Sample Survey formulae

1. Simple random sampling (without replacement)

Notation: Population values = \{Y_1, \ldots, Y_N\}, \mu_Y = \sum_{i=1}^N Y_i/N, \tau_Y = \sum_{i=1}^N Y_i/\tau \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, \ldots, 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 mean \mu_Y

Estimator of \mu_Y is \hat{\mu}_Y = \bar{y}.

Its variance is \var(\hat{\mu}_Y) = \frac{S_Y^2}{n} (1 - f) \quad \text{where } f = \frac{n}{N}.

Variance estimator is \hat{\var}(\bar{y}) = \frac{s_Y^2}{n} (1 - f).

Estimating a proportion, P

Define a 1-0 variable Y so that P = \Pr(Y = 1). Then, \mu_Y = P and \sigma_Y^2 = P(1 - P).

Estimator of P is \hat{P} = \bar{p}, the sample proportion.

Its variance is \var(\hat{P}) = \frac{N-n}{N-1} \frac{P(1-P)}{n}.

Variance estimator is \hat{\var}(\bar{p}) = \frac{p(1-p)}{n-1} (1 - f).

Estimating a ratio, R = \frac{\mu_Y}{\mu_X}

Estimator of R is \hat{r} = \frac{\bar{y}}{\bar{x}}.

Its variance is \var(\hat{r}) \approx \frac{1-f}{n\mu_X^2} \frac{1}{N-1} \sum_{\ell=1}^N (Y_\ell - RX_\ell)^2.

Variance estimator is \hat{\var}(\hat{r}) = \frac{1-f}{n\mu_X^2} \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2.

If \mu_X is not known, the sample estimate \bar{x} is substituted in the denominator.

Sample size for estimating mean \mu_Y

\[ n = \frac{N\sigma_Y^2}{(N-1)D + \sigma_Y^2} \quad \text{where } D = [\text{required s.e.}(\hat{\mu}_Y)]^2. \]

\[ n_0 = \frac{\sigma_Y^2}{D}, \quad \text{ignoring fpc.} \]

\[ n = \frac{n_0}{1+(n_0-1)/N}. \]

2. Stratified Random Sampling

Notations: \(N_h, \mu_h, \sigma_h^2, S_h^2\) and \(P_h\) are as above but for stratum \(h\); \(W_h = N_h/N\).

\(n_h, \bar{y}_h, s_h^2, p_h\) and \(f_h = n_h/N_h\) are as above but for the subsample from stratum \(h\).

\(\mu_Y, N\) are as before and refers to the whole population. \(n\) is the whole sample size.

Estimation of \(\mu\)

Estimator of \(\mu\) is \(\hat{\mu}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h = \sum_{h=1}^L W_h \bar{y}_h\).

Its variance is \(\var(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 s_h^2 (1 - f_h) = \sum_{h=1}^L W_h^2 \frac{s_h^2}{n_h} - \sum_{h=1}^L W_h^2 \frac{s_h^2}{N_h}\)

Variance estimator is \(\hat{\var}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 \frac{s_h^2}{n_h} (1 - f_h).\)
Sample size for estimating $\mu_Y$

$$n = \frac{\sum_{h=1}^{L} (N_h^2 S_h^2 / w_h)}{N^2 V + \sum_{h=1}^{L} N_h S_h^2}$$

where $V = \text{var}(\hat{\mu})$, $w_h = \frac{n_h}{n}$

Proportional allocation

$$\frac{n_h}{n} = \frac{N_h}{N}, \quad \text{var}_{\text{prop}}(\hat{\mu}_{\text{st}}) = \frac{1}{n} \sum_{h=1}^{L} W_h S_h^2$$

Optimal allocation

For cost function $C = c_0 + \sum c_h n_h$,

- the cost $C$ is minimized for a specified variance $\text{var}(\hat{\mu}_{\text{st}})$, and
- the variance $\text{var}(\hat{\mu}_{\text{st}})$ is minimized for a fixed cost $C$

if $n_h \propto W_h S_h / \sqrt{c_h}$. That is,

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum (W_h S_h / \sqrt{c_h})}.$$  

Thus,

$$n = \frac{(C - c_0) \sum (N_h S_h / \sqrt{c_h})}{\sum (N_h S_h / \sqrt{c_h})}, \quad \text{if cost C is fixed.}$$

$$n = \frac{(\sum W_h S_h / \sqrt{c_h}) \left( \sum W_h S_h \sqrt{c_h} \right)}{V + (1/N) \sum W_h S_h^2}, \quad \text{if } V = \text{var}(\hat{\mu}_{\text{st}}) \text{ is fixed.}$$

Neyman allocation

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

This is optimal allocation when sample size is fixed and $c_h = c$ for all $h$.

Post-stratification

Let $m_h$ be the number of units in stratum $h$.

$$\text{var}_p(\hat{\mu}_{\text{st}}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{m_h} (1 - f_h) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{m_h} - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2.$$  

3. Cluster sampling

Notations:
Population: $\mu_Y =$ population mean per element, $M =$ number of elements in population;
Clusters: $N =$ number of clusters in population, $m_h =$ size of cluster $h$,

$\mu_h =$ mean of cluster $h$. $\sigma_h^2 =$ variance of cluster $h$. $\sigma_b^2 =$ between cluster variance.
Sample: $n =$ number of clusters in sample, $\bar{y}_i =$ mean of cluster $i$ in sample.
One-stage cluster sampling with equal-sized clusters

Estimator of $\mu_Y$ is $\hat{\mu}_{cl} = \frac{1}{n} \sum_{h=1}^{n} \bar{y}_h$.

Its variance is $\text{var}(\hat{\mu}_{cl}) = \frac{S_t^2}{n}(1 - \frac{n}{N})$ where $S_t^2 = \frac{N\sigma^2}{N-1}$.

Variance estimator is $\hat{\text{var}}(\hat{\mu}_{cl}) = \frac{s_t^2}{n}(1 - \frac{n}{N})$,
where $s_t^2$ is the variance of the selected cluster means.

One-stage cluster sampling with unequal-sized clusters

Let $t_i = m_i\bar{y}_i$ denote the cluster total of cluster $i$ in the sample.

Estimator of $\mu_Y$ is $\hat{\mu}_{clr} = \bar{t}/\bar{m}$ where $\bar{t} = (\sum_i t_i)/n, \bar{m} = (\sum_i m_i)/n$.

Its variance estimator is $\text{var}(\hat{\mu}_{clr}) = \frac{1-f}{nM^2} \frac{1}{n-1} \sum_{l=1}^N (\bar{t}_l - \hat{\mu}_{clr}m_l)^2$

$= \frac{1-f}{nM^2} \frac{1}{n-1} \left( \sum_i t_i^2 - 2\hat{\mu}_{clr} \sum_i t_im_i + \hat{\mu}_{clr}^2 \sum_i m_i^2 \right)$.

where $\hat{M} = (\sum_h m_h)/N$, the population mean cluster size.

Estimating population total $\tau_Y$
When $M$ is known, $\hat{\tau}_Y = M\hat{\mu}$, ($\hat{\mu}$ may be $\hat{\mu}_{cl}$ or $\hat{\mu}_{clr}$.)

When $M$ is unknown, $\hat{\tau}_Y = NL$.

4. Ratio and regression estimators

SRS with data $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ collected. $\mu_X$ is known.

Ratio estimator

Estimator of $\mu_Y$ is $\hat{\mu}_{\text{ratio}} = r\hat{\mu}_X$ where $r = \bar{y}/\bar{x}$.

Its variance is $\text{var}(\hat{\mu}_{\text{ratio}}) \approx \frac{1-f}{n} \frac{1}{1 - \frac{1}{n}} \sum_{l=1}^N (\bar{Y}_l - RX_l)^2$

Variance estimator is $\hat{\text{var}}(\hat{\mu}_{\text{ratio}}) = \frac{1-f}{n} \frac{1}{1 - \frac{1}{n}} \sum_{i=1}^n (y_i - rx_i)^2 = \frac{1-f}{n} (s_y^2 - 2rs_{xy} + r^2s_x^2)$

$= \frac{1-f}{n} (s_y^2 - 2r\hat{\rho}s_ys_y + r^2s_x^2)$ where $\hat{\rho} = s_{xy}/(s_ys_y)$

Regression estimator

Estimator of $\mu_Y$ is $\hat{\mu}_r = \bar{y} + b(\mu_X - \bar{x})$.

• When $b = b_0$ is pre-assigned:
  The variance is $\text{var}(\hat{\mu}_r) = \frac{1-f}{n} (S_Y^2 - 2b_0S_{XY} + b_0^2S_X^2)$

  Variance estimator is $\hat{\text{var}}(\hat{\mu}_r) = \frac{1-f}{n} (s_y^2 - 2b_0s_{xy} + b_0^2s_x^2)$.

  The best value to assign $b$ is $\beta = \frac{S_{XY}}{S_X}$, which gives minimum $\text{var}(\hat{\mu}_r)$.

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

• When $b$ is estimated from the sample:
  Use $b = \hat{\beta} = \frac{s_{XY}}{S_X}$.

  The variance is $\text{var}[\hat{\mu}_r(\hat{\beta})] \approx \text{var}[\hat{\mu}_r(\beta)] = \frac{1-f}{n} \left( S_Y^2 - \frac{S_{XY}^2}{S_X^2} \right)$

  Variance estimator is $\hat{\text{var}}[\hat{\mu}_r(\hat{\beta})] = \frac{1-f}{n} \frac{n-1}{n-2} \left( s_y^2 - \frac{s_{XY}^2}{s_x^2} \right) = \frac{1-f}{n} \frac{n-1}{n-2} s_y^2 (1 - \hat{\rho}^2)$.

  where $\hat{\rho} = \frac{s_{xy}}{s_ys_y}$. 

Sample Survey Formulae — page 3