Related topics

Poststratification

If a stratified estimator is desired, but a simple random sample has been conducted instead, it is still possible to use a stratified estimator if the true population fractions $A_i = N_i/N$ are known. In this case, the estimator of the mean μ is:

$$\overline{y}_{st} = \sum_{i=1}^{L} A_i \overline{y}_i \ .$$

Even though this estimator is the same as $\overline{y}_{st}$ under proportional allocation, the variance is a bit more complicated. As shown in the text, an approximation to $E(\frac{1}{n_i})$ is used to obtain:

$$\widehat{V}_p(\overline{y}_{st}) \cong \left(\frac{N-n}{Nn}\right) \sum_{i=1}^{L} A_i s_i^2 + \frac{1}{n^2} \sum_{i=1}^{L} (1 - A_i) s_i^2$$

Example: Suppose that there are two strata with $n_1 = 20, \overline{y}_1 = 180, s_1 = 40, n_2 = 80, \overline{y}_2 = 110, s_2 = 25$, N is large and we know that $A_1 = A_2 = \frac{1}{2}$. Then we have $\overline{y}_{st} = (\frac{1}{2})(180) + (\frac{1}{2})(110) = 145$, and we can ignore the fpc to get:

$$\begin{aligned}
\widehat{V}_p(\overline{y}_{st}) &= \frac{1}{100}\left[(\frac{1}{2})(40)^2 + (\frac{1}{2})(25)^2\right] \\
&\quad + \frac{1}{100^2}\left[(1-\frac{1}{2})(40)^2 + (1-\frac{1}{2})(25)^2\right] \\
&= 11.125 + .11125 = 11.24,
\end{aligned}$$

giving a bound of $B = 2\sqrt{\widehat{V}_p(\overline{y}_{st})} = 6.70$. Notice that almost all of $\widehat{V}_p(\overline{y}_{st})$ comes from the first term of the sum. Poststratification works best when n and all n_i are large, which implies that the number of strata should not be very large.

Double Sampling

If the $A_i = N_i/N$ are unknown, then an alternate approach is to use double sampling. In double sampling a first large sample (Phase I) of size n' is taken, then we use the observed stratum sizes $a'_i = n'_i/n'$ as estimates

of the A_i. In Phase II of sampling, we sample n_i from the n_i' and measure the quantity of interest y_i from those units. Usually the information needed for stratification from the first phase is easy to obtain compared to the information collected in the second phase. From this data we estimate μ by:

$$\overline{y}_{st}' = \sum_{i=1}^{L} a_i' \overline{y}_i \ ,$$

with a variance estimate (assuming n' is large) of:

$$\widehat{V}(\overline{y}_{st}') \cong \sum_{i=1}^{L} \left[\frac{a_i' s_i^2}{n_i} + \frac{a_i'(\overline{y}_i - \overline{y}_{st}')^2}{n'} \right].$$

Example: Suppose for two strata we sample $n' = 500$ and get $n_1' = 240$ and $n_2' = 260$. Then on the second phase we get $n_1 = 20, \overline{y}_1 = 180, s_1 = 40, n_2 = 80, \overline{y}_2 = 110, s_2 = 25$. In this case we obtain $\overline{y}_{st}' = .48(180) + .52(110) = 143.6$ and

$$\begin{aligned}
\widehat{V}(\overline{y}_{st}') &= \left[\frac{(.48)^2(40)^2}{20} + \frac{.48(180-143.6)^2}{500} \right] \\
&+ \left[\frac{(.52)^2(25)^2}{80} + \frac{.52(110-143.6)^2}{500} \right] \\
&= 22.99,
\end{aligned}$$

giving a bound of $B = 2\sqrt{\widehat{V}(\overline{y}_{st}')} = 9.59$. This example is fairly similar to the one above for poststratification, and it illustrates the increase in variance that occurs due to double sampling.