Introduction to Sampling Theory

Lecture 14
Stratified Random Sampling

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Proportional Allocation:

For fixed *k*, select *n_i* such that it is proportional to stratum size *N_i*.

$$n_i \propto N_i$$

$n_i = \delta N_i$ **where δ is the constant of proportionality**

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \delta N_i$$

or $n = \delta N$

$$\Rightarrow \delta = \frac{n}{N}.$$

Thus $n_i = \left(\dfrac{n}{N}\right) N_i.$

Such allocation arises from the considerations like operational convenience.

Neyman or Optimum Allocation:

This allocation considers the size of strata as well as variability

$$n_i \propto N_i S_i$$

$$n_i = \delta^* N_i S_i$$

$n_i = \delta^* N_i S_i$ **where δ^* is the constant of proportionality**

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \delta^* N_i S_i$$

or $n = \delta^* \sum_{i=1}^{k} N_i S_i$

or $\delta^* = \dfrac{n}{\sum_{i=1}^{k} N_i S_i}.$

Thus $n_i = \dfrac{n N_i S_i}{\sum_{i=1}^{k} N_i S_i}, i = 1, 2, \ldots, k.$

Variances Under Proportional Allocation:

Under proportional allocation,

$$n_i = \frac{n}{N} N_i$$

and

$$Var(\bar{y})_{st} = \sum_{i=1}^{k} \left(\frac{N_i - n_i}{N_i n_i} \right) w_i^2 S_i^2$$

$$Var_{prop}(\bar{y})_{st} = \sum_{i=1}^{k} \left(\frac{N_i - \frac{n}{N} N_i}{N_i \frac{n}{N} N_i} \right) \left(\frac{N_i}{N} \right)^2 S_i^2$$

$$= \frac{N-n}{Nn} \sum_{i=1}^{k} \frac{N_i S_i^2}{N}$$

$$= \frac{N-n}{Nn} \sum_{i=1}^{k} w_i S_i^2 .$$

Variances Under Optimum Allocation:

$$\boxed{n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}, \; i = 1,2,\ldots,k}$$

Under optimum allocation,

$$V_{opt}(\overline{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)w_i^2S_i^2 = \sum_{i=1}^{k}\frac{w_i^2S_i^2}{n_i} - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[w_i^2S_i^2\left(\frac{\sum_{i=1}^{k}N_iS_i}{nN_iS_i}\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[\frac{1}{n}.\frac{N_iS_i}{N^2}\left(\sum_{i=1}^{k}N_iS_i\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2 - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}w_iS_i^2.$$

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

$$V_{SRS}(\bar{y}) = \frac{N-n}{Nn} S^2$$

$$V_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn} \sum_{i=1}^{k} \frac{N_i S_i^2}{N}.$$

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} w_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_i S_i^2 .$$

Assume that N_i is large enough to permit the approximation

$\frac{N_i - 1}{N_i} \approx 1$ **and** $\frac{N-1}{N} \approx 1.$

$$Var_{opt}(\bar{y}_{st}) \le Var_{prop}(\bar{y}_{st}) \le Var_{SRS}(\bar{y}).$$

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

(a) Proportional allocation:

$$V_{SRS}(\bar{y}) = \frac{N-n}{Nn} S^2$$

$$V_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn} \sum_{i=1}^{k} \frac{N_i S_i^2}{N}.$$

In order to compare $V_{SRS}(\bar{y})$ **and** $V_{prop}(\bar{y}_{st})$, **first we attempt to express** S^2 **as a function of** S_i^2 **.**

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

Consider

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(Y_{ij}-\bar{Y}_i)+(\bar{Y}_i-\bar{Y})\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{N_i}(\bar{Y}_i-\bar{Y})^2$$

$$= \sum_{i=1}^{k}(N_i-1)S_i^2+\sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2$$

$$\frac{N-1}{N}S^2 = \sum_{i=1}^{k}\frac{N_i-1}{N}S_i^2+\sum_{i=1}^{k}\frac{N_i}{N}(\bar{Y}_i-\bar{Y})^2.$$

For simplification, we assume that *N*$_i$ is large enough to permit the approximation

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

For simplification, we assume that *N*N_i is large enough to permit the approximation $\frac{N_i - 1}{N_i} \approx 1$ **and** $\frac{N-1}{N} \approx 1$

Thus

$$S^2 = \sum_{i=1}^{k} \frac{N_i}{N} S_i^2 + \sum_{i=1}^{k} \frac{N_i}{N} (\bar{Y}_i - \bar{Y})^2$$

Premultiply by $\frac{N - n}{Nn}$ **on both sides**

or $\frac{N-n}{Nn} S^2 = \frac{N-n}{Nn} \sum_{i=1}^{k} \frac{N_i}{N} S_i^2 + \frac{N-n}{Nn} \sum_{i=1}^{k} \frac{N_i}{N} (\bar{Y}_i - \bar{Y})^2$

$$Var_{SRS}(\bar{y}) = V_{prop}(\bar{y}_{st}) + \frac{N-n}{Nn} \sum_{i=1}^{k} w_i (\bar{Y}_i - \bar{Y})^2$$

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

$$Var_{SRS}(\bar{y}) = V_{prop}(\bar{y}_{st}) + \frac{N-n}{Nn}\sum_{i=1}^{k} w_i(\bar{Y}_i - \bar{Y})^2$$

Since $\sum_{i=1}^{k} w_i(\bar{Y}_i - \bar{Y})^2 \geq 0,$

$$\Rightarrow Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y}).$$

A larger gain in the difference is achieved when $\bar{Y}_i$ **differs from** $\bar{Y}$ **more.**

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

(b) Optimum allocation

$$Var_{opt}(\overline{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} w_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_i S_i^2.$$

Consider

$$Var_{prop}(\overline{y}_{st}) - Var_{opt}(\overline{y}_{st}) = \left[\left(\frac{N-n}{Nn}\right)\sum_{i=1}^{k} w_i S_i^2\right] - \left[\frac{1}{n}\left(\sum_{i=1}^{k} w_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_i S_i^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} w_i S_i^2 - \left(\sum_{i=1}^{k} w_i S_i\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{k} w_i\, S_i^2 - \frac{1}{n}\overline{S}^2$$

$$= \frac{1}{n}\sum_{i=1}^{k} w_i (S_i - \overline{S})^2 \quad \textbf{where } \overline{S} = \sum_{i=1}^{k} w_i S_i.$$

Comparison of Variances of Sample Mean Under SRS with Stratified Mean under Proportional and Optimal Allocation:

$$Var_{prop}(\bar{y}_{st}) - Var_{opt}(\bar{y}_{st}) = \frac{1}{n}\sum_{i=1}^{k} w_i (S_i - \bar{S})^2$$

$$\Rightarrow Var_{prop}(\bar{y}_{st}) - Var_{opt}(\bar{y}_{st}) \geq 0$$

or $Var_{opt}(\bar{y}_{st}) \leq Var_{prop}(\bar{y}_{st}).$

The larger gain in efficiency is achieved when *S* S_i **differs from** $\bar{S}$ **more.**

Combining the results in (a) and (b), we have

$$Var_{opt}(\bar{y}_{st}) \leq Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y})$$

Confidence Intervals of Population Mean:

Assume $\bar{y}_{st}$ is normally distributed, and $\sqrt{\widehat{Var}(\bar{y}_{st})}$ is well determined so that *t* can be read from normal distribution tables.

If only few degrees of freedom are provided by each stratum, then *t* values are obtained from the table of student's *t*-distribution. The confidence limits of $\bar{Y}$ can be obtained as

$$\bar{y}_{st} \pm t\sqrt{\widehat{Var}(\bar{y}_{st})}$$

Confidence Intervals of Population Mean:

The distribution of $\sqrt{\widehat{Var}(\overline{y}_{st})}$ **is generally complex.**

Assume $\boldsymbol{y_{ij}}$**'s are normally distributed.**

An approximate method of assigning an effective number of degrees of freedom $\boldsymbol{n_e}$ **to** $\sqrt{\widehat{Var}(\overline{y}_{st})}$ **is**

$$n_e = \frac{\left(\sum_{i=1}^{k} g_i s_i^2\right)^2}{\sum_{i=1}^{k} \frac{g_i^2 s_i^4}{n_i - 1}}$$

$$g_i = \frac{N_i(N_i - n_i)}{n_i}$$

$$Min(n_i - 1) \le n_e \le \sum_{i=1}^{k} (n_i - 1)$$

Modification of Optimal Allocation:

Sometimes in the optimal allocation, the size of subsample exceeds the stratum size. In such a case,

replace *n_i* by *N_i*

and re-compute the rest of *n_i* ‘s by the revised allocation.

For example, if *$n_i > N_i$*,

then take the revised *n_i* ‘s as

$$\tilde{n}_1 = N_1$$

and

$$\tilde{n}_i = \frac{(n - N_1)w_i S_i}{\sum_{i=2}^{k} w_i S_i} \; ; \quad i = 2, 3, \ldots, k$$

provided $\tilde{n}_i \le N_i$ for all *i* = 2,3,…,*k*.

Modification of Optimal Allocation:

Suppose in revised allocation, we find that $\tilde{n}_2 > N_2$ then the revised allocation would be

$$\tilde{n}_1 = N_1$$

$$\tilde{n}_2 = N_2$$

$$\tilde{n}_i = \frac{(n - N_1 - N_2)w_i S_i}{\sum_{i=3}^{k} w_i S_i}; i = 3, 4, \ldots, k.$$

provided $\tilde{n}_i < N_i$ for all *i* = 3, 4,…, *k.*

We continue this process until every $\tilde{n}_i < N_i$.

Modification of Optimal Allocation:

In such cases, the formula for the minimum variance of $\bar{y}_{st}$ need to be modified as

$$Min\ Var(\bar{y}_{st}) = \frac{(\sum^{*} w_i S_i)^2}{n^*} - \frac{\sum^{*} w_i S_i^2}{N}$$

where $\sum^{*}$ denotes the summation over the strata in which $\tilde{n}_i \leq N_i$ and *n is the revised total sample size in the strata.**

Estimation of the Gain in Precision due to Stratification:

An obvious question crops up that what is the advantage of stratifying a population in the sense that instead of using SRS, the population is divided into various strata?

This is answered by estimating the variance of estimators of population mean under SRS (without stratification) and stratified sampling by evaluating

$$\frac{\widehat{Var}_{SRS}(\bar{y})-\widehat{Var}(\bar{y}_{st})}{\widehat{Var}(\bar{y}_{st})}.$$

This gives an idea about the gain in efficiency due to stratification.

Estimation of the Gain in Precision due to Stratification:

Since

$$Var_{SRS}(\overline{y}) = \frac{N-n}{Nn} S^2,$$

so there is a need to express S^2 in terms of S_i^2.

How to estimate S^2 based on a stratified sample?

Estimation of the Gain in Precision due to Stratification:

Consider

$$
\begin{aligned}
(N-1)S^2 &= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(Y_{ij}-\bar{Y}_i)+(\bar{Y}_i-\bar{Y})\right]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2+\sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 \\
&= \sum_{i=1}^{k}(N_i-1)S_i^2+\sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 \\
&= \sum_{i=1}^{k}(N_i-1)S_i^2+N\left[\sum_{i=1}^{k}w_i\bar{Y}_i^{\,2}-\bar{Y}^{\,2}\right].
\end{aligned}
$$

In order to estimate *S*², we need to estimates of S_i^2, $\bar{Y}_i^{\,2}$ **and** $\bar{Y}^{\,2}$.

We consider their estimation one by one.

Estimation of the Gain in Precision due to Stratification:

(I) For an estimate of S_i^2, we have

$$E(s_i^2) = S_i^2$$

So $\hat{S}_i^2 = s_i^2$.

Estimation of the Gain in Precision due to Stratification:

(II) For estimate of $\overline{Y}_i^2$, we know

$$Var(\overline{y}_i) = E(\overline{y}_i^2) - [E(\overline{y}_i)]^2$$

$$= E(\overline{y}_i^2) - \overline{Y}_i^2$$

or $\overline{Y}_i^2 = E(\overline{y}_i^2) - Var(\overline{y}_i).$

An unbiased estimate of $\overline{Y}_i^2$ is

$$\hat{\overline{Y}}_i^2 = \overline{y}_i^2 - \widehat{Var}(\overline{y}_i)$$

$$= \overline{y}_i^2 - \left(\frac{N_i - n_i}{N_i n_i}\right) s_i^2 .$$

Estimation of the Gain in Precision due to Stratification:

(III) For the estimation of $\bar{Y}^2$, **we know**

$$\begin{aligned} Var(\bar{y}_{st}) &= E(\bar{y}_{st}^2) - [E(\bar{y}_{st})]^2 \\ &= E(\bar{y}_{st}^2) - \bar{Y}^2 \\ \Rightarrow \bar{Y}^2 &= E(\bar{y}_{st}^2) - Var(\bar{y}_{st}) \end{aligned}$$

So, an estimate of $\bar{Y}^2$, **is**

$$\begin{aligned} \hat{\bar{Y}}^2 &= \bar{y}_{st}^2 - \widehat{Var}(\bar{y}_{st}) \\ &= \bar{y}_{st}^2 - \sum_{i=1}^{k} \left(\frac{N_i - n_i}{N_i n_i} \right) w_i^2 s_i^2 . \end{aligned}$$

Estimation of the Gain in Precision due to Stratification:

Substituting these estimates in the expression (n-1)S^2 as follows, the estimate of S^2 is obtained as

$$(N-1)S^2 = \sum_{i=1}^{k}(N_i-1)S_i^2 + N\left[\sum_{i=1}^{k} w_i \bar{Y}_i^{\,2} - \bar{Y}^{\,2}\right].$$

Then

$$\hat{S}^2 = \frac{1}{N-1}\sum_{i=1}^{k}(N_i-1)\hat{S}_i^2 + \frac{N}{N-1}\left[\sum_{i=1}^{k} w_i \hat{\bar{Y}}_i^{\,2} - \hat{\bar{Y}}^{\,2}\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{k}\left(N_i-1\right)s_i^2\right] + \frac{N}{N-1}\left[\left(\sum_{i=1}^{k} w_i\left(\bar{y}_i^2 - \left(\frac{N_i-n_i}{N_i n_i}\right)s_i^2\right)\right) - \left(\bar{y}_{st}^2 - \sum_{i=1}^{k}\frac{N_i-n_i}{N_i n_i} w_i^2 s_i^2\right)\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{k}\left(N_i-1\right)s_i^2\right] + \frac{N}{N-1}\left[\sum_{i=1}^{k} w_i(\bar{y}_i - \bar{y}_{st})^2 - \sum_{i=1}^{k} w_i(1-w_i)\frac{N_i-n_i}{N_i n_i}s_i^2\right].$$

Estimation of the Gain in Precision due to Stratification:

Thus

$$\widehat{Var}_{SRS}(\bar{y}) = \frac{N-n}{Nn}\hat{S}^2$$

$$= \frac{N-n}{N(N-1)n}\left[\sum_{i=1}^{k}(N_i-1)s_i^2\right] + \frac{N(N-n)}{nN(N-1)}\left[\sum_{i=1}^{k} w_i(\bar{y}_i-\bar{y}_{st})^2 - \sum_{i=1}^{k} w_i(1-w_i)\frac{N_i-n_i}{N_i n_i}s_i^2\right]$$

and

$$\widehat{Var}(\bar{y}_{st}) = \sum_{i=1}^{k}\frac{N_i-n_i}{N_i n_i}w_i^2 s_i^2 .$$

Substituting these expressions in

$$\frac{\widehat{Var}_{SRS}(\bar{y}) - \widehat{Var}(\bar{y}_{st})}{\widehat{Var}(\bar{y}_{st})},$$

the gain in efficiency due to stratification can be obtained.

If any other particular allocation is used, then substituting the appropriate *n_i* under that allocation, such gain can be estimated.

Stratified Sampling for Proportions:

If the characteristic under study is qualitative in nature, then its values will fall into one of the two mutually exclusive complimentary classes *C* and *C'*.

Ideally, only two strata are needed in which all the units can be divided depending on whether they belong to *C* or its complement *C'*.

Thus is difficult to achieve in practice.

So the strata are constructed such that the proportion in *C* varies as much as possible among strata.

Stratified Sampling for Proportions:

Define an indicator variable

$$Y_{ij} = \begin{cases} 1 & \textbf{when } j^{th} \textbf{ unit belongs to the } i^{th} \textbf{ stratum is in } C \\ 0 & \textbf{otherwise} \end{cases}$$

and $\bar{y}_{st} = p_{st}$.

Let

$P_i = \dfrac{A_i}{N_i}$: **Proportion of units in *C* in the *i^th^* stratum**

$p_i = \dfrac{a_i}{n_i}$: **Proportion of units in *C* in the sample from the *i^th^* stratum**

Stratified Sampling for Proportions:

An estimate of population proportion based on the stratified sampling is

$$p_{st} = \sum_{i=1}^{k} \frac{N_i p_i}{N}$$

Here $S_i^2 = \frac{N_i}{N_i - 1} P_i Q_i$ **where** $Q_i = 1 - P_i$.

Also $Var(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i - n_i}{N_i n_i} w_i^2 S_i^2.$

So $Var(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^2 (N_i - n_i)}{N_i - 1} \frac{P_i Q_i}{n_i}.$

If the finite population correction can be ignored, then

$$Var(p_{st}) = \sum_{i=1}^{k} w_i^2 \frac{P_i Q_i}{n_i}$$

Stratified Sampling for Proportions:

If the proportional allocation is used for *n_i*, then the variance of *p_{st}* is

$$Var_{prop}(p_{st}) = \frac{N-n}{N}\frac{1}{Nn}\sum_{i=1}^{k}\frac{N_i^2 P_i Q_i}{N_i - 1}$$

$$= \frac{N-n}{Nn}\sum_{i=1}^{k} w_i P_i Q_i$$

and its estimate is

$$\widehat{Var}_{prop}(p_{st}) = \frac{N-n}{Nn}\sum_{i=1}^{k} w_i \frac{p_i q_i}{n_i - 1}.$$

Post Stratification:

Sometimes the stratum to which a unit belongs may be known after the field survey only.

For example, the age of persons, their educational qualifications etc. can not be known in advance. In such cases, we adopt the post-stratification procedure to increase the precision of the estimates.

Post Stratification:

In post-stratification,

- **draw a sample by simple random sampling from the population and carry out the survey.**
- **After the completion of the survey, stratify the sampling units to increase the precision of the estimates.**

Note*: *This topic will be covered after the ratio method of estimation.