

Module 3: Stratified Sampling

Introduction

Under SRS each unit of the population is selected in the sample with the same probability. Such a scheme may not be appropriate for sampling from a population that is heterogeneous.

Example

Populations All households in India
(approximately $N = 300$ million)
 ≈ 30 crores

Goal: To estimate average household income of a particular year.

Under SRS:

- (i) Each household has the same probability of getting selected in the sample. But income distribution is highly skewed with urban households having significantly higher income than rural households.
- (ii) A simple random sample may:
 - Overrepresent rural households (which are more in number; say 80% of the population is rural)
 - Underrepresent urban households (which have higher variance)
 - Variance of the estimation may be high because of heterogeneity between rural and urban groups

Solution: (i) Stratify (divide) the population into:

- Urban households
 - Rural households
- (ii) Take independent samples from each stratum (e.g. proportional to their size or optimized for variance)
 - (iii) Hence each stratum is homogeneous (lower within stratum variance)

(IV) Ensures adequate representation of urban households, which have higher variability and impact on overall mean.

The SRS is, although simple to implement, not appropriate for sampling from heterogeneous population. Stratified sampling may be more appropriate in those situations.

$$U = (U_1, \dots, U_H) \quad (\text{Population})$$

$$Y = (Y_1, \dots, Y_H) \quad (\text{Study variable})$$

Divide the whole population U into a L subpopulations, where units within each subpopulation are homogeneous.

$$\text{Subpopulation } U_h = (U_{h1}, \dots, U_{hN_h})$$

$$Y_h = (Y_{h1}, \dots, Y_{hN_h}) \quad ; \quad h=1, 2, \dots, L$$

N_h : Subpopulation size of h^{th} subpopulation (stratum), $h=1, 2, \dots, L$

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} \quad ; \quad h^{\text{th}} \text{ subpopulation total, } h=1, \dots, L$$

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi} = \frac{Y_h}{N_h} \quad ; \quad h^{\text{th}} \text{ subpopulation mean, } h=1, \dots, L$$

$$\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \quad ; \quad h^{\text{th}} \text{ subpopulation variance, } h=1, \dots, L$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 = \frac{N_h \sigma_h^2}{N_h - 1}, \quad h=1, \dots, L$$

$\bar{Y} = \frac{1}{N} \sum_{h=1}^L N_h \bar{Y}_h = \sum_{h=1}^L W_h \bar{Y}_h$ (SRSWOR or SRWOR), where $W_h = \frac{N_h}{N}$, $\sum_{h=1}^L W_h = 1$

Independent random samples of sizes $n_1, \dots, n_L$ are taken from strata 1, 2, ..., L, respectively.

Let $n = \sum_{h=1}^L n_h$: total sample size.

h^{th} stratum random sample::

$$\underline{S}_h = (S_{h1}, \dots, S_{h n_h}), \quad h=1, \dots, L, \quad S_{hi} \in \{U_{h1}, \dots, U_{hL}\}, \quad h=1, \dots, L$$

Let $\underline{Z}_h = (Z_{h1}, \dots, Z_{h n_h})$: Study variable corresponding to $\underline{S}_h = (S_{h1}, \dots, S_{h n_h}), \quad h=1, \dots, L$

Then

$$\hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi}$$

is an unbiased estimator of $\bar{Y}_h, \quad h=1, \dots, L$, under both the SRSWR and SRSWOR.

An estimator of population mean $\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h$

$$\hat{\bar{Y}}_{ST} = \sum_{h=1}^L W_h \hat{\bar{Y}}_h$$

$$E(\hat{\bar{Y}}_{ST}) = \sum_{h=1}^L W_h \bar{Y}_h = \bar{Y}$$

Define

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{h=1}^L n_h \hat{\bar{Y}}_h \rightarrow \text{Sample mean}$$

$$E(\hat{\bar{Y}}) = \frac{1}{n} \sum_{h=1}^L n_h \bar{Y}_h \neq \bar{Y}$$

The sample mean $\hat{\bar{Y}}$ is a biased estimator of population mean $\bar{Y}$

$$\text{Var}(\hat{\bar{Y}}_{ST}) = \begin{cases} \sum_{h=1}^L W_h^2 \frac{\sigma_h^2}{n_h}, & \text{Under SRSWR} \\ \sum_{h=1}^L W_h^2 (1 - f_h) \frac{S_h^2}{n_h}, & \text{Under SRSWOR} \end{cases}$$

Where $f_h = \frac{n_h}{N}, \quad h=1, \dots, L$

$$\text{Var}(\hat{\bar{Y}}) = \begin{cases} \frac{1}{n^2} \sum_{h=1}^L n_h \sigma_h^2, & \text{Under SRSWR} \\ \frac{1}{n^2} \sum_{h=1}^L n_h (1 - \frac{n_h}{N}) S_h^2, & \text{Under SRSWOR} \end{cases}$$

if $\frac{1/n_h}{1/N} \leq \frac{n_h}{N}$, $h=1, \dots, L$, then $\text{Var}(\hat{\bar{y}}_{ST}) \leq \text{Var}(\hat{\bar{y}})$

→ Disproportionate allocation

Define $\hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \hat{\bar{y}}_h)^2$, $h=1, \dots, L$.

Then

$$\hat{\text{Var}}(\hat{\bar{y}}_{ST}) = \begin{cases} \sum_{h=1}^L W_h^2 \frac{\hat{S}_h^2}{n_h}, & \text{Under SRSWR} \\ \sum_{h=1}^L W_h^2 (1 - b_h) \frac{\hat{S}_h^2}{n_h}, & \text{Under SRSWOR} \end{cases}$$

$$\hat{\text{Var}}(\hat{\bar{y}}) = \begin{cases} \frac{1}{N^2} \sum_{h=1}^L n_h \hat{S}_h^2, & \text{Under SRSWR} \\ \frac{1}{N^2} \sum_{h=1}^L n_h (1 - \frac{n_h}{N}) \hat{S}_h^2, & \text{Under SRSWOR} \end{cases}$$

Sample Size Allocation

Issue: How to allocate sample sizes to different strata?

(I) Proportional Allocation. (fixed total sample size)

Suppose that the total sample size n is fixed.
Under proportional allocation

$$n_h \propto N_h, \quad h=1, \dots, L$$

$$\Rightarrow n_h = k N_h, \quad h=1, \dots, L, \quad \text{for some constant } k.$$

$$\Rightarrow \sum_{h=1}^L n_h = k \sum_{h=1}^L N_h \Rightarrow k = \frac{n}{N}$$

$$\Rightarrow \frac{n_h}{n} = \frac{N_h}{N}, \quad h=1, \dots, L$$

$$n_h = \frac{n}{N} N_h, \quad h=1, \dots, L$$

or $n_h = n W_h, \quad h=1, \dots, L$

5/27

5/18

Under SRSWR

$$V_{\text{prop}}(\hat{q}_{\text{ST}}) = \sum_{h=1}^L W_h^2 \frac{\sigma_h^2}{n_h} = \sum_{h=1}^L \frac{N_h^2}{N^2} \cdot \frac{\sigma_h^2}{\frac{n}{N} N_h}$$

$$V_{\text{prop}}(\hat{q}_{\text{ST}}) = \frac{1}{n} \sum_{h=1}^L W_h \sigma_h^2$$

Under SRSWOR

$$\begin{aligned} V_{\text{prop}}(\hat{q}_{\text{ST}}) &= \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ &= \frac{1}{n} \sum_{h=1}^L W_h \left(1 - \frac{n_h W_h}{N W_h}\right) S_h^2 \\ &= \left(\frac{1-b}{n}\right) \sum_{h=1}^L W_h S_h^2 \end{aligned}$$

$$V_{\text{prop}}(\hat{q}_{\text{ST}}) = \begin{cases} \frac{1}{n} \sum_{h=1}^L W_h \sigma_h^2, & \text{Under SRSWR} \\ \left(\frac{1-b}{n}\right) \sum_{h=1}^L W_h S_h^2, & \text{Under SRSWOR} \end{cases}$$

Special Case:

$$\begin{aligned} S_1^2 &= \dots = S_L^2 = S_0^2, \text{ say} \\ V_{\text{prop}}(\hat{q}_{\text{ST}}) &= \frac{1-b}{n} S_0^2 \quad (\text{Under SRSWOR}) \end{aligned}$$

Under proportional allocation, a pooled unbiased estimator of S_0^2 is

$$\hat{S}_0^2 = \frac{\sum_{h=1}^L (n_h - 1) \hat{S}_h^2}{n - L}$$

$$V_{\text{prop}}(\hat{q}_{\text{ST}}) = \frac{1-b}{n} \hat{S}_0^2$$

Under SRSWOR

Proportional allocation (Proportional allocation with fixed cost) Cost based allocation

C_h : Cost of Collecting information from single unit of h^{th} stratum, $h=1, 2, \dots, L$

C_0 : fixed or overhead cost

Total Survey Cost: $C = C_0 + \sum_{h=1}^L C_h n_h$

Under proportional allocation

$$n_h = k N_h, \quad h=1, \dots, L$$

$$\Rightarrow \sum_{h=1}^L C_h n_h = k \sum_{h=1}^L C_h N_h$$

$$\Rightarrow k = \frac{\sum_{h=1}^L C_h n_h}{\sum_{h=1}^L C_h N_h} = \frac{C - C_0}{\sum_{h=1}^L C_h N_h}$$

$$\Rightarrow n_h = \frac{C - C_0}{\sum_{h=1}^L C_h N_h} N_h, \quad h=1, \dots, L$$

$$\Rightarrow n = \sum_{h=1}^L n_h = \frac{C - C_0}{\sum_{h=1}^L C_h N_h} N$$

$$n_h = \frac{C - C_0}{\sum_{h=1}^L C_h N_h} N_h, \quad h=1, \dots, L$$

$$n = \frac{C - C_0}{\sum_{h=1}^L C_h N_h} N$$

* ←

Under SRSWOR

$$V_{\text{Prob}}(\hat{Y}_{\text{ST}}) = \sum_{h=1}^L \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

$$= \sum_{h=1}^L \frac{N_h}{N^2} \left(\frac{\sum_{h=1}^L C_h N_h}{C-G} \right) S_h^2 - \sum_{h=1}^L \frac{N_h}{N^2} S_h^2$$

$$= \frac{1}{C-G} \left(\sum_{h=1}^L W_h S_h^2 \right) \left(\sum_{h=1}^L W_h C_h \right) - \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

$$V_{prop}(\hat{\gamma}_{st}) = \frac{1}{N} \left(\sum_{h=1}^L W_h S_h^2 \right) \left(\frac{N}{C-G} \sum_{h=1}^L W_h C_h - 1 \right)$$

$$\hat{V}(\hat{\gamma}_{st}) = \frac{1}{N} \left(\sum_{h=1}^L W_h \hat{S}_h^2 \right) \left(\frac{N}{C-G} \sum_{h=1}^L W_h C_h - 1 \right)$$

Optimum Allocation

Proportional allocation only focuses on strata size and ignores other important features of stratum, such as precision/variance of estimator, strata cost etc.

Goal: Minimize precision/variance of the estimator
Subject to fixed sample size or fixed cost.

Neyman Allocation For SRSWOR

Here the goal is to minimize the variance of estimator subject to a fixed total sample size
 $n (= \sum_{h=1}^L n_h)$

$$V(\hat{\gamma}_{st}) = \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_h^2}{n_h}$$

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

8/24

8/18

Our goal is to minimize $V(\hat{Y}_{ST})$ w.r.t. $n_1, \dots, n_L$
 Under the constraint $\sum_{h=1}^L n_h = n$ (n is a fixed integer)

By the method of Lagrange's multiplier, if
 $(\hat{n}_1, \dots, \hat{n}_L)$ is an extremum then for some $\lambda \in \mathbb{R}$

$$\frac{\partial}{\partial n_h} V(\hat{Y}_{ST}) = \lambda \frac{\partial}{\partial n_h} (n - \sum_{h=1}^L n_h), \quad h=1, \dots, L$$

$$e) \quad -\frac{1}{N^2} \cdot \frac{W_h^2 S_h^L}{n_h^2} = -\lambda$$

$$\Rightarrow n_h = \frac{W_h S_h}{N \sqrt{\lambda}}, \quad h=1, \dots, L$$

$$\begin{aligned} \sum_{h=1}^L n_h &= n \\ \Rightarrow n &= \frac{1}{\sqrt{\lambda}} \sum_{h=1}^L W_h S_h \Rightarrow \sqrt{\lambda} = \frac{\sum_{h=1}^L W_h S_h}{n} \end{aligned}$$

$$\Rightarrow n_h = \frac{W_h S_h}{\sum_{h=1}^L W_h S_h} n, \quad h=1, \dots, L$$

Allocate larger sample size to stratum if W_h is large
 and/or S_h is large.

$$V_{\text{Neyman}}(\hat{Y}_{ST}) = \frac{1}{N^2} \sum_{h=1}^L \frac{W_h^2 S_h^L}{W_h S_h n} \left(\sum_{h=1}^L W_h S_h \right) = \frac{1}{N^2} \sum_{h=1}^L N_h S_h^L$$

$$V_{\text{Neyman}}(\hat{Y}_{ST}) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^L$$

Fixed Cost Optimum Allocation

Let $c, c_0, c_1, \dots, c_L$ be fixed and

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

Goal is to minimize

$$V(\hat{\tau}) = \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

Subject to the constraint

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

Again, by the method of Lagrange's multiplier, optimum solution is obtained

$$\frac{\partial V(\hat{\tau})}{\partial n_h} = \lambda \left(C - C_0 - \sum_{h=1}^L C_h n_h \right), \quad h=1, \dots, L$$

$$\Rightarrow -\frac{W_h^2 S_h^2}{n_h^2} = -\lambda C_h, \quad h=1, \dots, L$$

$$\Rightarrow n_h = \frac{W_h S_h}{\sqrt{\lambda} \sqrt{C_h}}, \quad h=1, \dots, L$$

$$\Rightarrow \sum_{h=1}^L n_h C_h = \frac{1}{\sqrt{\lambda}} \sum_{h=1}^L W_h S_h \sqrt{C_h}$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum_{h=1}^L W_h S_h \sqrt{C_h}}{C - C_0}$$

$$\Rightarrow n_h = \frac{\frac{W_h S_h (C - C_0)}{\sqrt{C_h}}}{\sum_{h=1}^L W_h S_h \sqrt{C_h}}, \quad h=1, \dots, L$$

Allocate larger sample size to a stratum if

(i) Stratum size N_h is large

(ii) Stratum variance is large

(iii) Sampling from stratum is cheaper.

$$n = (C - C_0) \frac{\sum_{h=1}^L \frac{W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^L W_h S_h \sqrt{C_h}}$$

$$V_{OPT}(\hat{\theta}_{ST}) = \sum_{h=1}^L \frac{W_h^2 S_h^2}{\frac{W_h S_h}{\sqrt{C_h}} (C - C_0)} \left(\sum_{h=1}^L W_h S_h \sqrt{C_h} \right) - \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

$$V_{OPT}(\hat{\theta}_{ST}) = \frac{1}{C - C_0} \left(\sum_{h=1}^L W_h S_h \sqrt{C_h} \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

Remark One can also consider minimizing the cost for a fixed precision/variance V_0 , i.e.

$$\text{minimize } C = C_0 + \sum_{h=1}^L n_h c_h$$

Subject to the constraint

$$\sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 = V_0$$

Then the optimum solution satisfies

$$\frac{\partial C}{\partial n_h} = \lambda \frac{\partial}{\partial n_h} \left(V_0 - \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} + \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \right), \quad h=1, \dots, L$$

$$\Rightarrow C_h = \lambda \frac{W_h^2 S_h^2}{n_h^2}, \quad h=1, \dots, L$$

$$n_h = \sqrt{\lambda} \frac{W_h S_h}{\sqrt{C_h}}, \quad h=1, \dots, L$$

$$\sum_{h=1}^L \frac{W_h^2 S_h^2 \sqrt{C_h}}{\sqrt{\lambda} W_h S_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 = V_0$$

$$\sqrt{\lambda} = \frac{\sum_{h=1}^L W_h S_h \sqrt{c_h}}{V_0 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$$

$$\Rightarrow h_h = \left(\frac{\sum_{h=1}^L W_h S_h \sqrt{c_h}}{V_0 + \frac{1}{N} \sum_{h=1}^L W_h S_h^2} \right) \frac{W_h S_h}{\sqrt{c_h}}, \quad h=1, \dots, L$$

Theorem

We have, for large H_h 's (and hence large $N = \sum_{h=1}^L N_h$)

$$V_{\text{ran}}(\hat{\eta}) \approx V_{\text{opt}}(\hat{\eta}_{\text{ST}}) + \frac{1}{n} \sum_{h=1}^L W_h (S_h - \bar{S})^2 + \frac{1}{n} \sum_{h=1}^L W_h (\bar{\gamma}_h - \bar{\gamma})^2$$

$$\text{i.e. } V_{\text{ran}}(\hat{\eta}) \approx V_{\text{prop}}(\hat{\eta}_{\text{ST}}) + \frac{1}{n} \sum_{h=1}^L W_h (\bar{\gamma}_h - \bar{\gamma})^2$$

$$\text{i.e. } V_{\text{ran}}(\hat{\eta}) \approx V_{\text{prop}}(\hat{\eta}_{\text{ST}}) \geq V_{\text{opt}}(\hat{\eta}_{\text{ST}}).$$

We have

Proof.

$$V_{\text{ran}}(\hat{\eta}) = (1-b) \frac{S^2}{n}, \quad \text{where } b = \frac{n}{N} \approx 1$$

$$V_{\text{prop}}(\hat{\eta}_{\text{ST}}) = \frac{(1-b)}{n} \sum_{h=1}^L W_h S_h^2 \approx \frac{1}{n} \sum_{h=1}^L W_h S_h^2$$

$$\text{and } V_{\text{opt}}(\hat{\eta}_{\text{ST}}) = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h^2 \right) - \frac{1}{N} \left(\sum_{h=1}^L W_h S_h \right)^2 \approx \frac{1}{n} \left(\sum_{h=1}^L W_h S_h^2 \right)$$

$$(N-1) S^2 = \sum_{i=1}^N (\gamma_i - \bar{\gamma})^2$$

$$= \sum_{h=1}^L \sum_{i=1}^{N_h} (\gamma_{hi} - \bar{\gamma})^2$$

$$= \sum_{h=1}^L \sum_{i=1}^{N_h} |\gamma_{hi} - \bar{\gamma}_h + \bar{\gamma}_h - \bar{\gamma}|^2$$

$$= \sum_{h=1}^L \sum_{i=1}^{N_h} (\gamma_{hi} - \bar{\gamma}_h)^2 + \sum_{h=1}^L N_h (\bar{\gamma}_h - \bar{\gamma})^2$$

$$= \sum_{h=1}^L \frac{N_h}{n} + \sum_{h=1}^L N_h (\bar{\gamma}_h - \bar{\gamma})^2$$

For large H_h 's and N_h , $N_h \approx N$, $(1-b) = (1 - \frac{n}{N}) \approx 1$,
 $\sigma_h^2 \approx S_h^2$ ($\sigma_h^2 = \frac{(N_h - 1)}{N_h} S_h^2$)

$$S^2 \approx \sum_{h=1}^L W_h S_h^2 + \sum_{h=1}^L W_h (\bar{y}_h - \bar{y})^2$$

$$V_{\text{Ran}}(\hat{y}) \approx \frac{1}{n} \sum_{h=1}^L W_h S_h^2 + \frac{1}{n} \sum_{h=1}^L W_h (\bar{y}_h - \bar{y})^2$$

$$V_{\text{prop}}(\hat{y}_{\text{ST}}) \approx \frac{1}{n} \sum_{h=1}^L W_h S_h^2; \quad V_{\text{opt}}(\hat{y}_{\text{ST}}) \approx \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2$$

Let $\bar{S} = \sum_{h=1}^L W_h S_h$. Then

$$\left(\sum_{h=1}^L W_h S_h \right)^2 = \sum_{h=1}^L W_h S_h^2 - \sum_{h=1}^L W_h (S_h - \bar{S})^2 \quad \left[(E(X))^2 = E(X^2) - V(X) \right]$$

$$V_{\text{opt}}(\hat{y}_{\text{ST}}) \approx \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2$$

$$= \frac{1}{n} \sum_{h=1}^L W_h S_h^2 - \frac{1}{n} \sum_{h=1}^L W_h (S_h - \bar{S})^2$$

$$= V_{\text{prop}}(\hat{y}_{\text{ST}}) - \sum_{h=1}^L W_h (S_h - \bar{S})^2$$

$$\begin{aligned} V_{\text{Ran}}(\hat{y}) &\approx V_{\text{prop}}(\hat{y}_{\text{ST}}) + \frac{1}{n} \sum_{h=1}^L W_h (\bar{y}_h - \bar{y})^2 \\ &\approx V_{\text{opt}}(\hat{y}_{\text{ST}}) + \frac{1}{n} \sum_{h=1}^L W_h (S_h - \bar{S})^2 + \frac{1}{n} \sum_{h=1}^L W_h (\bar{y}_h - \bar{y})^2 \end{aligned}$$

Remark (a) When we move from SRS to Stratified Sampling, there is a reduction in variance due to

- (i) elimination of differences among stratum standard deviations;
- (ii) elimination of differences among stratum means.

(b) We always have

$$V_{\text{opt}}(\hat{y}_{\text{ST}}) \leq V_{\text{prop}}(\hat{y}_{\text{ST}})$$

However we may have $V_{\text{prop}}(\hat{y}_{\text{ST}}) \geq V_{\text{Ran}}(\hat{y})$

$$(N-1)S^2 = \sum_{h=1}^L (N_h-1) S_h^2 + \sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2$$

$$V_{\text{Ran}}(\bar{y}) = (1-b) \cdot \frac{S^2}{n}$$

$$= \frac{N-n}{n(N-1)} \left[\sum_{h=1}^L (N_h-1) S_h^2 + \sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2 \right]$$

$$V_{\text{Ran}}(\bar{y}) - V_{\text{prop}}(\hat{\bar{y}}_{\text{ST}}) = \frac{N-n}{nN} \left[\frac{1}{N-1} \sum_{h=1}^L (N_h-1) S_h^2 + \frac{1}{N-1} \sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2 - \sum_{h=1}^L \frac{N_h}{N} S_h^2 \right]$$

$$= \frac{N-n}{nN} \left[\sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2 - \sum_{h=1}^L \frac{N_h - N_h}{N} S_h^2 \right]$$

$V_{\text{prop}}(\hat{\bar{y}}_{\text{ST}}) > V_{\text{Ran}}(\bar{y})$ is possible if

$$\sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2 < \frac{1}{N} \sum_{h=1}^L (N_h - N_h) S_h^2$$

Between Strata Variance

Within Strata Variance

For example if $S_1^2 = \dots = S_L^2 = S_0^2$, say, then

$V_{\text{prop}}(\hat{\bar{y}}_{\text{ST}}) > V_{\text{Ran}}(\bar{y})$ is possible if

$$\sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2 < (L-1) S_0^2$$

$$\text{or } S_0^2 > \frac{1}{L-1} \sum_{h=1}^L N_h (\bar{y}_h - \bar{y})^2$$

i.e. Mean square difference among strata is smaller than mean square difference within strata

Estimation of Subpopulation Total and Subpopulation Mean

$$U = (U_1, \dots, U_N), \quad Y = (Y_1, \dots, Y_N)$$

Whole population U is divided into h strata, $h=1, \dots, L$.

N_h : Size of h^{th} strata, $h=1, \dots, L$

$U_h = (U_{h1}, \dots, U_{hN_h})$: h^{th} strata, $h=1, \dots, L$

$Y_h = (Y_{h1}, \dots, Y_{hN_h})$: Study Variable of h^{th} strata, $h=1, \dots, L$

$U^* = (U_1^*, \dots, U_{N^*}^*)$: Subpopulation of U

N^* : # of units in the subpopulation ($N^* < N$)

Define

$$Y_{hi}^* = \begin{cases} Y_{hi} & \text{if } U_{hi} \in U^* \\ 0 & \text{otherwise} \end{cases}, \quad i=1, \dots, N_h, \quad h=1, \dots, L$$

$$N_h^* = \sum_{i=1}^{N_h} I(U_{hi} \in U^*) = \text{Subpopulation size of the } h^{\text{th}} \text{ strata, } h=1, \dots, L$$

$$Y_h^* = \sum_{i=1}^{N_h} Y_{hi}^* = h^{\text{th}} \text{ stratum subpopulation total, } h=1, \dots, L$$

$$Y^* = \sum_{h=1}^L Y_h^* = \text{Subpopulation total}$$

n_h = Sample size for h^{th} stratum, $h=1, 2, \dots, L$

n_h^* = number of units of subpopulation in the h^{th} stratum.

$Z_{h1}, \dots, Z_{hn_h}, \quad h=1, \dots, L$

→ independent random samples from h strata.

Unit corresponding to Z_{hi} is denoted by $\hat{U}_{hi}$, $i=1, \dots, n_h$, $h=1, \dots, L$

20/15 15/18

Then

$$n_h^* = \sum_{i=1}^{n_h} \mathbb{I}(\hat{U}_{hi} \in U^*), \quad h=1, \dots, L$$

Define

$$Z_{hi}^* = \begin{cases} Z_{hi} & \text{if } \hat{U}_{hi} \in U^* \\ 0 & \text{o.w.} \end{cases}, \quad i=1, \dots, n_h, \quad h=1, \dots, L$$

Then

$$E\left(\frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*\right) = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}^*$$

$$E\left(\frac{N_h}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*\right) = Y_h^*, \quad h=1, \dots, L$$

$$\Rightarrow E\left(\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*\right) = Y^*$$

$$\hat{Y}^* = \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*$$

Then

$$E(\hat{Y}^*) = Y^*$$

$$\text{Var}(\hat{Y}^*) = \sum_{h=1}^L N_h^2 \text{Var}\left(\frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*\right)$$

$$\text{Var}(\hat{Y}^*) = \sum_{h=1}^L N_h^2 \left(1 - \frac{n_h}{N_h}\right) S_h^{*2}$$

$$\text{where } S_h^{*2} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(Y_{hi}^* - \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}^*\right)^2, \quad h=1, \dots, L$$

$$\hat{Y}^* = \frac{\hat{Y}}{N^*} = \frac{1}{N^*} \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*$$

$$E(\hat{Y}^*) = Y^*$$

Reg 15

Estimation of H_h^*

Define

$$Y_{hi}^* = \begin{cases} 1, & \text{if } U_{hi} \in U^* \\ 0, & \text{o.w.} \end{cases}, \quad i=1, \dots, N_h, \quad h=1, \dots, L$$

Then

$$H_h^* = \sum_{i=1}^{N_h} Y_{hi}^*$$

$$\hat{H}_h^* = \frac{N_h}{n_h} \sum_{i=1}^{n_h} Z_{hi}^* \quad \text{is an unbiased estimator of } H_h^*, \quad h=1, \dots, L$$

$$\begin{aligned} \text{Var}(\hat{H}_h^*) &= N_h^2 \text{Var}\left(\frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi}^*\right) \\ &= N_h^2 \left(1 - \frac{n_h}{N_h}\right) S_h^{*2}, \quad h=1, \dots, L \end{aligned}$$

Stratified Sampling For Proportions

Define

$$Y_{hi} = \begin{cases} 1, & \text{if } U_{hi} \text{ possesses attribute } A \\ 0, & \text{o.w.} \end{cases}, \quad i=1, \dots, N_h, \quad h=1, \dots, L$$

$$P_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi} = \text{proportion of units in } h^{\text{th}} \text{ stratum possessing attribute } A, \quad h=1, \dots, L$$

$$P = \frac{1}{N} \sum_{h=1}^L N_h P_h = \sum_{h=1}^L U_h P_h = \text{proportion of units in the population possessing attribute } A$$

Independent random samples from L strata:

$$Z_{h1}, \dots, Z_{hn_h}, \quad h=1, \dots, L$$

$$\hat{P}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi} = \text{proportion of units in the sample from } h^{\text{th}} \text{ stratum possessing attribute } A, \quad h=1, \dots, L$$

$$\hat{p}_{ST} = \sum_{h=1}^L w_h \hat{p}_h$$

$$E(\hat{p}_{ST}) = p$$

$$\begin{aligned} \text{Let } S_h^2 &= \frac{1}{N_h-1} \sum_{i=1}^{N_h} (Y_{hi} - p_h)^2 \\ &= \frac{1}{N_h-1} \left[\sum_{i=1}^{N_h} Y_{hi}^2 - N_h p_h^2 \right] \\ &= \frac{1}{N_h-1} \left[\sum_{i=1}^{N_h} Y_{hi} - N_h p_h^2 \right] \\ &= \frac{1}{N_h-1} \left[N_h p_h - N_h p_h^2 \right] \\ &= \frac{N_h}{N_h-1} p_h (1 - p_h), \quad h=1, \dots, L \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{p}_{ST}) &= \sum_{h=1}^L w_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\ &= \frac{\sum_{h=1}^L W_h^2 (1 - f_h) p_h (1 - p_h)}{N_h - 1} \end{aligned}$$

→ Proportional allocation
Cost based proportional allocation
Optimum allocation