Chapter 3: Stratified Sampling

Partition of population U into H **strata**: $h = 1, ..., H$ (Sec. 3.2)

$$U = U_1 \cup U_2 \cup \cdots \cup U_H$$
$$N = N_1 + N_2 + \cdots + N_H$$

according to *stratification (or stratifying) variables.*

Treat each stratum as a population of its own (Table 3.1)

- Separate sampling: $s = \cup_{h=1}^{H} s_h$ and $n = \cup_{h=1}^{H} n_h$
- Separate estimation: e.g. $\bar{y}_h$ and $\text{SE}(\bar{y}_h)$ for $h = 1, ..., H$
- Combine results: e.g. $\hat{\bar{y}}_U = \sum_h \frac{N_h}{N} \bar{y}_h$, $V(\hat{\bar{y}}_U) = \sum_h \left(\frac{N_h}{N}\right)^2 V(\bar{y}_h)$

Notation: add h to everything

Stratified simple random sampling (StrSRS)

Within-stratum estimates (p. 78):

mean: $\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i$

total: $\hat{t}_h = N_h \bar{y}_h$

variance: $s_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_{hi} - \bar{y}_h)^2$

Estimates for whole population (pp. 78-79):

mean: $\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h \quad$ where $W_h = \frac{N_h}{N}$

sampling variance: $\hat{V}(\bar{y}_{str}) = \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{s_h^2}{n_h}$

total: $\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h$

sampling variance: $\hat{V}(\hat{t}_{str}) = N^2 \hat{V}(\bar{y}_{str})$

All *point* and *variance* estimators are unbiased

Reasons for stratification (Sec. 3.1)

1. To spread the sample across the population, to ensure coverage of specific sub-populations; to avoid badly unbalanced samples

2. Strata form *domains of study* for which separate estimates of given precision for dissemination are required

3.1. Different sampling frames available for different sub-populations, e.g. the United Kingdom
3.2. Different data collection methods suitable for different units, e.g. web, telephone, person interview, etc.

4. To reduce sampling variance: variance within vs. between-strata, possibly varying stratum sampling fraction, e.g.

$$(N, n) = (6, 2) : \{y_1, y_2, y_3, y_4, y_5, y_6\} = \{2, 1, 1, 2, 2, 2\}$$

Variance decomposition (Table 3.3)

Stratum population mean $\bar{y}_{hU}$ and variance

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_{hi} - \bar{y}_{hU})^2$$

Decomposition: total = within-strata + between-strata

$$\text{total } (N-1)S^2 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_U)^2$$

$$\text{within-strata } = \sum_{h=1}^{H} \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_{hU})^2 = \sum_{h=1}^{H} (N_h - 1) S_h^2$$

$$\text{between-strata } = \sum_{h=1}^{H} N_h (\bar{y}_{hU} - \bar{y}_U)^2$$

Aim: more alike within stratum and more different between strata

Sample allocation given strata (Sec. 3.4)

Proportional allocation: $n_h/N_h = n/N \Leftrightarrow n_h/n = N_h/N$

Stratum sample size proportional stratum population size

$$\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y}$$

i.e. self-weighting (p. 40), stratified sample mean = sample mean

Compared to $\bar{y}$ under SRS: S^2 replaced by $\sum_{h=1}^{H} W_h S_h^2$

By variance decomposition:

$$\sum_{h=1}^{H} W_h S_h^2 \approx \sum_{h=1}^{H} \frac{N_h - 1}{N - 1} S_h^2 < S^2$$

unless $\bar{y}_{hU} \equiv \bar{y}_U$, where $\frac{N_h - 1}{N - 1} \approx \frac{N_h}{N} = W_h$ for large N_h

StrSRS with proportional allocation more efficient than SRS

Neyman (optimal) allocation (p. 89)

Stratum sample size and corresponding variance:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{g=1}^{H} N_g S_g}$$

$$V_{opt}(\bar{y}_{str}) = \frac{1}{n}(\sum_{h=1}^{H} W_h S_h)^2 - \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2$$

When S_h^2 is the same in every stratum, we obtain

$$n_h = n \cdot \frac{N_h}{N} \text{ (proportional allocation)}$$

Thus, for fixed n, we have

$$V_{opt}(\bar{y}_{str}) \leq V_{prop}(\bar{y}_{str}) \leq V_{srs}(\bar{y})$$

Given cost $C = c_0 + \sum_{h=1}^{H} n_h c_h$ (pp. 88-89), minimum $V(\bar{y}_{str})$ by

$$n_h = n \cdot \frac{W_h S_h / \sqrt{c_h}}{\sum_{g=1}^{H} W_g S_g / \sqrt{c_g}}$$

Practical issues (see also Sec. 3.5)

True S_h^2 unknown in advance:

- stratum variance of proxy values, e.g. previous y
- take a small pilot sample
- categoric y: proportional allocation nearly optimal
  e.g. $y_i = 0, 1$: $S_h^2 = \bar{y}_{hU}(1 - \bar{y}_{hU}) = 0.25$ if $\bar{y}_{hU} = 0.5$, or 0.21 if $\bar{y}_{hU} = 0.3$

Multiple study variables: compromise necessary

- 'average' the different optimal allocations, or optimise for a variable as weighted average of all variables
- settle for the key variable: minimax strategy
- proportional allocation as a compromise in household surveys

$n_h > N_h$: set $n_h = N_h$, so-called certainty stratum (or take-all, self-representing stratum), allocate $n - N_h$ to the rest strata

NB. dual problem: strata formation & sample allocation

Post-stratification (Sec. 4.4)

Stratifying variables may not be known throughout the population, or available at the time of sampling. Suppose

- stratifying variables are collected in the sample (e.g. hsh size)
- stratum population sizes available for estimation

Post-stratification: Form the strata in the sample, and estimate as if these were design strata to start with

Post-stratification involves *reweighting*, e.g. under SRS

- sampling weight $w_i \equiv N/n$, yielding $\bar{y}$ and $N\bar{y}$
- post-stratified weight $w_{hi} = N_h/n_h$, for realised n_h, yielding

$$\bar{y}_{pst} = \sum_{h=1}^{H} W_h \bar{y}_h \quad \text{and} \quad \hat{t}_{pst} = \sum_{h} N_h \bar{y}_h$$

Post-stratification (Sec. 4.4)

Under SRS, conditional on realised $\vec{n} = \{n_1, ..., n_H\}$,

$$V(\bar{y}_{pst}|\vec{n}) = \sum_{h=1}^{H} W_h^2 \big(1 - \frac{n_h}{N_h}\big) \frac{S_h^2}{n_h} \qquad [= V(\bar{y}_{st})]$$

i.e. variance under stratified sampling with n_h as stratum sample size

NB. $\vec{n}$ fixed under StrSRS, not otherwise. Unconditionally,

$$V(\bar{y}_{pst}) = E[V(\bar{y}_{pst}|\vec{n})] = \sum_{h=1}^{H} W_h^2 \Big[E\big(\frac{1}{n_h}\big) - \frac{1}{N_h}\Big] S_h^2$$

$$\approx \frac{1}{n}(1 - \frac{n}{N}) \sum_{h=1}^{H} W_h S_h^2 + \frac{1}{n^2} \sum_{h} (1 - W_h) S_h^2$$

i.e. StrSRS with proportional allocation + a lower-order term

NB. **Strategy** (SRS, $\bar{y}_{pst}$) nearly as efficient as (StrSRS$_{prop}$, $\bar{y}_{str}$)

Variance estimation: replace S_h^2 by post-stratum sample variance s_h^2

Prediction by one-way ANOVA model (Sec. 3.6)

One-way analysis-of-variance (ANOVA) model

$$Y_{hi} = \mu_h + \epsilon_{hi} \quad \text{and} \quad \epsilon_{hi} \overset{\text{ind}}{\sim} (0, \sigma_h^2)$$

i.e. group homogeneity model, independent ϵ_{hi}'s [within and between]

$\bar{y}_h$ is BLUP of $\bar{y}_{hU}$, with $\text{MSEP}(\bar{y}_h|s) = \left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}$ in stratum h

$$\text{MSEP}(\bar{y}_{pst}) = \sum_{h=1}^{H} W_h^2 \cdot \text{MSEP}(\bar{y}_h)$$

$$\text{NB. } \bar{y}_{pst} - \bar{y}_U = \sum_{h=1}^{H} W_h(\bar{y}_h - \bar{y}_{hU})$$

$$\text{NB. } E_M\Big(\sum_{h=1}^{H}\sum_{g\neq h} W_h W_g(\bar{Y}_h - \bar{Y}_{hU})(\bar{Y}_g - \bar{Y}_{gU})\Big) = 0$$

Quota sampling (Sec. 3.7)

Each *quota* of sub-population (n_h) as allocated stratum sample size

Difference to StrSRS: unknown within-stratum incl. prob. π_{hi}

Necessary with model-based approach

NB. similarly for non-probability sample without quotas

Super-population approach, e.g. 1-way ANOVA model

$$E_M(Y_{hi}) = \mu_h \quad \Rightarrow \quad E_M(\bar{Y}_h - \bar{Y}_{hU}) = 0$$

Quasi-randomisation approach — model of Z_{hi}'s instead, e.g.

$$E_M(Z_{hi}|\vec{y}) = \pi_h \qquad \text{given } \vec{y} = \{y_{hi} : h = 1, ..., H; i = 1, ..., N_h\}$$

$$\Rightarrow \quad \pi_h = n_h/N_h \text{ under quota sampling}$$

$$E_M(\bar{y}_h|\vec{y}) = E_M(\frac{\sum_{i \in U_h} Z_{hi} y_{hi}}{n_h}|\vec{y}) = \frac{\pi_h N_h \bar{y}_{hU}}{n_h} = \bar{y}_{hU}$$