Implementation in practice: sampling weights, variance approximations, and the design effect

We have learned about many important topics in designing sample surveys, such as stratification, clustering, and the use of auxillary information with ratio estimation. These ideas can be combined in many ways to obtain very complex multi-stage sampling designs, and our chapter on two-stage cluster sampling gave us a glimpse of the complexity that can be involved in more sophisticated designs. As we saw in the chapter on two-stage cluster sampling, the usual estimators (particularly variance estimators) can become very complicated. In actual sampling studies, researchers often simplify calculation of estimates by using sampling weights and computational approximations for variance calculations. Often the concept of the design effect is also used to simplify calculations for planning studies.

Use of sample weights: Here we will introduce sampling weights by rewriting the expression for the estimator of the mean in stratified random sampling. Recall in the chapter on stratified random sampling that the estimator of μ is:

$$\widehat{\mu} = \overline{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \overline{y}_i \text{ which can be written as}$$

$$= \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{n_i} \frac{N_i}{n_i} y_{ij} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{i=1}^{L}\sum_{j=1}^{n_i} w_{ij}},$$

where $w_{ij} = N_i/n_i$ is the weight for the j^{th} observation in group i, and has the interpretation that each observation in the sample represents $w_{ij} = N_i/n_i$ members of the population. Thus if a population of $N = 1000$ elements are divided into four strata each equal to $N_i = 250$, and if equal sample sizes of $n_i = 100$ are used for each stratum, then each observation in the sample represents $N_i/n_i = 2.5$ elements from the population. The general idea is that a sampling weight is a reciprocal of a selection probability, so for the StRS example above, $\delta_{ij} = 1/w_{ij} = n_i/N_i$ is the probability of being sampled for a member of the i^{th} stratum. For a multistage sampling design, the probabilities of selection are obtained my multiplying probabilities from

each stage. For example, suppose we have a two-stage cluster sample where n clusters are sampled from a population of N clusters, and then for the i^{th} sampled cluster, m_i elements are sampled out of a total of M_i elements. We generally call the clusters the primary sampling unit (psu) and the elements within clusters the secondary sampling unit (ssu). Then we can calculate the probability of the j^{th} ssu in the i^{th} psu being selected:

$$\begin{aligned}
&P(j^{th} \text{ ssu in } i^{th} \text{ psu being selected}) \\
&= P(i^{th} \text{ psu being selected}) P(j^{th} \text{ ssu is selected} | i^{th} \text{ psu is selected}) \\
&= (\frac{n}{N})(\frac{m_i}{M_i})
\end{aligned}$$

Now the weights are again the reciprocals of these probabilities. In many sampling studies, the sampling weights are calculated as the sampling design is developed. Once the sampling weights are calculated, any quantity of interest can be calculated as a weighted sum as exemplified in the StRS expression above. Also, the weights can be adjusted for nonresponse or other reasons as discussed in Lohr (1999).

Variance estimation in complex surveys: As we saw for two-stage cluster sampling, variance expressions become more complex as we add multiple stages of sampling. Stratification also adds to the complexity of variance estimation. Although we can in principle calculate complicated variance expressions for different sampling designs, in practice it is now common to use computationally intensive methods such as balanced repeated replication (BRR), jackknife, and bootstrap methods as well as some simplifying assumptions to calculate variance approximations. The SAS procedure Proc SURVEYMEANS can be used to calculate estimates and variance estimates for parameters for many sampling designs. It can use BRR, jackknife, or Taylor series methods to calculate an approximate variance estimate. When using multistage designs, it calculates approximate variances by only using the variation between primary sampling units. In the SAS examples of Proc SURVEYMEANS on the website, note how the variance estimate for the two-stage cluster example is slightly smaller than what we calculated, because only the variation between psu's is used.

Use of a design effect: Computing sample sizes for complex surveys that are repeated over time is made easier with the concept of a **design effect**

(denoted by deff). The design effect for a sampling plan and a statistic of interest is defined to be the ratio of the estimated variance of the statistic under the sampling plan to the estimated variance of the statistic under simple random sampling. As an example, consider estimating a proportion from a complex multistage design. The design effect for the complex design would be:

$$\begin{aligned}\text{deff}(\text{complex design}, \widehat{p}) &= \frac{\widehat{V}(\text{estimate from complex design})}{\widehat{V}(\text{SRS with same sample size})} \\ &= \frac{\widehat{V}(\text{estimate from complex design})}{\widehat{p}(1-\widehat{p})/n}.\end{aligned}$$

The design effect is similar to a relative efficiency, and measures the loss (or gain) in efficiency of the complex design relative to an SRS design. This is extremely useful when computing sample sizes for a future sample survey. For a future survey, the sample size estimate is just the estimate for a SRS sample for a given bound multiplied by the design effect. For example, suppose a multistage sampling plan that involved clustering and stratification was used to estimate a proportion, and the design effect was 1.7. Then for the next survey, the sample size for an SRS sample and the given bound can be calculated and multiplied by 1.7 to give a sample size for the complex sample design.

Reference: Lohr, S.L. 1999. *Sampling: Design and Analysis.* Pacific Grove, CA: Brooks/Cole.