

Sample Survey Formulae

Simple Random Sampling (Without Replacement)

Population: $\{Y_1, \dots, Y_N\}$, $\mu_Y = \sum_{i=1}^N Y_i/N$, $\tau_Y = \sum_{i=1}^N Y_i$, $\sigma_Y^2 = \sum_{i=1}^N (Y_i - \mu_Y)^2/N$, $S_Y^2 = N\sigma_Y^2/(N-1)$.

Sample: $\{y_1, \dots, y_n\}$, $\bar{y} = \sum_{i=1}^n y_i/n$, $s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$.

Estimating the mean

Estimator: $\hat{\mu}_Y = \bar{y}$.

Variance: $\text{Var}(\hat{\mu}_Y) = S_Y^2(1-f)/n$ where $f = n/N$.

Variance estimator: $\hat{\text{Var}}(\hat{\mu}_Y) = s_y^2(1-f)/n$.

Sample size: $n = N\sigma_Y^2/((N-1)D^2 + \sigma_Y^2)$, where $D^2 = \text{Var} \hat{\mu}_Y$.

Stratified Random Sampling

Population: N_h, μ_h, σ_h^2 and S_h^2 are as above but for stratum h ; $W_h = N_h/N$. μ_Y, N are as before and refer to the whole population.

Sample: n_h, \bar{y}_h, s_h^2 and $f_h = n_h/N_h$ are as above but for the subsample from stratum h . n is the whole sample size.

Estimating the mean

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

Variance: $\text{Var}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 S_h^2(1-f_h)/n_h$.

Variance estimator: $\hat{\text{Var}}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 s_h^2(1-f_h)/n_h$.

Sample size: $n = \sum_{h=1}^L (N_h^2 S_h^2 / w_h) / (N^2 V + \sum_{h=1}^L N_h S_h^2)$ where $V = \text{Var}(\hat{\mu}_{st})$, $w_h = n_h/n$.

Proportional allocation

Put $n_h/n = N_h/N$ then

Variance: $\text{Var}_{prop}(\hat{\mu}_{st}) = ((1-f)/n) \sum_{h=1}^L W_h S_h^2$

Optimal allocation

For cost function $C = c_0 + \sum c_h n_h$, the cost C is minimised for a specified variance $\text{Var}(\hat{\mu}_{st})$, and the variance $\text{Var}(\hat{\mu}_{st})$ is minimised for a fixed cost C , if

$$n_h \propto W_h S_h / \sqrt{c_h} \quad \text{that is} \quad \frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum (W_h S_h / \sqrt{c_h})}.$$

Thus,

$$n = \frac{(C - c_0) \sum (N_h S_h / \sqrt{c_h})}{\sum (N_h S_h \sqrt{c_h})}, \quad \text{if cost } C \text{ is fixed.}$$
$$n = \frac{(\sum W_h S_h / \sqrt{c_h}) (\sum W_h S_h \sqrt{c_h})}{V + (1/N) \sum W_h S_h^2}, \quad \text{if } V = \text{Var}(\hat{\mu}_{st}) \text{ is fixed.}$$

Neyman allocation

Optimal allocation when $c_h = c$ for all h .

$$\frac{n_h}{n} = \frac{W_h S_h}{\sum (W_h S_h)} = \frac{N_h S_h}{\sum (N_h S_h)}, \quad \text{Var}_{opt}(\hat{\mu}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}.$$

Post-stratification

Let m_h be the number of units in stratum h .

$$\begin{aligned} \text{Var}_p(\hat{\mu}_{st} | m_h) &= \sum_{h=1}^L W_h^2 \frac{S_h^2}{m_h} (1 - f_h) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{m_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \\ \text{Var}_p(\hat{\mu}_{st}) &\approx \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^L (1 - W_h) S_h^2. \end{aligned}$$

Cluster sampling

Population: μ_Y = population mean per element, M = number of elements in population;

Clusters: N = number of clusters in population, m_h = size of cluster h , μ_h = mean of cluster h ,
 τ_h = total of cluster h , σ_h^2 = variance of cluster h . σ_b^2 = between cluster variance.

Sample: n = number of clusters in sample, \bar{y}_i = mean of cluster i in sample, t_i = total of cluster i in sample.

One-stage cluster sampling with equal-sized clusters

Estimator: $\hat{\mu}_{cl} = \sum_{h=1}^n \bar{y}_h / n$.

Variance: $\text{Var}(\hat{\mu}_{cl}) = S_b^2(1-f)/n$ where $f = n/N$, $S_b^2 = N\sigma_b^2/(N-1)$.

Variance estimator: $\hat{V}ar(\hat{\mu}_{cl}) = s_b^2(1-f)/n$ where s_b^2 is the sample variance of the selected cluster means

One-stage cluster sampling with unequal-sized clusters

Estimator: $\hat{\mu}_{clr} = \bar{t}/\bar{m}$ where $\bar{t} = (\sum_i t_i)/n$, $\bar{m} = (\sum_i m_i)/n$ (sample averages).

Variance: $\text{Var}(\hat{\mu}_{clr}) \approx (N/M)^2((1-f)/n) \sum_{h=1}^N (\tau_h - \mu_Y m_h)^2 / (N-1)$.

Variance estimator: $\hat{V}ar(\hat{\mu}_{clr}) = (1/\bar{m})^2((1-f)/n) \sum_{i=1}^n (t_i - \hat{\mu}_{clr} m_i)^2 / (n-1)$

Ratio and regression estimators

Population: μ_X is known, μ_Y unknown, $R = \mu_Y / \mu_X$.

Sample: SRS with data $\{(x_i, y_i); i = 1, 2, \dots, n\}$, $r = \bar{y}/\bar{x}$.

Ratio estimator

Estimator: $\hat{\mu}_{ratio} = r\mu_X$.

Variance: $\text{Var}(\hat{\mu}_{ratio}) \approx ((1-f)/n) \sum_{\ell=1}^N (Y_\ell - RX_\ell)^2 / (N-1)$

Variance estimator: $\hat{\text{Var}}(\hat{\mu}_{ratio}) = ((1-f)/n) \sum_{i=1}^n (y_i - rx_i)^2 / (n-1)$
 $= ((1-f)/n)(s_y^2 - 2r\hat{\rho}s_x s_y + r^2 s_x^2)$ where $\hat{\rho} = s_{xy}/(s_x s_y)$

Regression estimator

Estimator: $\hat{\mu}_{lr} = \bar{y} + b(\mu_X - \bar{x})$.

- When $b = b_0$ is pre-assigned:

Variance: $\text{Var}(\hat{\mu}_{lr}) = ((1-f)/n)(S_Y^2 - 2b_0 S_{XY} + b_0^2 S_X^2)$

Variance estimator: $\hat{\text{Var}}(\hat{\mu}_{lr}) = ((1-f)/n)(s_y^2 - 2b_0 s_{xy} + b_0^2 s_x^2)$.

The best value to assign b is $\beta = S_{XY}/S_X^2$, which minimises $\text{Var}(\hat{\mu}_{lr})$.

$$\min_b [\text{Var}(\hat{\mu}_{lr})] = \frac{(1-f)}{n} \left(S_Y^2 - \frac{S_{XY}^2}{S_X^2} \right) = \frac{(1-f)}{n} S_Y^2 (1 - \rho^2).$$

- When b is estimated from the sample:

Estimator for b : $\hat{\beta} = \frac{s_{XY}}{s_x^2}$.

Variance: $\text{Var}[\hat{\mu}_{lr}(\hat{\beta})] \approx \text{Var}[\hat{\mu}_{lr}(\beta)] = ((1-f)/n) (S_Y^2 - S_{XY}^2/S_X^2)$

Variance estimator: $\hat{\text{Var}}[\hat{\mu}_{lr}(\hat{\beta})] = ((1-f)/n) ((n-1)/(n-2)) (s_y^2 - s_{xy}^2/s_x^2)$
 $= ((1-f)/n) ((n-1)/(n-2)) s_y^2 (1 - \hat{\rho}^2)$ where $\hat{\rho} = s_{xy}/(s_x s_y)$.