Stratified Sampling

Statistics 110

Summer 2006

Stratified Sampling

As we've seen, we can get more precise estimates by changing the estimator of a parameter (e.g. $\bar{Y}_R$ vs $\bar{Y}$). We can also get more precise estimation by changing the sampling scheme.

Stratified sampling is one approach that may (but not always) give more precise estimates. Its based on the idea of iterated expectations. Let Y be a discrete random variable taking values $y_1, y_2, \ldots, y_L$ with probabilities $p_1, p_2, \ldots, p_L$. Then

$$E[X] = E[E[X|Y]] = \sum_{l=1}^{L} E[X|Y = y_l]p_l$$

Suppose your that the population can be broken up into L groups, known as **Strata**. Suppose that for stratum l, there are N_l units from the population ($\sum_{l=1}^{L} N_L = N$) and the value for the units in stratum l are $x_{1l}, x_{2l}, \ldots, x_{N_l l}$.

Let

$$W_l = \frac{N_l}{N} \qquad \mu_l = \frac{1}{N_l}\sum_{i=1}^{N_l} x_{il}$$

Then

$$\mu = \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l} x_{il} = \frac{1}{N}\sum_{l=1}^{L} N_l\mu_l = \sum_{l=1}^{L} W_l\mu_l$$

Then instead of taking a SRS of n units from the total population, we can take a SRS of size n_l from each stratum ($\sum_{l=1}^{L} n_l = n$).

Here $\mu_l = E[X|\text{Stratum } l]$ and $W_l = P[\text{Stratum } l]$, so the overall mean satisfies the setup of an iterated expectation.

Examples of situations where you might want to use stratified sampling

- In the soy bean yield example, you could stratify on farm size: Small ($<$ 100 acres), Medium (between 100 and 200 acres), Large($>$ 200 acres).
- Health care costs - stratify on age
- Income of university graduates - stratify on major
- TV ratings - stratify on age and gender ($+$???)

When choosing a factor to stratify on, you want something that is associated by the variable of interest as this should make $\mathrm{Var}(X|\text{Stratum } l)$ small.

Setup: Let $X_{1l}, X_{2l}, \ldots, X_{n_l l}$ be the sample from stratum l and

$$\bar{X}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_{il} \qquad S_l^2 = \frac{1}{n_l - 1}\sum_{i=1}^{n_l}(X_{il} - \bar{X}_l)^2$$

be the sample mean and variance.

Then an estimate of the population mean μ is

$$\bar{X}_S = \sum_{l=1}^{L} \frac{N_l}{N}\bar{X}_l = \sum_{l=1}^{L} W_l \bar{X}_l$$

Does this approach work?

Theorem. *$\bar{X}_S$ is an unbiased estimate of μ, i.e. $E[\bar{X}_S] = \mu$.*

Proof.

$$E[\bar{X}_S] = \sum_{l=1}^{L} W_l E[\bar{X}_l] = \sum_{l=1}^{L} W_l \mu_l = \mu$$

□

Theorem.

$$\mathrm{Var}(\bar{X}_S) = \sum_{l=1}^{L} W_l^2 \frac{1}{n_l}\left(1 - \frac{n_l - 1}{N_l - 1}\right)\sigma_l^2$$

where

$$\sigma_l^2 = \frac{1}{N_l}\sum_{i=1}^{N_l}(x_{il} - \mu_l)^2$$

Proof. Since the $\{\bar{X}_l\}$ are independent,

$$\mathrm{Var}(\bar{X}_S) = \sum_{l=1}^{L} W_l^2 \mathrm{Var}(\bar{X}_l) = \sum_{l=1}^{L} W_l^2 \frac{1}{n_l}\left(1 - \frac{n_l - 1}{N_l - 1}\right)\sigma_l^2$$

□

Now whether stratified sampling is preferred depends on whether

$$\mathrm{Var}(\bar{X}_S) < \mathrm{Var}(\bar{X}) = \frac{\sigma_X^2}{n}\left(1 - \frac{n-1}{N-1}\right)$$

This depends on the choice of sample sizes $\{n_l\}$, the variation of the strata means $\{\mu_l\}$, and the strata variances $\{\sigma_l^2\}$

There are a number of different allocation schemes that can be used, some better than others

- Equal allocation: $n_1 = n_2 = \ldots = n_L = \frac{n}{L}$.
- Proportional allocation: $\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_L}{N_L}$ which leads to

$$n_l = n\frac{N_l}{N} = nW_l$$

- Optimal allocation: Choose $n_1, \ldots, n_L$ to minimize $\mathrm{Var}(\bar{X}_S)$ for a given total sample size n, which gives

$$n_l = n\frac{W_l\sigma_l}{\sum_{k=1}^{L} W_k\sigma_k}$$

  Note that this allocation scheme leads to more observations in the more variable strata.

Lets assume that for all of these schemes $n_l \ll N_l$ for all l so the $FPC \approx 1$ so we can ignore it. Also, lets ignore that the formulas may not give integer samples sizes. In that case round to the nearest integer when implementing.

The the sampling variance for these allocation schemes is

- Equal allocation:

$$\mathrm{Var}(\bar{X}_{SE}) = \frac{L}{n}\sum_{l=1}^{L} W_l^2\sigma_l^2$$

- Proportional Allocation:

$$\mathrm{Var}(\bar{X}_{SP}) = \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2$$

- Optimal Allocation:

$$\mathrm{Var}(\bar{X}_{SO}) = \frac{1}{n}\left(\sum_{l=1}^{L} W_l \sigma_l\right)^2$$

Before showing when stratified sampling works better, we need to figure out the population variance. One way to get this is by

$$\begin{aligned}
\mathrm{Var}(X) &= E[\mathrm{Var}(X|\text{Stratum})] + \mathrm{Var}(E[X|\text{Stratum}]) \\
&= \sum_{l=1}^{L} W_l \sigma_l^2 + \sum_{l=1}^{L} W_l (\mu_l - \mu)^2 \\
&= \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} (x_{il} - \mu)^2 = \sigma^2
\end{aligned}$$

Theorem.

$$\mathrm{Var}(\bar{X}_{SP}) \leq \mathrm{Var}(\bar{X})$$

That is, proportional sampling never does worse than a single SRS of the same total sample size n.

Proof.

$$\begin{aligned}\mathrm{Var}(\bar{X}) - \mathrm{Var}(\bar{X}_{SP}) &= \frac{1}{n}\left(\sum_{l=1}^{L} W_l\sigma_l^2 + \sum_{l=1}^{L} W_l(\mu_l - \mu)^2\right) - \frac{1}{n}\sum_{l=1}^{L} W_l\sigma_l^2 \\ &= \frac{1}{n}\sum_{l=1}^{L} W_l(\mu_l - \mu)^2 = \frac{1}{n}\mathrm{Var}(E[X|\text{Stratum}]) \geq 0\end{aligned}$$

□

This result implies that the more separated the strata are, in terms of strata means, the better proportional sampling will do.

This implies that when trying to pick a stratification variable, you want to pick one that is strongly associated with your variable of interest.

The advantage of optimal allocation over proportional allocation can be seen with

$$\begin{aligned}
\mathrm{Var}(\bar{X}_{SP}) - \mathrm{Var}(\bar{X}_{SO}) &= \frac{1}{n}\sum_{l=1}^{L} W_l \sigma_l^2 - \frac{1}{n}\left(\sum_{l=1}^{L} W_l \sigma_l\right)^2 \\
&= \frac{1}{n}\sum_{l=1}^{L} W_l (\sigma_l - \bar{\sigma})^2 \\
&= \frac{1}{n}\mathrm{Var}(SD(X|\text{Stratum})) \geq 0
\end{aligned}$$

This result implies that you get a bigger advantage from optimal allocation over proportional allocation when the variability in the different strata is highly variable.

Usually the gains from going from a single SRS to proportion allocation is much bigger than going from proportional allocation to optimal allocation.

There is also the additional problem in that you need to know the variances for each strata to get the sample sizes which is problematic. However for proportional sampling, you only need to know the fraction of units falling into each strata. This information is much more readily available, or at least easier to approximate.

Related to this, is calculation of standard errors. As the strata variances usually aren't available, they need to estimated with the stratum sample variances S_l^2. Using these gives the estimated variance

$$S^2_{\bar{X}_S} = \sum_{l=1}^{L} W_l^2 \frac{1}{n_l}\left(1 - \frac{n_l}{N_l}\right) S_l^2$$

This gives the following CI for μ

$$\bar{X}_S \pm z(\alpha/2) S_{\bar{X}_S}$$

Example: TV viewing time per household for 3 towns

	Town A	Town B	Town C
N_l	155	62	93
W_l	0.5	0.2	0.3
n_l	20	8	12
$\bar{X}_l$	33.900	25.125	19.000
S_l^2	35.358	232.411	87.636

$$\bar{X}_S = 0.5 \times 33.900 + 0.2 \times 25.125 + 0.3 \times 19.000 = 27.675$$

$$\begin{aligned} S^2_{\bar{X}_S} &= 0.5^2 \, \frac{135}{155} \, \frac{35.358}{20} + 0.2^2 \, \frac{54}{62} \, \frac{232.411}{8} + 0.3^2 \, \frac{81}{93} \, \frac{87.636}{12} \\ &= 1.97 \end{aligned}$$

So a 95% CI for the mean household viewing time is

$$27.675 \pm 1.96\sqrt{1.97} = 27.675 \pm 2.751$$

Now suppose that the sample strata means and variances are the true population strata means and variances. Lets see the advantages of stratified sampling under this assumption.

This gives $\sigma^2 = 133.7045$. Then for a sample of size $n = 40$,

$$S^2_{\bar{X}} = \frac{270}{310}\,\frac{133.7045}{40} = 2.92$$

So the efficiency of this stratified sampling scheme is

$$\frac{S^2_{\bar{X}}}{S^2_{\bar{X}_S}} = \frac{2.92}{1.97} = 1.48$$

So this stratified sampling scheme is as about as efficient as a SRS of 59 households ($= 40 \times 1.48$).

Under this same assumption, optimal allocation is

	Town A	Town B	Town C
Optimal n_l	13	14	13
Actual n_l	20	8	12
N_l	155	62	93
W_l	0.5	0.2	0.3
$\bar{X}_l$	33.900	25.125	19.000
S_l^2	35.358	232.411	87.636

Note that a proportional scheme was used in the study.

If this optimal allocation was used, $S_{\bar{X}^2_{SO}} = 1.66$, giving an efficiency

$$\frac{S^2_{\bar{X}_S}}{S^2_{\bar{X}_{SO}}} = \frac{1.97}{1.66} = 1.19$$

To get the same variance with optimal allocation would need about 34 obs.