Sampling Design Notes

Jack H

June 5, 2025

Introduction

Coefficient of Variation

$$\mathrm{cv} = \frac{\sqrt{\mathrm{Var}(\hat{\theta})}}{E(\hat{\theta})}$$

It measures the density of given data.

Mean Square Error

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{bias}(\hat{\theta}))^2$$

Point Estimation Error Control: MSE(V,C)
When $\hat{\theta}$ is unbiased, $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$, for a given upper bound of variance V, or a given upper bound of $\mathrm{cv}: C$, such that $V \geq \mathrm{var}(\hat{\theta})$ or $C \geq \mathrm{cv}(\hat{\theta})$

Margin of Error:
If $\hat{\theta}$ is a point estimator of θ, for a given $\alpha \in (0, 1)$, if

$$P(|\hat{\theta} - \theta| \leq d) = 1 - \alpha$$

we call d the margin of error of $\hat{\theta}$ at confidence level α.
Relative Margin of Error:
If

$$P(\frac{|\hat{\theta} - \theta|}{\theta} \leq r) = 1 - \alpha$$

we call r the relative margin of error of $\hat{\theta}$ at confidence level α.

Error Limit (d,r) Estimation Control:
For a given $\alpha \in (0,1)$, for given absolute margin of error d or relative margin of error r, such that $P(|\hat{\theta}-\theta| \leq d) = 1-\alpha$ or $P(\frac{|\hat{\theta}-\theta|}{\theta} \leq r) = 1-\alpha$.

Assumptions in this course:

1. Consistent Estimation: $\hat{\theta}_n \xrightarrow{P} \theta, n \to \infty$
   Definition: If $\hat{\theta}$ is a consistent estimator of θ, then for any $\epsilon > 0$,

$$P(|\hat{\theta}-\theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

2. Asymptotically Normal Distribution CAN: $\frac{\hat{\theta}_n - E(\hat{\theta})}{\sqrt{var(\hat{\theta}_n)}} \xrightarrow{d} N(0,1)$
   Denote $\sqrt{var(\hat{\theta}_n)}$ as $\mathrm{sd}(\hat{\theta}_n)$

You need to prove UE or AUE in this course
Theorem: When $\hat{\theta}$ is consistent and asymptotically normal, if $\hat{\theta}$ is an unbiased estimator(UE) or asymptotically unbiased estimator(AUE), then the distribution of $\hat{\theta}$ is approximately normal.

$$\frac{\hat{\theta}-\theta}{\sqrt{var(\hat{\theta})}} = \frac{\hat{\theta}-\theta}{\mathrm{sd}(\hat{\theta})} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty$$

Therefore

$$P\left(\frac{\hat{\theta}-\theta}{\mathrm{sd}(\hat{\theta})} \leq z_{\alpha/2}\right) = 1-\alpha$$

$d = z_{\alpha/2}\mathrm{sd}(\hat{\theta}), r = z_{\alpha/2}\frac{\mathrm{sd}(\hat{\theta})}{\theta}$

$\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\mathrm{sd}(\hat{\theta}), \hat{\theta}_R = \hat{\theta} + z_{\alpha/2}\mathrm{sd}(\hat{\theta})$
If $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R) = 1-\alpha$, then $\hat{\theta}_L$ and $\hat{\theta}_R$ are the endpoints of a $(1-\alpha)$ confidence interval for θ.
When n is large,

$$P(\theta \in [\hat{\theta} \pm z_{\alpha/2}\mathrm{sd}(\hat{\theta})]) \approx 1-\alpha$$

In `ci.r`

```
conf.interval=function(para.hat, SD.hat, alpha)
```

Simple Random Sampling

In this cource, we consider picking n units out of a population of N without replacement, each pick has probability $p = 1/C_N^n$

In `srs sampling.r`

```
## simple random sampling without replacement
mysrs=sample(1:N, n)
print(mysrs)
## simple random sampling with replacement
mysrs=sample(1:N, n, replace = TRUE)
print(mysrs)
```

Mean :

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Total:

$$Y_T = N\bar{Y}$$

Variance:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Estimation of Population Mean $\bar{Y}$

1. **Point Estimation**

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

2. **Unbiased estimator of $\bar{Y}$**

$$\mathrm{E}(\bar{y}) = \bar{Y} \quad (UE)$$

3. **Variance of estimation:**

$$Var(\bar{y}) = \frac{1-f}{n} S^2$$

   where $f = \frac{n}{N}$ and S^2 is the variance of population Y (unknown)

4. **Estimation of variance:**

$$\hat{\mathrm{Var}}(\bar{y}) = \frac{1-f}{n} s^2$$

   where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

5. **Confidence Interval**

$$\left[\bar{y} \pm z_{\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\bar{y})}\right]$$

$$d = \sqrt{\widehat{\mathrm{Var}}(\bar{y})}$$

$$r = \frac{d}{\bar{y}}$$

In `srs.r`

```
srs.mean=function(N, mysample, alpha)
```

Proof of $E(s^2) = S^2$

Step 1: Express S^2 and s^2

The population variance S^2 is defined as:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

The sample variance s^2 is:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Step 2: Expand the Sum of Squares

First, note that:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Step 3: Take the Expectation of s^2

Compute $E(s^2)$:

$$E(s^2) = E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(y_i^2) - nE(\bar{y}^2)\right]$$

Step 4: Compute $E(y_i^2)$ and $E(\bar{y}^2)$

For any y_i:

$$E(y_i^2) = \mathrm{Var}(y_i) + [E(y_i)]^2 = S^2\left(1 - \frac{1}{N}\right) + \bar{Y}^2$$

For $\bar{y}$:

$$E(\bar{y}^2) = \text{Var}(\bar{y}) + [E(\bar{y})]^2 = \frac{1-f}{n}S^2 + \bar{Y}^2$$

Where $f = \frac{n}{N}$.

Step 5: Substitute Back into $E(s^2)$

$$E(s^2) = \frac{1}{n-1}\left[n\left(S^2\left(1-\frac{1}{N}\right)+\bar{Y}^2\right) - n\left(\frac{1-f}{n}S^2+\bar{Y}^2\right)\right]$$

Simplify the expression:

$$E(s^2) = \frac{1}{n-1}\left[nS^2\left(1-\frac{1}{N}\right) - (1-f)S^2\right]$$

$$= \frac{1}{n-1}\left[nS^2 - \frac{n}{N}S^2 - S^2 + \frac{n}{N}S^2\right]$$

$$= \frac{1}{n-1}\left[(n-1)S^2\right] = S^2$$

Conclusion

Thus, we have shown that:

$$E(s^2) = S^2$$

Estimation of Population Total $Y_T = N\bar{Y} = \sum_{i=1}^{N} Y_i$

1. **Point Estimation**

$$\hat{y}_T = N\bar{y}$$

2. **Unbiased Estimator**

$$\mathrm{E}(\hat{y}_T) = N \cdot \mathrm{E}(\bar{y}) = N\bar{Y} = Y_T$$

3. **Variance of Estimation**

$$Var(\hat{y}_T) = N^2 Var(\bar{y}) = N^2 \frac{1-f}{n} S^2$$

4. **Estimation of Variance**

$$\hat{V}ar(\hat{Y}_T) = N^2 \frac{1-f}{n} s^2$$

5. **Confidence Interval**

$$\left[\hat{y}_T \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{y}_T)}\right]$$

$$d = z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{y}_T)}$$

$$r = \frac{d}{\hat{y}_T}$$

In `srs.r`

```
srs.total=function(N, mysample, alpha)
```

Estimation of Population Proportion P

Define:

- Population Proportion $P = \frac{1}{N} \sum_{i=1}^{N} Y_i = \bar{Y}$
- Population Total $A = \sum Y_i = NP$
- Population Variance

$$S^2 = \frac{N}{N-1} P(1-P) = \frac{N}{N-1} PQ \quad \text{where } Q = 1-P$$

Let the observed $y_1, \ldots, y_n$ have property with count a

1. $\hat{p} = \bar{y} = \frac{a}{n}$
2. UE
3. Variance of Estimation

$$Var(\hat{p}) = \frac{1-f}{n}(\frac{N}{N-1}PQ)$$

4. Estimation of Variance

$$\hat{V}ar(\hat{p}) = \frac{1-f}{n-1}\hat{p}\hat{q}$$

In `srs.r`

```
srs.prop=function(N=NULL, n, event.num, alpha)
```

Estimation of Population total A

1. $\hat{A} = N\bar{y} = N\hat{p}$
2. UE
3. $$Var(\hat{A}) = N^2\frac{1-f}{n}\frac{N}{N-1}PQ$$
4. $$\hat{V}ar(\hat{A}) = N^2\frac{1-f}{n}\frac{n}{n-1}\hat{p}\hat{q}$$

In `srs.r`

```
srs.num=function(N=NULL, n, event.num, alpha)
```

Determining the Sample size

The sample size is determined by the accuracy needed

$$(V, C, d, r) \implies n_{\min}$$

V: Variance upper bound

C: CV upper bound

d: Error upper bound

r: Relative error upper bound

Sample Size $n_{\min}$ for Estimating Population Mean $\bar{Y}$

Step 1 Calculate n_0
Here S^2 and $\bar{Y}$ are given from historical data.

$$n_0 = \frac{S^2}{V} = \begin{cases} \frac{S^2}{V} & V = V \\ \frac{S^2}{C^2 \bar{Y}^2} & C = \sqrt{V} / \bar{Y} \\ \frac{z_{\alpha/2}^2 S^2}{d^2} & d = z_{\alpha/2} \sqrt{V} \\ \frac{z_{\alpha/2}^2 S^2}{r^2 \bar{Y}^2} & r = z_{\alpha/2} \sqrt{V} / \bar{Y} \end{cases}$$

Step 2

$$n_{\min} = \begin{cases} \frac{n_0}{1+\frac{N_0}{N}} & \text{given reasonable } N \\ n_0 & \text{when } N \text{ is very big} \end{cases}$$

In `srs size.r`

```
size.mean=function(N=NULL, Mean.his=NULL, Var.his, method, bound, alpha)
```

Sample Size for Estimating Proportion P

Here P and $Q = 1 - P$ are given from historical data.

$$n_0 = \frac{PQ}{V} = \begin{cases} \frac{PQ}{V} \\ \frac{Q}{C^2 P} \\ \frac{z_{\alpha/2}^2 PQ}{d^2} \\ \frac{z_{\alpha/2}^2 Q}{r^2 P} \end{cases}$$

$$n_{\min} = \begin{cases} \frac{n_0}{1+\frac{n_0 - 1}{N}} & \text{Given } N \\ n_0 & N >> n_0 \end{cases}$$

In `srs size.r`

```
size.prop=function(N=NULL, Prop.his, method, bound, alpha)
```

Sample Size for Estimating Population Total Y_T

Use `size.mean` and adjust inputs

Apply the **Sample Number $n_{\min}$ for Estimating Population Mean $\bar{Y}$** Methods

Bounding Total is the same as bounding $\bar{Y}$ with different bounds:

$$Var(\hat{y}_T) \le V \iff Var(\bar{y}) \le \frac{V}{N^2}$$

$$CV(\hat{y}_T) \le C \iff CV(\bar{y}) \le C$$

$$\text{Error}(\hat{y}_T) \le d \iff \text{Error}(\bar{y}) \le \frac{d}{N}$$

$$\text{Absolute Error}(\hat{y}_T) \le r \iff \text{Absolute Error}(\bar{y}) \le r$$

Stratified Random Sampling

Stratified Random Sampling Formulas

Concept	Population ($Y_{h1}, \ldots, Y_{hN_h}$)	Sample ($y_{h1}, \ldots, y_{hn_h}$)
Size (Size)	N_h ($\sum_{h=1}^{L} N_h = N$)	n_h ($\sum_{h=1}^{L} n_h = n$)
Mean	$\overline{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$	$\overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$
Variance	$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \overline{Y}_h)^2$	$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2$
Stratum weight	$W_h = \frac{N_h}{N}$	$f_h = \frac{n_h}{N_h}$

Esimation of Population Mean $\bar{Y}$

1. $$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

2. $$E(\bar{y}_{st}) = \bar{Y}$$

3. $$Var(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} S_h^2$$

4. $$\hat{V}ar(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h} s_h^2$$

See `stratified mean.r`

```
stra.srs.mean1=function(Nh, nh, yh, s2h, alpha)
stra.srs.mean2=function(Nh, mysample, stra.index, alpha)
```

Estimation of Population Total Y_T

1. **Total Estimate:**

$$\hat{y}_{st} = N \cdot \bar{y}_{st} = N \left(\sum_{h=1}^{L} W_h \bar{y}_h \right)$$

2. **Expected Value of the Estimator:**

$$E(\hat{y}_{st}) = \hat{Y}$$

3. **Variance of the Estimator:**

$$\mathrm{Var}(\hat{y}_{st}) = N^2 \left(\sum_{h=1}^{L} W_h^2 \cdot \frac{1-f_h}{n_h} \cdot S_h^2 \right)$$

4. **Estimated Variance:**

$$\widehat{\mathrm{Var}}(\hat{y}_{st}) = N^2 \left(\sum_{h=1}^{L} W_h^2 \cdot \frac{1-f_h}{n_h} \cdot s_h^2 \right)$$

See `stratified mean.r`

Estimation of Proportion

Symbol	Population ($Y_{h1}, \ldots, Y_{hN_h}$)	Sample ($y_{h1}, \ldots, y_{hn_h}$)
Size	N_h ($N = \sum_{h=1}^{L} N_h$)	n_h ($n = \sum_{h=1}^{L} n_h$)
Proportion with attribute	A_h	a_h
Proportion	$P_h = \frac{A_h}{N_h}$	$\hat{p}_h = \frac{a_h}{n_h}$
Variance	$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$	$s_h^2 = \frac{n_h}{n_h - 1} \hat{p}_h \hat{q}_n$
Weight	$W_h = \frac{N_h}{N}$	$f_h = \frac{n_h}{N_h}$

Stratified Sampling Estimation of Population Proportion P

1. **Estimator for Population Proportion**:

$$\hat{p}_{st} = \sum_{h=1}^{L} W_h \hat{p}_h = \sum_{h=1}^{L} W_h \cdot \frac{a_h}{n_h}$$

2. **Expected Value**:

$$E(\hat{p}_{st}) = P$$

3. **Variance**:

$$Var(\hat{p}_{st}) = \sum_{h=1}^{L} W_h^2 Var(\hat{p}_h) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_h}{n_h} \cdot \frac{N_h}{N_h - 1} P_h Q_h \right)$$

4. **Estimated Variance**:

$$\widehat{Var}(\hat{p}_{st}) = \sum_{h=1}^{L} W_h^2 \widehat{Var}(\hat{p}_h) = \sum_{h=1}^{L} W_h^2 \left(\frac{1 - f_h}{n_h} \cdot \frac{n_h}{n_h - 1} \hat{p}_h \hat{q}_h \right)$$

5. **Confidence Interval (CI)**:

$$CI \quad d.r.$$

Stratified Sampling Estimation for Total A

1. **Estimator for Population Total**:

$$\hat{A}_{st} = N\left(\sum_{h=1}^{L} W_h \hat{p}_h\right) = \sum_{h=1}^{L} W_h \cdot \frac{a_h}{n_h}$$

2. **Expected Value**:

$$E(\hat{A}_{st}) = A$$

3. **Variance**:

$$Var(\hat{A}_{st}) = N^2 \sum_{h=1}^{L} W_h^2 Var(\hat{p}_h) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \cdot \frac{N_h}{N_h - 1} P_h Q_h\right)$$

4. **Estimated Variance**:

$$\widehat{Var}(\hat{A}_{st}) = N^2 \sum_{h=1}^{L} W_h^2 \widehat{Var}(\hat{p}_h) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \cdot \frac{n_h}{n_h - 1} \hat{p}_h \hat{q}_h\right)$$

5. **Confidence Interval (CI)**:

$$CI \quad d.r.$$

Determining Sample Size

When given n, determine n_h for each stratum

Use

```
strata.weight=function(Wh, S2h, Ch=NULL, allocation)

return(wh)
```

`allocation = "Prop"` or `"Opt"` or `"Neyman"`

Use

```
strata.size=function(n, Wh, S2h, Ch=NULL, allocation)

return(list(n=n, allocation=allocation, wh=wh, nh=ceiling(nh)))
```

The sample size for each stratum, n_h, can be determined using different allocation methods. The general formula is:

$$n_h = W_h \cdot n$$

where W_h is the stratum weight and n is the total sample size.

1. Proportional Allocation (Prop):

The stratum weight W_h is proportional to the stratum size N_h:

$$W_h = \frac{N_h}{N}$$

Thus, the sample size for stratum h is:

$$n_h = \frac{N_h}{N} \cdot n$$

This method ensures that the sample size in each stratum is proportional to the stratum's size in the population.

2. Optimal Allocation (Opt):

The stratum weight W_h is adjusted based on the stratum's variability and cost. The formula is:

$$W_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}$$

Thus, the sample size for stratum h is:

$$n_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}} \cdot n$$

This method minimizes the variance of the estimator by allocating more samples to strata with higher variability or lower costs.

3. Neyman Allocation:

The stratum weight W_h is adjusted based on the stratum's variability. The formula is:

$$W_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}$$

If the cost per unit is the same across all strata ($c_h = c$), this simplifies to:

$$W_h = \frac{\frac{N_h S_h}{\sqrt{c}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c}}} = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

Thus, the sample size for stratum h is:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \cdot n$$

This method minimizes the variance of the estimator by allocating more samples to strata with higher variability.

Summary

- **Proportional Allocation**: Simple and easy to implement, but does not account for variability.
- **Optimal Allocation**: Minimizes variance by considering both variability and cost.
- **Neyman Allocation**: A special case of optimal allocation when costs are equal across strata.

$$\boxed{\begin{array}{l} \text{Proportional Allocation: } n_h = \frac{N_h}{N} \cdot n \\ \text{Optimal Allocation: } n_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}} \cdot n \\ \text{Neyman Allocation: } n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \cdot n \quad (\text{when } c_h = c) \end{array}}$$

See `stratified size.r`

When given V, C, d, r of $\bar{Y}$, determine n and n_h

Use

```
strata.mean.size=function(Nh, S2h, Ch=NULL, allocation, method, bound, Ybar=NULL, alpha
```

Step 1 Calculate w_h with different allocation methods.

$$n_h = w_h n$$

$$w_h = \begin{cases} W_h & \text{prop} \\ \frac{WhS_h/\sqrt{C_h}}{\sum_h W_h Sh/\sqrt{C_h}} & \text{opt} \\ \frac{W_h S_h}{\sum W_h S_h} & \text{Neyman} \end{cases}$$

Step 2
Calculate $n_{\min}$:

$$n_{\min} = \frac{\sum_h W_h^2 S_h^2 / w_h}{V + \frac{1}{N} \sum_h W_h S_h^2}$$

where

$$V = \begin{cases} V & V \\ C^2 \bar{Y}^2 & C \\ (d/z_{\alpha/2})^2 & d \\ (r\bar{Y}/z_{\alpha/2})^2 & r \end{cases}$$

$S_h^2, \bar{Y}$ are given from historical data.

Step 3

$$n_{h\min} = w_h n_{\min}$$

Given V, C, d, r of P, determine n and n_h

Use

```
strata.prop.size=function(Nh, Ph, Ch=NULL, allocation, method, bound, Ybar=NULL, alpha=
```

Here

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$$

Given V, C, d, r of Total Y_T, determine n and n_h

Adjust the input bound parameter of calculating n and n_h of given parameters of $\bar{Y}$.
Use

```
strata.mean.size=function(Nh, S2h, Ch=NULL, allocation, method, bound, Ybar=NULL, alpha
```

Population Total (Y_T) $\quad \hat{y}_T$	V	C	d	r
Population Mean $(\bar{Y})$ $\quad \bar{y}_{st}$	$\frac{V}{N^2}$	C	$\frac{d}{N}$	r

Given V, C, d, r of Total A, determine n and n_h

Adjust the input bound parameter of calculating n and n_h of given parameters of $\bar{Y}$.
Use

```
strata.prop.size=function(Nh, Ph, Ch=NULL, allocation, method, bound, Ybar=NULL, alpha=
```

Population Total (A) $\quad \hat{a}$	V	C	d	r
Population Mean (P) $\quad \hat{p}_{st}$	$\frac{V}{N^2}$	C	$\frac{d}{N}$	r

Design Efficiency - Comparison of Sampling Methods

Comparing the variance of your method versus Simple Random Sampling under the same sampling size, the design efficiency is defined as the fraction.

$$\text{Deff} = \frac{Var(\hat{\theta}_p)}{Var(\hat{\theta}_{SRS})}$$

Ratio Estimation and Regression Estimation

Notations

For population use UPPER CASE characters and for sample use lower case.

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \qquad (Y \text{ Total Variance})$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \qquad (X \text{ Total Variance})$$

$$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X}) \qquad (\text{Y, X Total Covariance})$$

$$\rho = \frac{S_{yx}}{\sqrt{S_y^2 \cdot S_x^2}} = \frac{S_{yx}}{S_y \cdot S_x} \qquad (\text{Y, X Total Correlation})$$

$$C_y^2 = \frac{S_y^2}{\bar{Y}^2} \qquad (\text{Y Total Relative Variance})$$

$$C_x^2 = \frac{S_x^2}{\bar{X}^2} \qquad (\text{X Total Relative Variance})$$

$$C_{yx} = \rho \cdot \frac{S_y}{\bar{Y}} \cdot \frac{S_x}{\bar{X}} \qquad (\text{Y X Relative Covariance})$$

Estimation of Ratio

Ratio is defined as

$$R = \frac{\bar{Y}}{\bar{X}} = \frac{Y_T}{X_T}$$

1. Point Estimation

$$\hat{R} = \frac{\bar{y}}{\bar{x}}$$

2. AUE

$$\lim_{n\to\infty} E(\hat{R}) = R$$

3. Variance of Estimation

> **Proposition** :
>
> $$MSE(\hat{R}) \overset{AUE}{\simeq} Var(\hat{R}) \overset{n\to\infty}{\simeq} \frac{1-f}{n\bar{X}^2}\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2$$
>
> Where:
>
> $$\begin{aligned} S_g^2 &\overset{0}{=} \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2 \\ &\overset{1}{=} S_y^2 + R^2 S_x^2 - 2RS_{yx} \\ &\overset{2}{=} \bar{Y}^2(C_y^2 + C_x^2 - 2C_{yx}) \end{aligned}$$

4. Estimation of Variance

Method 1 When $\bar{X}$ is given

$$\begin{aligned} \widehat{\mathrm{Var}}_1(\hat{R}) &\overset{0}{=} \frac{1-f}{n}\cdot\frac{1}{\bar{X}^2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2 \\ &\overset{1}{=} \frac{1-f}{n}\cdot\frac{1}{\bar{X}^2}\cdot(S_y^2 + \hat{R}^2 S_x^2 - 2\hat{R}S_{yx}) \end{aligned}$$

Method 2 When $\bar{X}$ is unknown, we use $\bar{x}$ from the sample

$$\begin{aligned} \widehat{\mathrm{Var}}_2(\hat{R}) &\overset{0}{=} \frac{1-f}{n}\cdot\frac{1}{\bar{X}^2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2 \\ &\overset{1}{=} \frac{1-f}{n}\cdot\frac{1}{\bar{X}}\cdot(S_y^2 + \hat{R}^2 S_x^2 - 2\hat{R}S_{yx}) \end{aligned}$$

Note: When $\bar{X}$ is given, we can use both methods 1 and 2. When $\bar{X}$ is unknown, use method 2.

5. Confidence Interval

CI1, CI2, CI3

> placeholder for confidence interval

Use `ratio.r`

```r
ratio = function(y.sample, x.sample, N=NULL, auxiliary=FALSE, Xbar=NULL, alpha)
#when auxiliary = false , Xbar =null ; when auxiliary = true, Xbar = Xbar
```

Ratio Estimation of Population Mean $\bar{Y}$ and Total Y_T

SRSF (Simple Random Sampling with Fixed Ratio Estimation) of Population Mean $\bar{Y}$

1. **Estimator for the Population Mean**:

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \cdot \bar{X} = \hat{R} \cdot \bar{X}$$

2. **Expected Value of the Estimator**:

$$E(\bar{y}_R) = E(\hat{R}) \cdot \bar{X} \approx R \cdot \bar{X} = \bar{Y} \quad \text{(AUE)}$$

3. **Variance of the Estimator**:

$$\text{Var}(\bar{y}_R) = \bar{X}^2 \cdot \text{Var}(\hat{R})$$

4. **Estimated Variance of the Estimator**:

$$\widehat{\text{Var}}(\bar{y}_R) = \bar{X}^2 \cdot \widehat{\text{Var}}_1(\hat{R})$$

5. **Confidence Interval**:

$$\text{CI} = \left[\bar{X} \cdot \text{left}, \, \bar{X} \cdot \text{right}\right]$$

In `ratio.r` use:

```r
ratio.mean=function(y.sample, x.sample, N=NULL, Xbar, alpha)
```

Example:

```
mean.simple.result=srs.mean(N, y.sample, alpha)

mean.ratio.result=ratio.mean(y.sample, x.sample, N, Xbar, alpha)

var.result=c(mean.simple.result$ybar.var, mean.ratio.result$ybarR.var)
deff.result=deff(var.result)

rownames(deff.result)=c("Simple", "Ratio")
print(deff.result)
```

SRSF Estimation of Population Total Y_T

1. **Estimator for the Population Total**:

$$\hat{Y}_R = N \cdot \bar{y}_R$$

2. **Approximately Unbiased Estimator (AUE)**:

$$E(\hat{Y}_R) \approx Y_T$$

3. **Variance of the Estimator**:

$$\mathrm{Var}(\hat{Y}_R) = N^2 \cdot \mathrm{Var}(\bar{y}_R)$$

4. **Estimated Variance of the Estimator**:

$$\widehat{\mathrm{Var}}(\hat{Y}_R) = N^2 \cdot \widehat{\mathrm{Var}}(\bar{y}_R)$$

5. **Confidence Interval**:

$$\mathrm{CI} = [N \cdot \mathrm{left},\ N \cdot \mathrm{right}]$$

use:

```
ratio.total=function(y.sample, x.sample, N, Xbar, alpha)
```

Example

```r
total.simple.result=srs.total(N, y.sample, alpha)

total.ratio.result=ratio.total(y.sample, x.sample, N, Xbar, alpha)

var.result=c(total.simple.result$ytot.var, total.ratio.result$ytotal.var)
deff.result=deff(var.result)
rownames(deff.result)=c("Simple", "Ratio")
print(deff.result)
```

Design Efficiency

Ratio and Regression Estimation are called **complex** estimation methods, while Simple Random Sampling is called **simple** estimation method. When comparing complex methods to simple methods, design efficiency is defined as the fraction.

$$\text{Deff} = \frac{Var(\bar{y}_R)}{Var(\bar{y})} = \begin{cases} < 1 & \bar{y}_R \text{ is more efficient} \\ \geq 1 & \bar{y} \text{ is more efficient} \end{cases}$$

When

$$\rho > \frac{C_x}{2C_y}$$

$\bar{y}_R$ if more efficient then $\bar{y}$.
When Y and X are highly correlated, $\bar{y}_R$ is more efficient than $\bar{y}$.

Determining Sample Size

Step 1
When given bound (V, C, d, r) of $\bar{Y}$, determine the simple sample size n_{simple}
Using the function `size.mean`
Step 2
Determine the ratio sample size n_R

$$n_R = \text{Deff} \cdot n_{\text{simple}}$$

Use `deff=function(var.result)` to calculate the design efficiency and use `deff.size=function(deff.result, n.simple)` to calculate the size.

Example

```
mean.simple.result=srs.mean(N, y.sample, alpha)
mean.ratio.result=ratio.mean(y.sample, x.sample, N, Xbar, alpha)

var.result=c(mean.simple.result$ybar.var, mean.ratio.result$ybarR.var)
deff.result=deff(var.result)

n.simple=size.mean(N, Mean.his=NULL, Var.his=var(y.sample), method="d", bound=0.05, alp
size.result=deff.size(deff.result, n.simple)

rownames(size.result)=c("Simple", "Ratio")
print(size.result)
```

Regression Estimation of Population Mean $\bar{Y}$ and Total Y_T

The Linear Regression Estimator is defined as

$$\bar{y}_{lr} = \bar{y} + \beta(\bar{X} - \bar{x})$$

$$\hat{Y}_{lr} = N\bar{y}_{lr}$$

> Normally β is either constant or the regression coefficient B of X on Y.

> When $\beta = 1$, we obtain the difference estimator **Diff**
>
> $$\bar{y}_d = \bar{y} + (\bar{X} - \bar{x})$$
>
> When $\beta = 0$, it degenerates to the simple estimator $\bar{y}$.
> When $\beta = \bar{y}/\bar{x} = \hat{R}$, we obtain the ratio estimator $\bar{y}_R$.

Regression Estimation of Population Mean $\bar{Y}$

Case 1: $\beta = \beta_0$ is constant

1. Estimator for the Population Mean

$$\bar{y}_{lr}(\beta_0) = \bar{y} + \beta_0(\bar{X} - \bar{x})$$

2. Unbiased Estimator

$$E(\bar{y}_{lr}) = \bar{Y} + \beta_0(\bar{X} - E(\bar{x})) = \bar{Y} \quad \text{(UE)}$$

3. Variance of the Estimator

$$\text{Var}(\bar{y}_{lr}) = \frac{1-f}{n}\left(S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{yx}\right)$$

> Minimum Variance Condition
>
> $$\text{Minimum when } \beta_0 = B = \frac{S_{yx}}{S_x^2} \Rightarrow \text{Var}_{\min} = \frac{1-f}{n} S_e^2$$
>
> Here B is the population regression coefficient of X on Y.
>
> $$B = \frac{S_{yx}}{S_x^2}$$
>
> $$\text{Var}_{\min}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2(1-\rho^2)$$
>
> $$S_e^2 \triangleq S_y^2(1-\rho^2), \qquad \rho = \frac{S_{yx}}{SyS_x}$$

4. Estimated Variance of the Estimator

$$\widehat{\text{Var}}(\bar{y}_{lr}) = \frac{1-f}{n}\left(s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{yx}\right)$$

5. Confidence Interval

$$\left[\bar{y}_{lr} \pm z_\alpha \sqrt{\widehat{\text{Var}}(\bar{y}_{lr})}\right]$$

Case 2: $\beta = \hat{b}$ is the sample regression coefficient of x and y

$$\beta = \hat{b} = \frac{s_{yx}}{s_x^2}$$

1. Estimator for the Population Mean

$$\bar{y}_{lr} = \bar{y} + \hat{b}(\bar{X} - \bar{x})$$

2. Approximate Unbiased Estimator

$$E(\bar{y}_{lr}) \approx \bar{Y} \quad (\text{AUE})$$

3. Mean Squared Error (MSE) and Variance

$$\text{MSE}(\bar{y}_{lr}) \approx \text{Var}(\bar{y}_{lr}) \approx \frac{1-f}{n} S_e^2$$

> This is the theoretically minimum variance estimator.

4. Estimated Variance of the Estimator

$$\widehat{\text{Var}}(\bar{y}_{lr}) = \frac{1-f}{n} s_e^2 = \frac{1-f}{n} \cdot \frac{n-1}{n-2} \left(s_y^2 - \frac{s_{yx}^2}{s_x^2} \right)$$

where

$$s_e^2 = \frac{n-1}{n-2} \left(s_y^2 - \frac{s_{yx}^2}{s_x^2} \right)$$

In `regression.r` :

```
regression.mean=function(y.sample, x.sample, N=NULL, Xbar, alpha, method="Min", beta0=N
```

Regression Estimation of Population Total Y_T

Notice that

$$\text{mean} \quad \bar{y}_{lr} \xrightarrow{N} \hat{y}_{lr} \quad \text{total}$$

```
regression.total=function(y.sample, x.sample, N=NULL, Xbar, alpha, method="Min", beta0=
```

Comparison of Simple, Ratio, and Regression Estimation

Their corresponding variances are:

$$\begin{aligned}
\mathrm{Var}(\bar{y}) &= \frac{1-f}{n} \cdot S_y^2 \\
\mathrm{Var}(\bar{y}_R) &\approx \frac{1-f}{n} \cdot (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) \\
\mathrm{Var}(\bar{y}_{lr}) &\approx \frac{1-f}{n} \cdot S_y^2(1-\rho^2)
\end{aligned}$$

The condition for the regression estimator to be more efficient than the ratio estimator is:

$$(B-R)^2 \geq 0$$

When n is not large, the estimations might be biased. Real-life experiments show that when n is small, the regression estimator can be more biased than the ratio estimator.

Example: Comparing Simple, Ratio, and Regression Estimation of Population Total Y_T

```
total.simple.reult=srs.total(N, y.sample, alpha)
print(total.simple.result)

total.ratio.result=ratio.total(y.sample, x.sample, N, Xbar, alpha)
print(total.ratio.result)

total.reg.result=regression.total(y.sample, x.sample, N, Xbar, alpha, method="Min", bet
print(total.reg.result)

var.result=c(total.simple.reult$ytot.var, total.ratio.result$ytotal.var, total.reg.resu
deff.result=deff(var.result)
rownames(deff.result)=c("Simple", "Ratio", "Regression")
print(deff.result)
```

Determining Sample Size

The design efficiency is defined as the fraction.

$$\text{Deff} = \frac{Var(\bar{y}_{lr})}{Var(\bar{y})}$$

Given the bound (V, C, d, r) of $\bar{y}$, determine the simple sample size n_{simple}, then

$$n_{lr} = \text{Deff} \cdot n_{\text{simple}}$$

Which is similar to ratio estimation.

Stratified Ratio and Regression Estimation

Two approaches for stratified estimation.

1. **Separated Estimation** First estimate for each stratum, then take weighted average or sum.
2. **Combined Estimation** First take the weighted average or sum, then estimate for the combined sample.

Stratified Ratio Estimation

For the h-th stratum ($h = 1, \dots, L$):

Notations	Population $\xrightarrow{SRS}$	Sample
	$\begin{pmatrix} Y_{h1} & \cdots & Y_{hN_h} \\ X_{h1} & \cdots & X_{hN_h} \end{pmatrix}$	$\begin{pmatrix} y_{h1} & \cdots & y_{hn_h} \\ x_{h1} & \cdots & x_{hn_h} \end{pmatrix}$
Mean	$\bar{Y}_h \quad \bar{X}_h$	$\bar{y}_h \quad \bar{x}_h$
Var, Cov, ρ	$S^2_{yh}, S^2_{xh}, S_{yxh}, \rho_h$	$s^2_{yh}, s^2_{xh}, s_{yxh}, \hat{\rho}_h$

Notations	Population $\xrightarrow{SRS}$	Sample
Seperate Ratio Estimation for each stratum	$R_h = \dfrac{\bar{Y}_h}{\bar{X}_h}$	$\hat{R}_h = \dfrac{\bar{y}_h}{\bar{x}_h}$
Combined Ratio Estimation	$R_c = \dfrac{\bar{Y}}{\bar{X}}$	$\hat{R}_c = \dfrac{\bar{y}_{st}}{\bar{x}_{st}}$

Separate Ratio Estimation of Population Mean $\bar{Y}$

1. **Estimator for the Population Mean**

$$\bar{y}_{RS} = \sum_h W_h \bar{y}_{Rh} = \sum_h W_h \left(\frac{\bar{y}_h}{\bar{x}_h} \cdot \bar{X}_h \right)$$

> Notice that
>
> $$\bar{y}_{Rh} = \frac{\bar{y}_h}{\bar{x}_h} \cdot \bar{X}_h$$
>
> is the ratio estimator of the h-th stratum.

2. **Approximate Unbiasedness**

$$E(\bar{y}_{RS}) \approx \bar{Y} \quad (\text{AUE})$$

3. **Variance of the Estimator**

$$\operatorname{Var}(\bar{y}_{RS}) \approx \sum_h W_h^2 \frac{1-f_h}{n_h} \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{yxh} \right)$$

4. **Estimated Variance of the Estimator**

$$\widehat{\operatorname{Var}}(\bar{y}_{RS}) \approx \sum_h W_h^2 \frac{1-f_h}{n_h} \left(s_{y_h}^2 + \hat{R}_h^2 s_{x_h}^2 - 2\hat{R}_h s_{yxh} \right)$$

where $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$

In `stra ratio.r`

```
separate.ratio.mean=function(Nh, y.sample, x.sample, stra.index, Xbarh, alpha)
```

Combined Ratio Estimation of Population Mean $\bar{Y}$

1. **Estimator for the Population Mean**

$$\bar{y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \cdot \bar{X} = \hat{R}_c \cdot \bar{X}$$

2. **Approximate Unbiasedness**

$$E(\bar{y}_{RC}) \approx \bar{Y} \quad (\text{AUE})$$

3. **Variance of the Estimator**

$$\text{Var}(\bar{y}_{RC}) = \sum_h W_h^2 \frac{1-f_h}{n_h} \left(S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h S_{yx_h}\right)$$

4. **Estimated Variance of the Estimator**

$$\widehat{\text{Var}}(\bar{y}_{RC}) = \sum_h W_h^2 \frac{1-f_h}{n_h} \left(s_{y_h}^2 + \hat{R}_c^2 s_{x_h}^2 - 2\hat{R}_c s_{yx_h}\right)$$

where:

$$\hat{R}_c = \frac{\bar{y}_{st}}{\bar{x}_{st}}$$

In `stra ratio.r`

```
combined.ratio.mean=function(Nh, y.sample, x.sample, stra.index, Xbar, alpha)
```

Stratified Regression Estimation

Separate Regression Estimation of Population Mean $\bar{Y}$

Case I: When β_h is constant

1. **Estimator for the Population Mean**

$$\bar{y}_{lrS} = \sum_h W_h \bar{y}_{lrh} = \sum_h W_h \left(\bar{y}_h + \beta_h(\bar{X}_h - \bar{x}_h)\right)$$

> Notice that
>
> $$\bar{y}_{lrh} = \bar{y}_h + \beta_h(\bar{X}_h - \bar{x}_h)$$
>
> is the regression estimator of the h-th stratum.

2. **Unbiasedness**

$$E(\bar{y}_{lrS}) = \bar{Y} \quad \text{(UE)}$$

3. **Variance of the Estimator**

$$\text{Var}(\bar{y}_{lrS}) = \sum_h W_h^2 \frac{1-f_h}{n_h} \left(S_{y_h}^2 + \beta_h^2 S_{x_h}^2 - 2\beta_h S_{yx_h}\right)$$

> **Minimum Variance Condition**
> When $\beta_h = B_h = \frac{S_{yx_h}}{S_{x_h}^2}$:
>
> $$\text{Var}_{\min} = \sum_h W_h^2 \frac{1-f_h}{n_h} S_{eh}^2$$
>
> where:
>
> $$S_{eh}^2 = S_{y_h}^2(1-\rho_h^2)$$

4. **Estimated Variance of the Estimator**

$$\widehat{\text{Var}}(\bar{y}_{lrS}) = \sum_h W_h^2 \frac{1-f_h}{n_h} \left(s_{y_h}^2 + \hat{\beta}_h^2 s_{x_h}^2 - 2\hat{\beta}_h s_{yx_h}\right)$$

Case II: When $\beta_h = \hat{b}_h = \frac{S_{yx_h}}{S_{x_h}^2}$ (Regression Coefficient)

1. Estimator for the Population Mean

$$\bar{y}_{lrS} = \sum_h W_h \left(\bar{y}_h + \hat{b}_h(\bar{X}_h - \bar{x}_h)\right)$$

2. **Asymptotically Unbiased Estimator**

$$E(\bar{y}_{lrS}) \approx \bar{Y} \quad \text{(AUE)}$$

3. **Variance of the Estimator**

$$\mathrm{Var}(\bar{y}_{lrS}) \approx \sum_h W_h^2 \frac{1-f_h}{n_h} S_{y_h}^2 (1-\rho_h^2)$$

4. **Estimated Variance of the Estimator**

$$\widehat{\mathrm{Var}}(\bar{y}_{lrS}) \approx \sum_h W_h^2 \frac{1-f_h}{n_h} \frac{n_h-1}{n_h-2} \left(s_{y_h}^2 - \frac{s_{yx_h}}{s_{x_h}^2}\right)^2$$

In `stra regression.r`

```
seperate.regression.mean=function(Nh, y.sample, x.sample, stra.index, Xbarh, alpha, met
```

Combined Regression Estimation of Population Mean $\bar{Y}$

Case I: When β is constant

1. **Estimator for the Population Mean**

$$\bar{y}_{lrC} = \bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st})$$

2. **Unbiasedness**

$$E(\bar{y}_{lrC}) = \bar{Y} \quad \text{(UE)}$$

3. **Variance of the Estimator**

$$\mathrm{Var}(\bar{y}_{lrC}) = \sum_h W_h^2 \frac{1-f_h}{n_h} \left(S_{yh}^2 + \beta^2 S_{xh}^2 - 2\beta S_{yxh}\right)$$

> **Minimum Variance Condition**
> When
>
> $$\beta = B_c = \frac{\sum_h W_h^2 \frac{1-f_h}{n_h} S_{yxh}}{\sum_h W_h^2 \frac{1-f_h}{n_h} S_{xh}^2}$$
>
> The Variance achieves its minimum.

4. **Estimated Variance of the Estimator**

$$\widehat{\mathrm{Var}}(\bar{y}_{lrC}) = \sum_h W_h^2 \frac{1-f_h}{n_h} \left(s_{yh}^2 + \hat{\beta}^2 s_{xh}^2 - 2\hat{\beta} s_{yxh}\right)$$

Case II

When

$$\beta = \hat{b}_c = \frac{\sum_h W_h^2 \frac{1-f_h}{n_h} S_{yx_h}}{\sum_h W_h^2 \frac{1-f_h}{n_h} S_{x_h}^2}$$

1. **Estimator for the Population Mean**

$$\bar{y}_{lrC} = \bar{y}_{st} + \hat{b}_c(\bar{X} - \bar{x}_{st})$$

2. **Approximate Unbiasedness**

$$E(\bar{y}_{lrC}) \approx \bar{Y} \quad (\mathrm{AUE})$$

3. **Variance of the Estimator**

$$\mathrm{Var}(\bar{y}_{lrC}) \approx \sum_h W_h^2 \frac{1-f_h}{n_h} \left(S_{y_h}^2 + B_c^2 S_{x_h}^2 - 2B_c S_{yx_h}\right)$$

4. **Estimated Variance of the Estimator**

$$\widehat{\mathrm{Var}}(\bar{y}_{lrC}) \approx \sum_h W_h^2 \frac{1-f_h}{n_h} \left(s_{y_h}^2 + \hat{b}_c^2 s_{x_h}^2 - 2\hat{b}_c s_{yx_h}\right)$$

In `stra regression.r`

```
combined.regression.mean = function(Nh, y.sample, x.sample, stra.index, Xbar, alpha, me
```

Estimation of Population Total Y_T

Notice that

$$\text{mean} \quad \bar{Y} \xrightarrow{N} Y_T \quad \text{total}$$

In `stra ratio.r`

```
separate.ratio.total=function(Nh, y.sample, x.sample, stra.index, Xbarh, alpha)
```

```
combined.ratio.total=function(Nh, y.sample, x.sample, stra.index, Xbar, alpha)
```

In `stra regression.r`

```
seperate.regression.total = function(Nh, y.sample, x.sample, stra.index, Xbarh, alpha,
```

```
combined.regression.total = function(Nh, y.sample, x.sample, stra.index, Xbar, alpha, m
```

Determining Sample Size

$$\text{Deff} = \frac{Var(\bar{y}_{\text{prop}})}{Var(\bar{y}_{\text{st}})} \implies n_{\text{prop}} = \text{Deff} \cdot n_{\bar{y}_{st}}$$

Where given bound (V, C, d, r) of $\bar{y}_{st}$, calculate sample size $n_{\bar{y}_{st}}$
$\bar{y}_{\text{prop}}$ is estimated by the methods (RS, RC, lrS, lrC).

Double Sampling

Double Sampling or **Two-phase Sampling** is a method with two phases.

1. First, sample from the population to obtain a big sample to obtain auxiliary information. In this course, the first-phase sampling is always SRS.
2. Second, sample with a small size. In this book, the second-phase is always sampled from the first-phase.

Process

Population:

$$Y_1, \ldots, Y_N$$

Step 1:

- **Sampling Method**: SRS (Simple Random Sampling)
- **Sample Drawn**:

$$y'_1, \ldots, y'_{n'} \quad \text{(First Sample)}$$

Step 2:

- **Second Sample**:

$$y_1, \ldots, y_n \quad \text{(Second Sample)}$$

Estimation:

- **Estimator**:

$$\hat{\theta} = \hat{\theta}(y_1, \ldots, y_n)$$

Expectation and Variance Decomposition

Expectation of the Estimator

$$\begin{aligned} E(\hat{\theta}) &= E\left(\hat{\theta}(y_1, \ldots, y_n)\right) \\ &= E_1\left[E_2\left(\hat{\theta}(y_1, \ldots, y_n) \middle| y'_1, \ldots, y'_{n'}\right)\right] \\ &= E_1\left[E_2(\hat{\theta})\right] \end{aligned}$$

Variance of the Estimator

The variance of the estimator $\hat{\theta}$ can be decomposed as:

$$\operatorname{Var}(\hat{\theta}) = \operatorname{Var}_1\left(E_2\left(\hat{\theta} \mid y'_1, \ldots, y'_{n'}\right)\right) + E_1\left(\operatorname{Var}_2\left(\hat{\theta} \mid y'_1, \ldots, y'_{n'}\right)\right)$$

Double Stratified Sampling

Sampling Process

Step 1

SRS sample from the population to obtain the first-phase samples. For known N and given n':

$$(Y_1, \ldots, Y_N) \xrightarrow{SRS} (y'_1, \ldots, y'_{n'})$$

Step 2

Stratify the first-phase samples $(y'_1, \ldots, y'_{n'})$ into L strata. The unit count for stratum h is n'_h.
The samples are: $(y'_{n'_1}, \ldots, y'_{n'_h}), \quad h = 1, \ldots, L$

Step 3

Estimate the stratum weight of stratum h, since $W_h = \frac{N_h}{N}$ is unkown.
Using samples from the first-phase, we have:

$$w'_h = \frac{n'_h}{n'}, \quad h = 1, \ldots, L$$

Step 4

Perform a stratified sampling from the first-phase samples $(y'_1, \ldots, y'_{n'})$ to obtain the second-phase samples.

$$(y'_{n'_1}, \ldots, y'_{n'_h}) \longrightarrow (y_{n_1}, \ldots, y_{n_{n_h}})$$

Two-Phase Sampling Formulas

1. Second-phase sampling proportion:

$$v_h = \frac{n_h}{n'_h}$$

- n_h: Size of the second-phase sample.
- n'_h: Size of the first-phase sample.

2. Second-phase Sample Mean for h-th stratum ($\bar{y}_h$):

$$\bar{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$$

- y_{hj}: Value of the target variable y for the j-th unit in the second-phase sample of stratum h.

3. Second-phase Variance for h-th stratum (S_h^2):

$$S_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

- $\bar{y}_h$: Sample mean of the target variable y for the second-phase sample of stratum h.

Double Stratified Sampling Estimation of Population Mean $\bar{Y}$

1. **Estimator for the Population Mean**:

$$\bar{y}_{stD} = \sum_{h=1}^{L} w'_h \cdot \bar{y}_h$$

2. **Unbiased Estimation**:

$$E(\bar{y}_{stD}) = \bar{Y} \quad \text{(UE)}$$

3. **Variance of the Estimator**:

$$\operatorname{Var}(\bar{y}_{stD}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \sum_{h=1}^{L} \frac{1}{n'_h} (v'_h - 1) w'_h S_h^2$$

4. **Estimated Variance of the Estimator**:

$$\widehat{\operatorname{Var}}(\bar{y}_{stD}) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{n'_h}\right) w'_h s_h^2 + \left(\frac{1}{n'} - \frac{1}{N}\right) \sum_{h=1}^{L} w'_h \left(\bar{y}_h - \bar{y}_{stD}\right)^2$$

In `two phase stra.r`

```
twophase.stra.mean1=function(N=NULL, nh.1st, nh.2nd, ybarh, s2h, alpha)

twophase.stra.total1=function(N, nh.1st, nh.2nd, ybarh, s2h, alpha)
```

Double Ratio Estimation and Double Regression Estimation

Sampling Process

Y is the target property and X is the auxiliary property.

Step 1

SRS sample from the population to obtain the first-phase samples. For known N and given n':

$$\begin{pmatrix} Y_1 & \cdots & Y_N \\ X_1 & \cdots & X_N \end{pmatrix} \xrightarrow{SRS} \begin{pmatrix} y_1' & \cdots & y_{n'}' \\ x_1 & \cdots & x_{n'}' \end{pmatrix}$$

Step 2

Since the auxiliary information $\bar{X}$ is unknown, use the first-phase samples to estimate $\bar{X}$:

$$\bar{X} = \frac{1}{n'} \sum_{j=1}^{n'} x_j'$$

Step 3

SRS from the first-phase samples to obtain the second-phase samples:

$$\begin{pmatrix} y_1' & \cdots & y_{n'}' \\ x_1 & \cdots & x_{n'}' \end{pmatrix} \xrightarrow{SRS} \begin{pmatrix} y_1 & \cdots & y_n \\ x_1 & \cdots & x_n \end{pmatrix}$$

Notations for second-phase samples:

$$\bar{y}, \bar{x}, s_y^2, s_x^2, s_{yx}$$

Double Ratio Estimation of Population Mean $\bar{Y}$

1. **Estimator for the Population Mean**:

$$\bar{y}_{RD} = \hat{R} \cdot \bar{x}' = \frac{\bar{y}'}{\bar{x}'} \cdot \bar{x}'$$

2. **Asymptotically Unbiased Estimation**:

$$E(\bar{y}_{RD}) \approx \bar{Y} \quad \text{AUE}$$

3. **Variance of the Estimator**:

$$\text{Var}(\bar{y}_{RD}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) (S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

4. **Estimated Variance of the Estimator**:

$$\widehat{\text{Var}}(\bar{y}_{RD}) = \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \left(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{yx}\right)$$

In `twophase ratio.r`

```
twophase.ratio.mean=function(N=NULL, n.1st, xbar.1st, y.sample, x.sample, alpha)
```

```
twophase.ratio.total=function(N, n.1st, xbar.1st, y.sample, x.sample, alpha)
```

Double Regression Estimation of Population Mean $\bar{Y}$

Case 1: When β is a Constant ($\beta = \beta_0$, i.e., $\beta = 1$)

1. **Estimator for the Population Mean**:

$$\bar{y}_{lrD} = \bar{y} + \beta(\bar{x}' - \bar{x})$$

2. **Unbiasedness**:

$$E(\bar{y}_{lrD}(\beta_0)) = \bar{Y} \quad \text{(UE)}$$

3. **Variance of the Estimator**:

$$\text{Var}(\bar{y}_{lrD}(\beta_0)) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \left(S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{yx}\right)$$

4. **Estimated Variance of the Estimator**:

$$\widehat{\text{Var}}(\bar{y}_{lrD}(\beta_0)) = \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \left(s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{yx}\right)$$

Here is the Markdown representation of the given mathematical expressions:

Case II: When β is the regression coefficient of the second-phase sample

$$\beta = \hat{b} = \frac{S_{yx}}{S_x^2}$$

1. **Estimator for the Population Mean**

$$\bar{y}_{lrD} = \bar{y} + \hat{b}(\bar{x}' - \bar{x})$$

2. **Asymptotically Unbiased Estimation**

$$E(\bar{y}_{lrD}) \approx \bar{Y} \quad \text{(AUE)}$$

3. **Variance of the Estimator**

$$\text{Var}(\bar{y}_{lrD}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_y^2(1-\rho^2)$$

4. **Estimated Variance of the Estimator**

$$\widehat{\text{Var}}(\bar{y}_{lrD}) = \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) s_e^2$$

where:

$$s_e^2 = \frac{n-1}{n-2}\left(s_y^2 - \frac{s_{yx}^2}{s_x^2}\right)$$

In `twophase regression.r`

```
twophase.regression.mean=function(N=NULL, n.1st, xbar.1st, y.sample, x.sample, alpha, b

twophase.regression.total=function(N=NULL, n.1st, xbar.1st, y.sample, x.sample, alpha,
```

Cluster Sampling

The population is formed by clusters. Cluster Sampling is to sample clusters and examine all the smaller units within the clusters.

Cluseter Sampling Estimation of Population Mean $\bar{Y}$

Sampling Process

Population is formed by N clusters:

$$\boxed{Y_{11}, \ldots, Y_{1M_1}}_1 \quad \cdots \quad \boxed{Y_{i1}, \ldots, Y_{iM_i}}_i \quad \cdots \quad \boxed{Y_{N1}, \ldots, Y_{NM_N}}_N$$

SRS from the cluster indices:

$$(1, \ldots, N) \xrightarrow{SRS} (1, \ldots, n)$$

We obtain the samples:

$$\boxed{y_{11}, \ldots, y_{1m_1}}_1 \quad \cdots \quad \boxed{y_{i1}, \ldots, y_{im_i}}_i \quad \cdots \quad \boxed{y_{n1}, \ldots, y_{nm_n}}_n$$

For a given n, the sample rate:

$$f = \frac{n}{N}$$

Clusters with the same size $(M_i = M = m_i)$

Notations

UPPER CASE: population; lower case: sample.

$$\overline{Y}_i = \frac{1}{M}\sum_{j=1}^{M} Y_{ij} \qquad \text{(cluster mean)}$$

$$\overline{\overline{Y}} = \frac{1}{MN}\left(\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}\right) \qquad \text{(unit mean)}$$

$$\overline{Y} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} Y_{ij}\right) \qquad \text{(mean by cluster)}$$

$$S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \overline{Y}_i\right)^2 \qquad \text{(Within-cluster variance)}$$

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \overline{Y}\right)^2 \qquad \text{(Total variance)}$$

$$S_b^2 = \frac{M}{N-1}\sum_{i=1}^{N}\left(\overline{Y}_i - \bar{Y}\right)^2 \qquad \text{(Between-cluster variance)}$$

$\bar{y}_i, \bar{\bar{y}}, \bar{y}, s_w^2, s^2, S_b^2$ can be defined similiarly.

> **Proposition:**
>
> - Decomposition of population variance S^2:

> $$\begin{aligned} S^2 &= \frac{1}{NM-1}\left[(N-1)S_b^2 + N(M-1)S_2^2\right] \\ &= \frac{N-1}{NM-1}S_b^2 + \frac{N(M-1)}{NM-1}S_w^2 \end{aligned}$$
>
> - Decomposition of sample variance s^2:
>
> $$s^2 = \frac{n-1}{nM-1}s_b^2 + \frac{n(M-1)}{nM-1}s_w^2$$

Estimation of the unit mean $\overline{\overline{Y}}$

1. **Estimation**

$$\overline{\overline{y}} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} = \left(\frac{1}{M}\right)\bar{y}$$

$$\left(= \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \mathrm{mean}(\bar{y}_1, \ldots, \bar{y}_n)\right)$$

2. **Unbiased**

$$E(\overline{\overline{y}}) = \overline{\overline{Y}} \quad \text{(UE)}$$

3. **Variance of Estimation**

$$\mathrm{Var}(\overline{\overline{y}}) = \frac{1-f}{nM}S_b^2$$

4. **Estimation of Variance**

$$\widehat{\mathrm{Var}}(\overline{\overline{y}}) = \frac{1-f}{nM}S_b^2$$

In `cluster srs.r`

```
cluster.srs.mean = function(N, M.ith, ybar.ith, s2.ith, alpha)
```

Estimation of Population Variance S^2

> **Proposition**
> Recall:
>
> $$S^2 = \frac{N-1}{NM-1} S_b^2 + \frac{N(M-1)}{NM-1} S_w^2$$
>
> $$s^2 = \frac{n-1}{nM-1} s_b^2 + \frac{n(M-1)}{nM-1} s_w^2$$
>
> We have:
>
> 1. s_b^2 is an Unbiased Estimator of S_b^2
> 2. s_w^2 is an Unbiased Estimator of S_w^2
> 3. s^2 is **NOT** an Unbiased Estimator of S^2
>
> The Unbiased Estimator of S^2 is:
>
> $$\begin{aligned} S^2 &\approx \frac{N-1}{NM-1} s_b^2 + \frac{N(M-1)}{NM-1} s_w^2 && (N \text{ is given}) \\ S^2 &\approx \frac{1}{M} s_b^2 + \frac{M-1}{M} s_w^2 && (N = +\infty) \end{aligned}$$

Design Efficiency

> **Definition**
> Within Cluster Correlation Coefficient:
>
> $$\begin{aligned} \rho_c &\overset{\text{def}}{=} \frac{2 \sum_{i=1}^{N} \sum_{j<k}^{M} (Y_{ij} - \overline{\overline{Y}})(Y_{ik} - \overline{\overline{Y}})}{(M-1)(NM-1)S^2} \\ &= 1 - \frac{NM S_w^2}{(NM-1)S^2} \end{aligned}$$
>
> Note that:

> $$\rho_c \in \left[-\frac{1}{M-1}, 1\right]$$

To Calculate the Design Efficiency, we need to calculate the variance of our cluster estimator versus the variance of SRS.

1. **Variance of the Cluster Estimator**

$$\begin{aligned}
\mathrm{Var}(\overline{\overline{y}}) &= \frac{1-f}{nM} S_b^2 \\
&= \frac{1-f}{nM} \cdot \frac{NM-1}{M(N-1)} S^2 (1 + (M-1)\rho_c) && (N \text{ is given}) \\
&= \frac{1-f}{nM} S^2 (1 + (M-1)\rho_c) && (N = +\infty)
\end{aligned}$$

Now lets tackle ρ_c, the estimation of ρ_c is:

$$\begin{aligned}
\rho_c &= 1 - \frac{NM s_2^2}{(NM-1)\hat{S}^2} && N \text{ is given} \\
&= \frac{s_b^2 - s_w^2}{s_b^2 + (M-1)s_w^2} && N = +\infty
\end{aligned}$$

2. **Variance for SRS from a population of NM with sample size nM**

$$\mathrm{Var}(\overline{y}_{SRS}) = \frac{1-f}{nM} S^2$$

Hence the Design Efficiency can be derived as:

$$\widehat{\mathrm{Deff}} = \frac{\mathrm{Var}(\overline{\overline{y}})}{\mathrm{Var}(\overline{y}_{SRS})} = \begin{cases} \frac{NM-1}{M(N-1)}(1 + (M-1)\hat{\rho}_c) & N \text{ is limitied} \\ 1 + (M-1)\hat{\rho}_c & N = +\infty \end{cases}$$

Determining the Sample Size

Given (V, C, d, r) for $\overline{y}_{SRS}$, we can determine the sample size n_{SRS}, therefore:

$$n_{\min} = \widehat{\mathrm{Deff}} \cdot n_{SRS}$$

The minimum number of clusters is:

$$n_{\text{cluster}} = \frac{n_{\min}}{M}$$

Clusters with different sizes

1. If M_i are close enough, use the mean $\bar{M}$ as a proxy for M:

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

2. When M_i are widely apart, use the stratified method for each cluster to obtain a similar stratum size. Then use the mean as a proxy.