum w_i \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 + \lambda^2 \cdot \sum c_i n_i$$

$$= \sum \frac{w_i S_i^2}{n_i} + \lambda^2 \sum c_i n_i - \sum \frac{w_i S_i^2}{N_i} =$$

$$= \sum \left(\frac{w_i S_i}{\sqrt{n_i}} - \lambda \sqrt{c_i n_i} \right)^2 + B$$

$\mathcal{L} \to \min$ at $\dfrac{w_i S_i}{\sqrt{n_i}} = \lambda \sqrt{c_i n_i}$

$$n_i = \frac{1}{\lambda} \cdot \frac{w_i S_i}{\sqrt{c_i}}$$

$\lambda \nearrow$ min var , fixed cost (i)

$\lambda \searrow$ min cost , fixed var (ii)

(i) $C = C_0^*$

$$\sum c_i n_i = C_0^*$$

$$\sum c_i \cdot \frac{w_i S_i}{\lambda \sqrt{c_i}} = C_0^*$$

$$\lambda_{FC} = \frac{\sum \sqrt{c_i} \cdot w_i S_i}{C_0^*}$$

$$n_i = \frac{1}{\lambda_{FC}} \cdot \frac{w_i S_i}{\sqrt{c_i}}$$

(ii)
$$\sum \left(\frac{1}{n_i} - \frac{1}{N_i} \right) w_i^2 S_i^2 = V_0$$

$$\sum \frac{\lambda \sqrt{c_i}}{w_i S_i} w_i^2 S_i^2 = V_0 + \sum \frac{w_i^2 S_i^2}{N_i}$$

$$\lambda_{FV} = \frac{V_0 + \sum \frac{w_i^2 S_i^2}{N_i}}{\sum w_i S_i \sqrt{c_i}} \qquad n_i = \frac{1}{\lambda_{FV}} \cdot \frac{w_i S_i}{\sqrt{c_i}}$$

Sample size	$\bar{X}$	$\hat{p}$
SRS	$\dfrac{z^2_{\alpha/2} \cdot \sigma^2}{e^2}$	$\dfrac{z^2_{\alpha/2} \cdot p \cdot (1-p)}{e^2}$
Prop	$\dfrac{z^2_{\alpha/2} \cdot \Sigma W_k \cdot \sigma_k^2}{e^2}$	$\dfrac{z^2_{\alpha/2}\ \Sigma W_k \cdot p_k(1-p_k)}{e^2}$
Opt	$\dfrac{z^2_{\alpha/2} \cdot (\Sigma W_k \cdot \sigma_k)^2}{e^2}$	$\dfrac{z^2_{\alpha/2} \cdot \left(\Sigma W_k \sqrt{p_k \cdot (1-p_k)}\right)^2}{e^2}$

if $\sigma_k = \sigma_j$, $n_k = n_j$ $\forall\ k, j$

$$\frac{(\Sigma W_k \sigma_k)^2}{e^2} = \frac{\left(\Sigma \frac{1}{k} \cdot \sigma\right)^2}{e^2} = \frac{k/k \cdot \sigma^2}{e^2}$$