STA304

Neil Montgomery
2016-03-03

stratified sampling: poststratification

Weights known, strata unavailable

Stratification can be practically difficult. The frame may not contain the required information for partition into strata.

But the weights W_i might be known. (This is key.) (Note: the book finally gets around to adopting the notion of weight and calls the weights A_i .))

For example, in Canada the male and female proportions are 0.496 and 0.504 respectively. (Many other population-level proportions are known as well.) But it may not be possible to stratify by sex.

It can be suitable to perform a simple random sample and divide the sample up into groups, adjusting the population parameter estimate accordingly.

Poststratification illustration

For example, a Statistics Canada regularly compiles salary data and publishes results by sex. Suppose in one particular survey the SRS results are as follows (in 000's of dollars)

Sex	n	mean	var	sd
Female	550	26.56	290.62	17.05
Male	450	42.12	1112.31	33.35

The SRS population mean income would be 33.56.

But we know the SRS sub-sample sizes are off. Here is the *poststratified* estimate of the mean income, reweighted for the known true weights:

$$\bar{y}_{post} = W_1\bar{y}_1 + W_2\bar{y}_2 = 0.504 \cdot 26.56 + 0.496 \cdot 42.12 = 34.28$$

The question is…what is $V(\bar{y}_{post})$?

poststratified variance - I

When the n_i are fixed we have from before (CORRECTED - $1/N$ was missing):

$$
\begin{aligned}
\hat{V}(\bar{y}_{st}) &= \sum_{i=1}^{L} W_i^2 \frac{s^2}{n_i} \frac{N_i - n_i}{N_i} \\
&= \sum_{i=1}^{L} W_i^2 \frac{s^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) \\
&= \sum_{i=1}^{L} W_i^2 \frac{s^2}{n_i} - \frac{1}{N} \sum_{i=1}^{L} W_i s_i^2
\end{aligned}
$$

What is fundamentally different this time?

poststratified variance - I

The procedure is to replace $1/n_i$ with $E(1/n_i)$. This is difficult to evaluate but can be approximated by:

$$E\left(\frac{1}{n_i}\right) = \frac{1}{nW_i} + \frac{1 - W_i}{n^2 W_i^2}$$

Essentally "almost what we expect" plus "something that might be small". The approximation is good as long as n is large and the weights are not too small.The resulting formula (see book for the three line derivation) is:

$$\hat{V}(\overline{y}_{post}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\sum_{i=1}^{L} W_i s_i^2 + \frac{1}{n^2}\sum_{i=1}^{L}(1 - W_i)s_i^2$$

example completed

The variance of the poststratified mean income estimate is comes from this summary of the situation:

Sex	n	mean	var	sd	W_i	W_i*s^2_i	(1-W_i)*s^2_i
Female	550	26.561	290.620	17.048	0.504	146.473	144.148
Male	450	42.118	1112.308	33.351	0.496	551.705	560.603