References

1. Cochran, W.G. (1963) *Sampling Techniques*, Wiley, New York.
2. Kish, L. (1995) *Survey Sampling*, Wiley Inter. Science.
3. Lohr, S.L. (1999) *Sampling: Design and Analysis*, Duxbury Press.
4. McLennan, W. (1999) *An Introduction to Sample Surveys*, A.B.S. Publications, Canberra.

Section outline

1. Simple random samples and stratification.
   Finite population correction factor. Sample size determination. Inference over subpopulations.
2. Stratified sampling.
   Optimal allocation.
3. Ratio and regression estimators.
   Ratio estimators. Hartley-Ross estimator. Ratio estimator for stratified samples. Regression estimator.
4. Systematic sampling and cluster sampling.
5. Sampling with unequal probabilities.
   Probability proportional to size(PPS) sampling. The Horvitz-Thompson estimator.

1 Simple Random Samples (SRS)

1.1 The Population

We have a finite number of elements, N where N is *assumed known*. The population is $Y_1 \ldots Y_N$, where Y_i is a numerical value associated with i-th element. We adopt the notation where capital letters refer to characteristics of the population; small letters are used for the corresponding characteristics of a sample.

Population Total: $Y = \sum_{i=1}^{N} Y_i,$

Population Mean: $\mu = \bar{Y} = \dfrac{Y}{N} = \dfrac{1}{N}\sum_{i=1}^{N} Y_i,$

Population Variance : $\sigma^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ and

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

These are fixed (population) quantities, to be estimated.

If we have to consider two numerical values: (Y_i, X_i), $i = 1, \cdots, N$ an additional population quantity of interest is

$$\boxed{R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}}$$

the **ratio of totals.**

1.2 Simple Random Sampling

Focus on the numerical values $Y_i, \quad i = 1, \cdots, N$. A random sample of size n is taken *without replacement* : the observed values $y_1, \cdots, y_n$ are *random variables* and are *stochastically dependent*. The *sampling frame* is a list of the values Y_i, $i = 1, \cdots, N$.

$$\boxed{\begin{aligned} &\text{The } \textit{natural estimator} \text{ for } \bar{Y} \text{ is} \quad \widehat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \\ &\text{and hence for } Y = N\mu \text{ is} \qquad \widehat{Y} = N\bar{y}. \end{aligned}}$$

Distributional properties of $\bar{y}$ are complicated by the dependence of the y_i's.

$$\boxed{\textit{Sample variance:} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Fundamental Results

$$\boxed{\begin{aligned} &E(\bar{y}) = \mu. \\ &\mathrm{Var}(\bar{y}) = \big(1 - \frac{n}{N}\big)\frac{S^2}{n} = (1-f)\frac{S^2}{n} = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \\ &\mathrm{var}(\bar{y}) = \big(1 - \frac{n}{N}\big)\frac{s^2}{n} \\ &E(s^2) = S^2 \end{aligned}}$$

where f is the sampling fraction and the finite population correction (f.p.c.) is $1 - f$.

Proof: Let $\overline{y} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i = \frac{1}{n}\sum_{i=1}^{N} y_i I_i$ where the sample membership indicator

$$I_i = \begin{cases} 1 & \text{if element } i \text{ is in the sample,} \\ 0 & \text{if otherwise.} \end{cases}$$

First, we have

$$\begin{aligned}
E(I_i) &= 0 \times \Pr(I_i = 0) + 1 \times \Pr(I_i = 1) = \pi_i = \frac{n}{N}, \\
E(I_i^2) &= 0^2 \times \Pr(I_i = 0) + 1^2 \times \Pr(I_i = 1) = \pi_i = \frac{n}{N}, \\
E(I_i I_j) &= 0 \times 0 \Pr(I_i = 0 \& I_j = 0) + 0 \times 1 \Pr(I_i = 0 \& I_j = 1) + \\
&\quad 1 \times 0 \Pr(I_i = 1 \& I_j = 0) + 1 \times 1 \Pr(I_i = 1 \& I_j = 1) \\
&= \pi_{ij} = \frac{n(n-1)}{N(N-1)} \\
\text{Var}(I_i) &= E(I_i^2) - E^2(I_i) = \pi_i(1-\pi_i) = \frac{n}{N}(1 - \frac{n}{N}), \\
\text{Cov}(I_i, I_j) &= E(I_i I_j) - E(I_i)E(I_j) = \pi_{ij} - \pi_i \pi_j = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2.
\end{aligned}$$

Then

$$E(\overline{y}) = \frac{1}{n}\sum_{i=1}^{N} y_i E(I_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \cdot \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \overline{Y}. \quad \text{Unbiased}$$

$$E[\text{var}(\overline{y})] \overset{def.}{=} E\left[\left(\frac{N-n}{N}\right)\frac{s_y^2}{n}\right] \overset{Pf.2\rightarrow}{=} \left(\frac{N-n}{N}\right)\frac{S_y^2}{n} \overset{\leftarrow Pf.1}{=} \text{Var}(\overline{y}) \quad \text{Unbiased.}$$

Proof 1: show that $\text{Var}(\overline{y}) = \left(\frac{N-n}{N}\right)\frac{S_y^2}{n}$.

$$
\begin{aligned}
\mathrm{Var}(\overline{y}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N} y_i I_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \mathrm{Var}(I_i) + 2\sum_i \sum_{j,i<j} y_i y_j \mathrm{Cov}(I_i, I_j)\right] \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{N} y_i^2 \left[\frac{n}{N}\left(1-\frac{n}{N}\right)\right] + 2\sum_i \sum_{j,i<j} y_i y_j \left[\frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2\right]\right\} \\
&= \frac{n}{n^2}\left\{\frac{1}{N}\left(1-\frac{n}{N}\right)\sum_{i=1}^{N} y_i^2 + 2\frac{1}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right)\sum_i \sum_{j,i<j} y_i y_j\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\frac{1}{N}\sum_{i=1}^{N} y_i^2 + 2\frac{1}{N}\left(1-\frac{n}{N}\right)^{-1}\frac{N(n-1)-n(N-1)}{N(N-1)}\sum_i \sum_{j,i<j} y_i y_j\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\left\{\frac{1}{N}\sum_{i=1}^{N} y_i^2 + 2\frac{1}{N}\frac{N}{N-n}\frac{n-N}{\underline{N(N-1)}}\sum_i \sum_{j,i<j} y_i y_j\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{\underline{N-1}}\left\{\frac{N-1}{N}\sum_{i=1}^{N} y_i^2 - 2\frac{1}{N}\sum_i \sum_{j,i<j} y_i y_j\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left\{\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i^2 + 2\sum_i \sum_{j,i<j} y_i y_j\right)\right\} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\left\{\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} y_i\right)\right\} \\
&= \left(1-\frac{n}{N}\right)\frac{S_y^2}{n} = \frac{N-n}{N}\frac{N}{N-1}\frac{\sigma_y^2}{n} = \left(\frac{N-n}{N-1}\right)\frac{\sigma_y^2}{n}
\end{aligned}
$$

Proof 2: show that $E(s_y^2) = S_y^2$ where $S_y^2 = \frac{1}{N-1}\sum\limits_{i=1}^{N}(y_i - \overline{Y})^2 = \frac{N}{N-1}\sigma_y^2$.

$$
E(s_y^2) \;=\; E\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2\right] = \frac{1}{n-1}E\left\{\sum_{i=1}^{n}[(y_i - \overline{Y}) - (\overline{y} - \overline{Y})]^2\right\}
$$

$$
\begin{aligned}
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(y_i-\overline{Y})^2 - n(\overline{y}-\overline{Y})^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(y_i-\overline{Y})^2 - nE(\overline{y}-\overline{Y})^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} \mathrm{Var}(y_i) - n\mathrm{Var}(\overline{y})\right] \\
&= \frac{1}{n-1}\left[n\sigma_y^2 - n\left(\frac{N-n}{N-1}\right)\frac{\sigma_y^2}{n}\right] \\
&= \left(n - \frac{N-n}{N-1}\right)\frac{\sigma_y^2}{n-1} = \left(\frac{nN-n-N+n}{N-1}\right)\frac{\sigma_y^2}{n-1} \\
&= \frac{N}{N-1}\sigma_y^2 = S_y^2
\end{aligned}
$$

Central Limit Property

For large sample size n ($n > 30$ say), and small to moderate f, we have the approximation

$$(\bar{y}-\mu)/\sqrt{\mathrm{Var}(\bar{y})} \sim \mathcal{N}(0,1).$$

Confidence Interval for $\mu = \bar{Y}$

Replacing S^2 by s^2, an approximate 95% C.I. for μ and Y are respectively

$$
\begin{aligned}
\bar{y} &\pm 1.96\frac{s}{\sqrt{n}}\sqrt{1-f}, \\
N(\bar{y} &\pm 1.96\frac{s}{\sqrt{n}}\sqrt{1-f})
\end{aligned}
$$

Read Tutorial 10 Q1.

Example: (Industrial firm) An industrial firm is concerned about the time spent each week by staff on certain tasks. The time-log sheets of a SRS of $n = 50$ employees show the average amount of time spent on these tasks is 10.31 hours, with a sample variance $s^2 = 2.25$. The company employs $N = 750$ staff. Estimate the total number of man-hours used each week on the tasks and construct a 95% CI for the estimate.

Solution: From $N = 750$ time-log sheets, a SRS of $n = 50$ sheets was obtained. The average amount of time used in the sample is $\overline{y} = 10.31$ hours/week. Since $n = 50$ is large, we take $z_{0.025} = 1.96$. Hence

$$
\begin{aligned}
\widehat{Y} &= N\overline{y} = 750 \times 10.31 = 7732.5 \text{ hours} \\
\text{var}(\widehat{Y}) &= N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n} = 750^2 \left(1 - \frac{50}{750}\right) \frac{2.25}{50} = 23,625 \\
\text{se}(\widehat{Y}) &= \sqrt{23,625} = 153.70 \\
95\% \text{ CI for } Y &= (N\overline{y} - 1.96 \times \text{se}(\widehat{Y}),\ N\overline{y} + 1.96 \times \text{se}(\widehat{Y})) \\
&= (7732.5 - 1.96 \times 153.70,\ 7732.5 + 1.96 \times 153.70) \\
&= (7431.2,\ 8033.8)
\end{aligned}
$$

1.3 Simple Random Sampling for Attributes.

The method for SRS can be applied to estimate the *total number*, or *proportion* (or %) of units which possess some qualitative attribute. Let this subset of the population be C.

Let

$$\begin{aligned} Y_i &= 1 \text{ if } i \in C \\ &= 0 \text{ if } i \notin C \end{aligned}$$

and similarly for y_i's .

Customary notation:

$$\begin{aligned} \mu &= \bar{Y} = P \text{ is the population proportion,} \\ Y &= \sum_{i=1}^{N} Y_i = NP \text{ is the population total count,} \\ \bar{y} &= p \text{ is the sample proportion.} \end{aligned}$$

Let $y = \sum_{i=1}^{n} y_i = np$, $Q = 1 - P$ and $q = 1 - p$.

Thus $$\boxed{\widehat{Y} = Np = Ny/n.}$$

Since

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2 = N\bar{Y} - N\bar{Y}^2 = NP(1-P),$$

we have

$$S^2 = NP(1-P)/(N-1) \approx P(1-P)$$

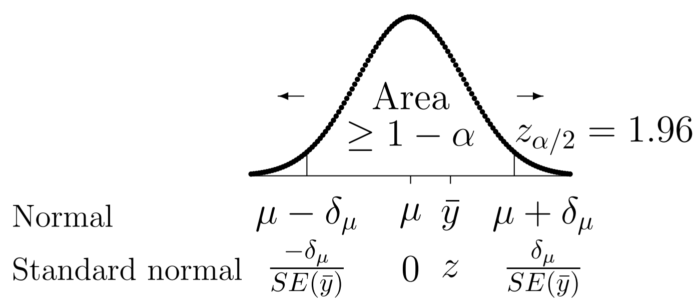
and similarly,

$$\begin{aligned} s^2 &= np(1-p)/(n-1) \approx p(1-p) \text{ and} \\ \frac{s^2}{n} &= \frac{np(1-p)}{n(n-1)} = \frac{p(1-p)}{n-1} \approx \frac{p(1-p)}{n} \end{aligned}$$

1.4 Sample Size Calculations

To calculate the sample size needed for sampling *yet to be carried out*, we want to be at least $100(1-\alpha)\%$ sure the estimate $\bar{y}$ of μ is within $100\delta\%$ of the actual value of μ (e.g. $1-\alpha = 0.95$, $z_{\alpha/2} = 1.96$). That is

$$
\begin{aligned}
& \Pr\{|\bar{y}-\mu| \le \delta_\mu\} \ge 1-\alpha \\
\Leftrightarrow\ & \Pr\left(\frac{|\bar{y}-\mu|}{\sqrt{\mathrm{Var}(\bar{y})}} \le \frac{\delta_\mu}{\sqrt{\mathrm{Var}(\bar{y})}}\right) \ \ge\ 1-\alpha \\
\Leftrightarrow\ & \delta_\mu/\sqrt{\mathrm{Var}(\bar{y})} \ge z_{\alpha/2} \\
\Leftrightarrow\ & \left(1-\frac{n}{N}\right)\frac{S^2}{n} \le \frac{(\delta_\mu)^2}{z^2_{\alpha/2}} \\
\Leftrightarrow\ & \frac{S^2}{n} - \frac{S^2}{N} \le \frac{(\delta_\mu)^2}{z^2_{\alpha/2}} \\
\Leftrightarrow\ & \frac{S^2}{n} \le \frac{(\delta_\mu)^2}{z^2_{\alpha/2}} + \frac{S^2}{N} \\
\Leftrightarrow\ & n \ge \frac{S^2}{\frac{(\delta_\mu)^2}{z^2_{\alpha/2}} + \frac{S^2}{N}} \\
\Leftrightarrow\ & \boxed{n \ge \frac{NS^2}{N(\delta_\mu)^2/z^2_{\alpha/2} + S^2}}
\end{aligned}
$$



Ignoring f.p.c. (i.e. taking $f = 0$ or $\mathrm{Var}(\bar{y}) = S^2/n$ when N is unknown)

$$\boxed{n \ge \frac{z^2_{\alpha/2}S^2}{(\delta_\mu)^2} \approx \frac{z^2_{\alpha/2}s^2}{(\delta_{\bar{y}})^2}}$$

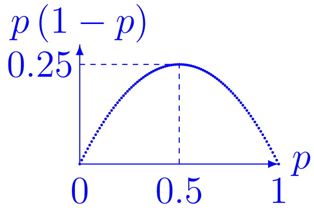
where s^2 and $\bar{y}$ are estimates from a pilot survey and $s^2 \approx p(1-p)$ for attributes.

Example: (blood group)

1. What size sample must be drawn from a population of size $N = 800$ in order to estimate the proportion with a given blood group to within 0.04 (i.e. an absolute error of 4%) with probability 0.95?
2. What sample size is needed if we know that the blood group is present in no more than 30% of the population?

Solution:

1. We have $N = 800$, $S^2 \simeq \hat{p}(1-\hat{p}) = 0.5^2 = 0.25$, $\delta_\mu = 0.04$ and $z_{0.025} = 1.96$. Note that $S^2 = p(1-p)$ is max at $p = 0.5$ and this gives the most conservative n.



$$n \geq \frac{NS^2}{N\delta_\mu^2/z_{\alpha/2}^2 + S^2} = \frac{800(0.25)}{800(0.04^2)/1.96^2 + 0.25} = 342.93$$

Take $n = 343$.

2. Now $S^2 = p(1-p) = 0.3(0.7) = 0.21$.

$$n \geq \frac{NS^2}{N\delta_\mu^2/z_{\alpha/2}^2 + S^2} = \frac{800(0.21)}{800(0.04^2)/1.96^2 + 0.21} = 309.28$$

Take $n = 310$.

Read Tutorial 10 Q2.

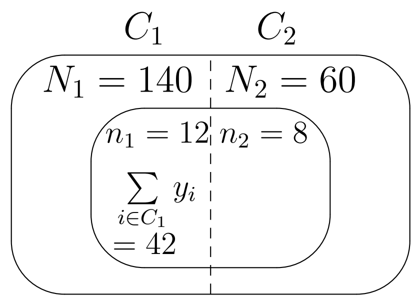
1.5 Inference over Subpopulations-Poststratification

Motivating example: (dentist) There are 200 children in a village. One dentist takes a simple random sample of 20 and finds 12 children with at least one decayed tooth and a total of 42 decayed teeth. Another dentist quickly checks all 200 children and finds 60 with no decayed teeth.

Estimate the total number of decayed teeth.

C_1: ≥ 1 decayed teeth	C_2: no decayed teeth	Total
$N_1 = 140$	$N_2 = 60$	$N = 200$
$n_1 = 12$	$n_2 = 8$	$n = 20$

$$\sum_{i \in C_1} y_i = 42 = \sum_{i=1}^{n} y_i'$$



From a population of N individuals, *one* simple random sample of n individuals y_i, $i = 1, \cdots, n$ is drawn. Separate estimates might be wanted for one of a number of subclasses $\{C_1, C_2, \cdots\}$ which are *subsets* of the population (sampling frame) using *post-stratification*.

Reasons:

1. Unavailability of a suitable sampling frame for each stratum even though the stratum sizes $N_1, \ldots, N_L$ are often obtainable from official statistics.

2. Inability to classify population elements into an appropriate stratum without actual contact,

   e.g. personal characteristics such as educational level and political preference and household characteristics such as owned/rented accommodation, income level and household size are unknown

3. Multi-variate and multi-purpose nature of most surveys.

4. Post-stratification is to correct the distorted sample proportion due to non-response.

Example:

POPULATION (SAMPLING FRAME)	SUBPOPULATION
Australian population	unemployed Queenslanders
retailers	supermarkets
the employed	the employed working overtime

Solution: The estimate for the *overall* average no. of decay teeth using *overall sample mean* is

$$\hat{\bar{Y}} = \bar{y}' = \frac{1}{n}\sum_{i\in C_1} y_i = \frac{42}{20} = 2.1.$$

The estimate for the *overall* or *conditional* total number of decayed teeth, Y is

$$\hat{Y}_{pst,1m1} = N\bar{y}' = 200 \times 2.1 = 420,$$

ignoring the information of $n_1 = 12$ and $N_1 = 140$ from the second dentist. Using these information and condition on *those with at least one decayed teeth*, the average no. of decay teeth is

$$\hat{\bar{Y}}_{pst,1m1} = \frac{N\bar{y}'}{N_1} = \frac{200(2.1)}{140} = 3$$

Alternative estimate for the average no. of decayed teeth using the *conditional sample mean* is

$$\hat{\bar{Y}}_{pst,1m2} = \bar{y}_1 = \frac{n\bar{y}'}{n_1} = \frac{42}{12} = 3.5$$

Hence the estimate for the total no. of decayed teeth is

$$\hat{Y}_{pst,1m2} = N_1\bar{y}_1 = 140(3.5) = 490$$

Read Tutorial 10 Q3.

Formulae: Denote the total number of items in class C_l by N_l. Note that N_l is generally *unknown* but we can *estimate* N_l by

$$\boxed{\widehat{N}_l = N n_l / n}$$

where $w_l = n_l/n$ is the sample proportion of units falling into C_l and the corresponding population proportion is $W_l = N_l/N$.

The same technique is used to estimate population mean & total in C_l:

$$\bar{Y}_l = \sum_{i \in C_l} y_i / N_l, \quad \text{and} \quad Y_l = \sum_{i \in C_l} y_i$$

using

$$y'_i = \begin{cases} y_i & \text{if } i \in C_l \\ 0 & \text{if } i \notin C_l \end{cases}$$

Data	C_1			C_2			Mean
y_i :	y_1	y_2	$\dots$	y_{n_1+1}	y_{n_1+2}	$\dots$	$\bar{y}$
y'_i :	y_1	y_2	$\dots$	0	0	$\dots$	$\bar{y}'_1$
y_i :	y_1	y_2	$\dots$	$-$	$-$	$\dots$	$\bar{y}_1$

Then $\boxed{\bar{y}' = \sum_{i=1}^{n} y'_i / n \text{ estimates the mean } \bar{Y}' = \sum_{i=1}^{N} y'_i / N.}$

The unbiased estimator for the total $Y_l = N\bar{Y}' = \sum_{i=1}^{N} Y'_i$ in C_l and its variance are

$$\boxed{\widehat{Y}_{pst,lm1} = N\bar{y}' = N \sum_{i=1}^{n} y'_i / n \quad \text{and} \quad \text{var}(\widehat{Y}_{pst,lm1}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_l'^2}{n}}$$

1. When N_l is *known*, the unbiased estimator for mean $\bar{Y}_l = Y_l/N_l$ in C_l and its variance estimate are

$$\boxed{\widehat{\bar{Y}}_{pst,lm1} = N\bar{y}'/N_l = \bar{y}'/W_l \quad \text{and} \quad \text{var}(\widehat{\bar{Y}}_{pst,lm1}) = \frac{1}{W_l^2} \left(1 - \frac{n}{N}\right) \frac{s_l'^2}{n}}$$

where $S_l'^2$ can be estimated by $s_l'^2$:

$$S_l'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i' - \bar{Y}')^2 \quad \text{and} \quad s_l'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i' - \bar{y}')^2.$$

Note: the random variable n_l are not included in these calculations.

2. When N_l is *unknown*, we can estimate the total Y_l by $N\bar{y}'$ but we cannot estimate $\bar{Y}_l$ by $\bar{y}'/W_l$. A natural way is to estimate $W_l = \dfrac{N_l}{N}$ by $w_l = \dfrac{n_l}{n}$:

$$\boxed{\hat{\bar{Y}}_{pst,lm2} = n\bar{y}'/n_l = \bar{y}_l \quad \text{and} \quad \text{var}(\hat{\bar{Y}}_{pst,lm2}) \simeq \frac{1}{W_l}\left(1 - \frac{n}{N}\right)\frac{{s_l}^2}{n}}$$

Similarly the total estimator and its variance estimate are

$$\boxed{\hat{Y}_{pst,lm2} = N_l\bar{y}_l = N_l n\bar{y}'/n_l \quad \text{and} \quad \text{var}(\hat{Y}_{pst,lm2}) \simeq N^2 W_l \left(1 - \frac{n}{N}\right)\frac{{s_l}^2}{n}}$$

where $s_l^2 = \dfrac{1}{n_l - 1}\displaystyle\sum_{i\in C_l}(y_i - \bar{y}_l)^2$ is the sample variance in C_l and $\dfrac{N_l^2}{W_i} = N^2 W_l$.

Theorem: For the estimator $\hat{\bar{Y}}_l = \bar{y}_l = (n/n_l)\bar{y}'$,

$$\begin{aligned} E(\bar{y}_l) &= \bar{Y}_l \\ \text{Var}(\bar{y}_l) &\simeq \frac{1}{W_l}\left(1 - \frac{n}{N}\right)\frac{S_l^2}{n} \end{aligned}$$

provided we define $E(\bar{y}_l) = \bar{Y}_l$ when $n_l = 0$.

Proof: Using conditional expectation,

$$\begin{aligned}
E(\bar{y}_l) &= E_{n_l}\{E(\bar{y}_l|n_l)\} \\
&= E(\bar{y}_l|n_l=0)\Pr(n_l=0) + \sum_{i\geq 1} E(\bar{y}_l|n_l=i)\Pr(n_l=i) \\
&= \bar{Y}_l \Pr(n_l=0) + \sum_{i\geq 1} \bar{Y}_l \Pr(n_l=i) = \bar{Y}_l
\end{aligned}$$

where for each fixed n_l, a simple random sample of size n_l is drawn from C_l. Further using conditional variance,

$$\begin{aligned}
\mathrm{Var}(\bar{y}_l) &= \mathrm{Var}(\underbrace{E(\bar{y}_l|n_l)}_{\bar{Y}_l}) + E(\mathrm{Var}\,(\bar{y}_l|n_l)) = 0 + E(\mathrm{Var}(\bar{y}_l|n_l)) \\
&= E\left[\left(1-\frac{n_l}{N_l}\right)\frac{S_l^2}{n_l}\right] = E\left(\frac{S_l^2}{n_l} - \frac{S_l^2}{N_l}\right) \simeq \frac{S_l^2}{nW_l} - \frac{S_l^2}{NW_l} \\
&= \frac{1}{W_l}\left(1-\frac{n}{N}\right)\frac{S_l^2}{n}
\end{aligned}$$

since $E\left(\frac{1}{n_l}\right) \simeq \frac{1}{nW_l} + \frac{1-W_l}{n^2W_l^2} \simeq \frac{1}{nW_l}$ where the ignored term:

$$\text{extra var.} = \frac{1-W_l}{n^2W_l^2}S_l^2$$

is the extra variability due to the random sample size n_l.

Theorem: $E\left(\frac{1}{n_l}\right) \simeq \frac{1}{nW_l} + \frac{1-W_l}{n^2W_l^2}$.

Proof:
$$\begin{aligned}
\frac{1}{n_l} &= (n_l - nW_l + nW_l)^{-1} = \frac{1}{nW_l}\left(1 + \frac{n_l - nW_l}{nW_l}\right)^{-1} \\
&= \frac{1}{nW_l}\left[1 - \left(\frac{n_l - nW_l}{nW_l}\right) + \left(\frac{n_l - nW_l}{nW_l}\right)^2 - \ldots\right]
\end{aligned}$$

$$\text{Hence}\quad E\left(\frac{1}{n_l}\right) = \frac{1}{nW_l}\left[1-\frac{E(n_l-nW_l)}{nW_l}+\frac{E(n_l-nW_l)^2}{n^2W_l^2}-\ldots\right]$$
$$\simeq \frac{1}{nW_l}+\frac{1}{nW_l}\left(\frac{nW_l(1-W_l)}{n^2W_l^2}\right)=\frac{1}{nW_l}+\frac{1-W_l}{n^2W_l^2}$$

since $E(n_l)=nW_l$ and $Var(n_l)=nW_l(1-W_l)$.

Corollary: The estimator $\widehat{Y}_{pst,lm2}=N_l\,\bar{y}_l$ is unbiased for Y_l and has $\text{Var}\,(\widehat{Y}_l)=N_l^2\,\text{Var}(\bar{y}_l)$.

When N_l is known, the estimator $\widehat{Y}_{pst,lm2}=N_l\,\bar{y}_l$ uses more information (both N_l & n_l) than $N\bar{y}'$ and so is a better unbiased estimator.

Using this method, the *overall* total and mean estimates are:

$$\boxed{\widehat{Y}_{pst}=N\sum_{l=1}^{L}W_l\,\bar{y}_l \ \text{ and }\ \text{var}\,(\widehat{Y}_{pst})\simeq N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{l=1}^{L}W_ls_l^2}$$

and

$$\boxed{\widehat{\bar{Y}}_{pst}=\sum_{l=1}^{L}W_l\,\bar{y}_l \ \text{ and }\ \text{var}\,(\widehat{\bar{Y}}_{pst})\simeq \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{l=1}^{L}W_ls_l^2}$$

respectively since

$$\widehat{\bar{Y}}_{pst}=\frac{1}{N}\sum_{l=1}^{L}\widehat{Y}_{pst,lm2}=\frac{1}{N}\sum_{l=1}^{L}N_l\,\bar{y}_l=\sum_{l=1}^{L}W_l\,\bar{y}_l$$

$$\text{var}\,(\widehat{\bar{Y}}_{pst})=\sum_{l=1}^{L}W_l^2\,\text{var}(\bar{y}_l)\simeq\sum_{l=1}^{L}W_l^2\,\frac{1}{W_l}\left(1-\frac{n}{N}\right)\frac{s_l^2}{n}\simeq\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{l=1}^{L}W_ls_l^2$$

Compare $\widehat{\bar{Y}}_{pst}$ to $\quad \widehat{\bar{Y}}=\bar{y}=\sum_{l=1}^{L}w_l\,\bar{y}_l$ and $\text{var}\,(\widehat{\bar{Y}})=\left(1-\frac{n}{N}\right)\frac{s^2}{n}$

for one SRS without post-stratification, we have

W_l replaces w_l and $\sum_{l=1}^{L} W_l s_l^2$ replaces s^2 to correct sample proportions.

Read Tutorial 11 Q1,c,d.

Post-stratified estimator based on 1 SRS

Par.	N_l	Estimator	Variance
$\bar{Y}_l$	known	$\hat{\bar{Y}}_{pst,lm1} = N\bar{y}'/N_l$	$\text{var}(\hat{\bar{Y}}_{pst,lm1}) = \frac{1}{W_l^2}\left(1-\frac{n}{N}\right)\frac{{s'_l}^2}{n}$
Y_l	unknown	$\hat{Y}_{pst,lm1} = N\bar{y}'$	$\text{var}(\hat{Y}_{pst,lm1}) = N^2\left(1-\frac{n}{N}\right)\frac{{s'_l}^2}{n}$
$\bar{Y}_l$	unknown	$\hat{\bar{Y}}_{pst,lm2} = \bar{y}_l$	$\text{var}(\hat{\bar{Y}}_{pst,lm2}) \simeq \frac{1}{W_l}\left(1-\frac{n}{N}\right)\frac{{s_l}^2}{n}$
Y_l	known	$\hat{Y}_{pst,lm2} = N_l\bar{y}_l$	$\text{var}(\hat{Y}_{pst,lm2}) \simeq N^2 W_l\left(1-\frac{n}{N}\right)\frac{{s_l}^2}{n}$
$\overline{Y}$	known	$\hat{\overline{Y}}_{pst} = \sum_{l=1}^{L} W_l\overline{y}_l$	$\text{var}(\hat{\overline{Y}}_{pst}) \simeq \left(1-\frac{n}{N}\right)\sum_{l=1}^{L} W_l\frac{s_l^2}{n}$
Y	known	$\hat{Y}_{pst} = \sum_{l=1}^{L} N_l\overline{y}_l$	$\text{var}(\hat{Y}_{pst}) \simeq N^2\left(1-\frac{n}{N}\right)\sum_{l=1}^{L} W_l\frac{s_l^2}{n}$

where $y'_i = y_i$ if $i \in C_l$ & 0 otherwise, $\bar{y}' = \frac{1}{n}\sum_{i\in C_l} y_i$, $\bar{y}_l = \frac{1}{n_l}\sum_{i\in C_l} y_i$,

$${s'_l}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y'_i - \bar{y}')^2, \quad \text{and} \quad s_l^2 = \frac{1}{n_l-1}\sum_{i\in C_l}(y_i - \bar{y}_l)^2.$$