Statistical modeling of data collected from a stratified random sampling.

- **Data**: $\underbrace{(X_{1,1}, X_{2,1}, \ldots, X_{n_1,1})}_{\in \mathbb{S}_1}, \ldots, \underbrace{(X_{1,L}, X_{2,L}, \ldots, X_{n_L,L})}_{\in \mathbb{S}_L}$,

sampling probability in S_ℓ: $1/N_\ell$

where $(X_{1,l}, \ldots, X_{n_l,l})$, $l = 1, \ldots, L$, is the data collected from the s.r.s. (either with or without replacement) taken within the lth stratum $\mathbb{S}_l$.

depends on
- n_ℓ's
- $F_{0,\ell}$'s

- distribution of data
  - $(X_{1,l}, \ldots, X_{n_l,l})$: since a s.r.s. is taken within each stratum (subpopulation), the joint distribution of the data from the stratum $\mathbb{S}_l$ is as that given in LNp.11-12, with F_0 replaced by $F_{0,l}$
  - data from different strata are independent

check Def.19 (LNp.62)

Q: What is the joint dist. of all data? $\prod_{\ell=1}^{L} p(x_{1,\ell}, \cdots, x_{n_\ell,\ell})$

Definition 20 (some intuitive estimators of the parameters of population and stratum)

- subpopulation $\mathbb{S}_l$: since a s.r.s. is taken within each stratum,
  - mean μ_l: estimated by the subsample mean $\overline{X}_l \equiv \frac{1}{n_l}\sum_{k=1}^{n_l} X_{k,l}$
  - total τ_l: estimated by subsample total $T_l \equiv N_l \overline{X}_l$

check Note8 (LNp 28)

- variance σ_l^2: estimated by $s_l^2 \equiv \frac{1}{n_l - 1}\sum_{k=1}^{n_l}(X_{k,l} - \overline{X}_l)^2$ under with replacement, and by $\left(1 - \frac{1}{N_l}\right)s_l^2$ under without replacement

(unbiased) sample variance

- (whole) population Ω: under a stratified random sample,
  - mean μ: estimated by the stratified sample mean

$$\overline{X}_{\mathbb{S}} \equiv \frac{1}{N}\sum_{l=1}^{L} N_l \overline{X}_l = \sum_{l=1}^{L} W_l \overline{X}_l = \frac{1}{N}\sum_{l=1}^{L}\frac{1}{(n_l/N_l)}\left(\sum_{k=1}^{n_l} X_{k,l}\right),$$

since $\mu = \sum_{l=1}^{L} W_l \mu_l$. τ_ℓ stratum fraction $\frac{N_\ell}{N}$ cf.   weights   sampling fraction in S_ℓ

Thm 20 (LNp.61)

(**Note**. $\overline{X}_{\mathbb{S}} \neq \frac{1}{n}\sum_{l=1}^{L}\sum_{k=1}^{n_l} X_{k,l} = \sum_{l=1}^{L}\frac{n_l}{n}\overline{X}_l$ in general, they are equal only when $\frac{n_l}{n} = \frac{N_l}{N}$, $l = 1, \ldots, L$.)

weights

iff $\frac{n}{N} = \frac{n_\ell}{N_\ell}$ sampling fraction ↑ weight ↓

Why should not equal? check the graph in LNp.60

  - total τ $(=N\mu)$: estimated by $T_{\mathbb{S}} \equiv N\,\overline{X}_{\mathbb{S}}$
  - **FYI**. An intuitive estimator of the population variance σ^2 can be developed, based on the relation between σ^2 and μ_l's, σ_l^2's (Thm. 20, LNp.61), by using the estimators $\overline{X}_l$'s and s_l^2's (or $(1 - \frac{1}{N_l})s_l^2$'s).

This explains why in LNp.63 within-stratum variations should be small

Theorem 22 (mean and variance of the stratified estimator of population mean)

- Under stratified random sampling, with or without replacement, $E(\overline{X}_{\mathbb{S}}) = \mu$. (unbiased)
- Under stratified random sampling,
  - with replacement, $Var(\overline{X}_{\mathbb{S}}) = \sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right)$.
  - without replacement, $Var(\overline{X}_{\mathbb{S}}) = \sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right)\left(1 - \frac{n_l - 1}{N_l - 1}\right)$

stratum fraction $\frac{N_\ell}{N}$

$Var(\overline{X}_\ell)$ Note 8, LNp.28

Proof: The expectation of the stratified estimator $\overline{X}_{\mathbb{S}}$ is

Thm20, LNp.61

$$E(\overline{X}_{\mathbb{S}}) = E\left(\sum_{l=1}^{L} W_l\,\overline{X}_l\right) = \sum_{l=1}^{L} W_l\,E(\overline{X}_l) = \sum_{l=1}^{L} W_l\,\mu_l = \mu.$$

∵ s.r.s. in each strata

indep.

Since the data from different strata are independent of one anther, the subsample means $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_L$ are independent random variables, and

$S_1 \to (X_{1,1}, X_{2,1}, \cdots, X_{n_1,1})$
$S_2 \to (X_{1,2}, X_{2,2}, \cdots, X_{n_2,2})$
$S_L \to (X_{1,L}, X_{2,L}, \cdots, X_{n_L,L})$

$$Var(\overline{X}_{\mathbb{S}}) = Var\left(\sum_{l=1}^{L} W_l\,\overline{X}_l\right) = \sum_{l=1}^{L} W_l^2\,Var(\overline{X}_l).$$

Since s.r.s. is taken within each stratum, the results follows from Thm.2 (LNp.17) and Thm.3 (LNp.18) respectively for with and without replacement.

Note 17 (Some notes about the mean and variance of the stratified estimator of μ)

- Under stratified random sampling, $\overline{X}_{\mathbb{S}}$ is an unbiased estimator of μ.
- If the sampling fractions (i.e., (n_l/N_l)'s) within all strata are small, then

with replacement $\approx$ without replacement,

similar probabilities

and

$n_l << N_l,\ l = 1, \cdots, N$

$$\sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right) \approx \sum_{l=1}^{L} W_l^2\left(\frac{\sigma_l^2}{n_l}\right)\left(1 - \frac{n_l-1}{N_l-1}\right).$$

≈ 1

Definition 22 (estimated standard error of the stratified estimator of population mean)

- Under stratified random sampling with replacement, since s_l^2 is an unbiased estimator of σ_l^2, the $Var(\overline{X}_{\mathbb{S}})$ can be estimated by

Def 20, LNp.64

∵ s.r.s in each strata and Thm9 (LNp.24)

unbiased → $$s_{\overline{X}_{\mathbb{S}}}^2 = \sum_{l=1}^{L} W_l^2\left(\frac{s_l^2}{n_l}\right).\ \ S^2_{\overline{X}_l}\ \ \left(\xrightarrow{\sqrt{\ }} s_{\overline{X}_{\mathbb{S}}}\right)$$

← estimated standard error of $\overline{X}_{\mathbb{S}}$

- Under stratified random sampling without replacement, since $(1 - \frac{1}{N_l})s_l^2$ is an unbiased estimator of σ_l^2, the $Var(\overline{X}_{\mathbb{S}})$ can be estimated by

∵ s.r.s in each strata and Thm9 (LNp.24)

unbiased → $$s_{\overline{X}_{\mathbb{S}}}^2 = \sum_{l=1}^{L} W_l^2\left(\frac{s_l^2}{n_l}\right)\left(1 - \frac{1}{N_l}\right)\left(1 - \frac{n_l - 1}{N_l - 1}\right)$$

same notation as with repl., but different formula

$$\left(\frac{N_l}{N}\right)^2 \quad = \sum_{l=1}^{L} W_l^2\left(\frac{s_l^2}{n_l}\right)\left(1 - \frac{n_l}{N_l}\right).\ \ S^2_{\overline{X}_l}\ \ \left(\xrightarrow{\sqrt{\ }} s_{\overline{X}_{\mathbb{S}}}\right)$$

sampling fraction in S_l

estimated standard error of $\overline{X}_{\mathbb{S}}$

Theorem 23 (mean and variance of the stratified estimator of population total)

Since $T_{\mathbb{S}} = N\,\overline{X}_{\mathbb{S}}$, we have $E(T_{\mathbb{S}}) = N\,E(\overline{X}_{\mathbb{S}})$ and $Var(T_{\mathbb{S}}) = N^2\,Var(\overline{X}_{\mathbb{S}})$.

- $E(T_{\mathbb{S}}) = N\,\mu = \tau$, i.e., $T_{\mathbb{S}}$ is an unbiased estimator of τ
- $$Var(T_{\mathbb{S}}) = \begin{cases} \sum_{l=1}^{L} N_l^2\left(\frac{\sigma_l^2}{n_l}\right), & \text{if with replacement,} \\ \sum_{l=1}^{L} N_l^2\left(\frac{\sigma_l^2}{n_l}\right)\left(1 - \frac{n_l-1}{N_l-1}\right), & \text{if without replacement,} \end{cases}$$

$N^2 W_l^2$

with / without

Note. The $Var(T_{\mathbb{S}})$ can be estimated by $s_{T_{\mathbb{S}}}^2 \equiv N^2\,s_{\overline{X}_{\mathbb{S}}}^2$. $\left(\xrightarrow{\sqrt{\ }} s_{T_{\mathbb{S}}} = N\,s_{\overline{X}_{\mathbb{S}}}\right)$

Example 20 (stratified random sampling, cont. Ex.17 in LNp.53)

∵ correlated with discharge

- Consider the population of 393 hospitals.

extra information (useful one)

- Assume that the number of beds in each hospital is known, and 4 strata are determined by the number of beds from small to large:

×: unknown in sampling survey

Stratum	N_l	$W_l = N_l/N$	μ_l increasing	$\sigma_l < \sigma$
A	98	0.249	×182.9	×103.4 ← smallest
B	98	0.249	×526.5	×204.8
C	98	0.249	×956.3	×243.5
D	99	0.252	×1591.2	×419.2 ← largest

N_l: about equal. W_l: about equal.

good strata? Check "Recall" in LNp.63

- For a without-replacement stratified random sample of size n, suppose we choose $n_1 = n_2 = n_3 = n_4 = n/4$. Neglecting the finite population correction, we have

①∵ $n_l \ll N_l$, $n \ll N$
②∵ $1-\frac{n_l-1}{N_l-1} \approx 1-\frac{n-1}{N-1}$

cf. Thm22, LNp.65 $\quad 1/n_l$

$$\sigma_{\overline{X}_S} = \sqrt{\frac{4}{n}\sum_{l=1}^{4} W_l^2 \sigma_l^2} = \frac{268.4}{\sqrt{n}}.$$

- For a without-replacement s.r.s. of size n, neglecting the finite population correction, we have (see Ex.4, LNp.20)

population variance $\sigma^2 = 589.7^2$ ← Thm3, LNp.18

$$\sigma_{\overline{X}} = \frac{589.7}{\sqrt{n}}.$$

- Note that the stratification has resulted in a tremendous gain in precision: $\sigma_{\overline{X}_S} \approx 0.455 \times \sigma_{\overline{X}} \Rightarrow \sigma^2_{\overline{X}_S}/\sigma^2_{\overline{X}} = 0.207$. The stratified estimator $\overline{X}_S$ based on a total sample size of $n/5$ is as precise as $\overline{X}$ based on a s.r.s. of size n. (**cf.** the reduction in variance due to stratification is comparable to that achieved by using a ratio estimator given in Ex.18, LNp.58). LNp.57

• Methods of allocation in stratified random sampling

- Q: Why and when can a stratification produce a dramatic improvement in precision?

• Why? ∵ exclude many biased samples
• When? i.e., n_l's = ? strata = ?