Design Effects in Surveys that Require Oversampling of Certain Subpopulations

K.P. Srinath. Abt SRBI

1. Introduction

The design effect is the ratio of the variance of an estimated parameter from a survey using a complex sampling design to the variance of the same estimate under simple random sampling with the same total sample size. This is a useful tool to plan sample sizes for future surveys using complex sampling designs. It also provides a measure of the precision of the complex design. To compute the design effect it is required to compute the variance of the estimator under the design used for the survey and the variance of the estimator under simple random sampling.

In surveys which require oversampling of certain subpopulations, we may want to compute the design effect which takes into account stratification, clustering and unequal probability sampling but excluding oversampling since oversampling is one of the basic objectives of the survey. In this paper, we derive the variance of the estimate when there is oversampling but the selection of the initial sample and then the oversample is with equal probabilities.

2. Estimator of a Population Total

First, the variance of the estimator for estimating a population total under simple random sampling with the requirement of oversampling is derived**.**

Let N denote the population size. The population consists of a subgroup from which we want an oversample. Let the size of the subgroup in the population be N_1. For example if we are sampling households, we may be interested in selecting an oversample of Hispanic households. The sampling procedure is as follows. First we select a simple random sample of n households. In this sample, let the number of households not belonging to the subgroup be n_1 and the number of households belonging to the subgroup of interest be n_2 . We have $n_1 + n_2 = n.$ From the remaining population of $N - n$ households, we select a sample of n^* households and screen this sample and select a sample of n_2^* households belonging to the subpopulation. An estimator of the population total of some household characteristic of interest is

$$\hat{Y} = \frac{N}{n}(n_1 \bar{y}_1) + \frac{N}{n+n^*}(n_2 \bar{y}_2 + n_2^* \bar{y}_2^*). \qquad (1)$$

We have $E(\hat{Y}) = E_1 E_2(\hat{Y})$.

$$E_2(\hat{Y}) = \frac{N}{n} n_1 E_2(\bar{y}_1) + \frac{N}{n+n^*}[(n_2 E_2(\bar{y}_2) + n_2^* E_2(\bar{y}_2^*)].$$

$$E_2(\hat{Y}) = \frac{N}{n} n_1 \overline{Y}_1 + \frac{N}{n+n^*}[(n_2 \overline{Y}_2 + n_2^* \overline{Y}_2)].$$

$\overline{Y}_1$ is the population mean excluding the subgroup identified for oversampling. $\overline{Y}_2$ is the population mean of the subgroup identified for oversampling.

$$E_1 E_2(\hat{Y}) = \frac{N}{n} \overline{Y}_1 E_1(n_1) + \frac{N}{n+n^*} \overline{Y}_2 [E_1(n_2) + E_1(n_2^*)].$$

We have $E_1(n_1) = n \frac{N_1}{N}$, $E_1(n_2) = n \frac{N_2}{N}$ and $E_1(n_2^*) = E_1 E_2(n_2^* / n_2)$.

N_1 is the size of the population excluding the subgroup and N_2 is the size of subgroup population.

$E_2(n_2^* / n_2) = n^* \frac{N - n_2}{N - n}$. Therefore, $E_1(n_2^*) = n^* \frac{N - E_1(n_2)}{N - n} = n^* \frac{N_2}{N}$.

Therefore, $E(\hat{Y}) = \frac{N}{n} n \frac{N_1}{N} \overline{Y}_1 + \frac{N}{n+n^*}(n+n^*) \frac{N_2}{N}$. This gives

$$E(\hat{Y}) = N_1 \overline{Y}_1 + N_2 \overline{Y}_2 = Y.$$

$\hat{Y}$ is unbiased for Y.

Variance of $\hat{Y}$ is

$$V(\hat{Y}) = E_1 V_2(\hat{Y}) + V_1 E_2(\hat{Y})$$

For deriving $V_1 E_2(\hat{Y})$, we assume that n^* and n_2^* are fixed in repeated sampling. We replace n_2^* by its expected value which is $n^* \frac{N_2 - n_2}{N - n}$. This is the desired size of the oversample.

Under this assumption

$$V_1 E_2(\hat{Y}) = V_1[\frac{N}{n} n_1 \bar{Y}_1 + \frac{N}{n+n^*}(n_2 + n^* \frac{N_2 - n_2}{N-n})\bar{Y}_2].$$

$$V_1 E_2(\hat{Y}) = \frac{N^2}{n^2}\bar{Y}_1^2 V(n_1) + \frac{N^2}{(n+n^*)^2}\frac{(N-n-n^*)^2}{(N-n)^2}\bar{Y}_2^2 V(n_2)$$

$$+ 2(\frac{N}{n})(\frac{N}{n+n^*})(\frac{N-n-n^*}{N-n})\bar{Y}_1\bar{Y}_2 Cov(n_1, n_2).$$

Ignoring the finite population correction (fpc), we have $V(n_1) = V(n_2) = n\frac{N_1 N_2}{N^2}$.

Also,

$Cov(n_1, n_2) = -V(n_1)$. Therefore,

$$V_1 E_2(\hat{Y}) = N^2 n \frac{N_1 N_2}{N^2}[(\frac{\bar{Y}_1}{n} - \frac{\bar{Y}_2}{(n+n^*)}\frac{(N-n-n^*)}{(N-n)}]^2 \qquad (2)$$

$$E_1 V_2(\hat{Y}) = E_1[\frac{N^2}{n^2} n_1^2 V_2(\bar{y}_1) + \frac{N^2}{(n+n^*)^2}((n_2^2 V_2(\bar{y}_2) + n_2^{*2} V_2(\bar{y}_2^*))].$$

Ignoring fpc, $V_2(\bar{y}_1) = \frac{S_1^2}{n_1}$, $V_2(\bar{y}_2) = \frac{S_2^2}{n_2}$ and $V_2(\bar{y}_2^*) = \frac{S_2^2}{n_2^*}$. Now we have

$$E_1 V_2(\hat{Y}) = N^2 E_1[\frac{1}{n^2} n_1 S_1^2 + \frac{1}{(n+n^*)^2}((n_2 + n_2^*)S_2^2)].$$

$$E_1 V_2(\hat{Y}) = N^2[\frac{1}{n^2} E_1(n_1) S_1^2 + \frac{1}{(n+n^*)^2}(E_1(n_2) + E_1(n_2^*))\, S_2^2$$

$$E_1 V_2(\hat{Y}) = N^2[\frac{1}{n}\frac{N_1}{N} S_1^2 + \frac{1}{(n+n^*)}\frac{N_2}{N} S_2^2] \ . \qquad (3)$$

$$V(\hat{Y}) = \text{(2) +(3)}$$

$$V(\hat{Y}) = N^2[\frac{1}{n}\frac{N_1}{N} S_1^2 + \frac{1}{(n+n^*)}\frac{N_2}{N} S_2^2] + N^2 n \frac{N_1 N_2}{N^2}[(\frac{\bar{Y}_1}{n} - \frac{\bar{Y}_2}{(n+n^*)}\frac{(N-n-n^*)}{(N-n)}]^2$$

Let $\alpha = \frac{N_2}{N}$ be the proportion of the subpopulation (population identified for oversampling) in the total population. is $(1-\alpha) = \frac{N_1}{N}$ the proportion of rest of the population. If we assume that $(N-n)$ is large relative to n^*, then we can write the variance as

$$V(\hat{Y}) = \quad N^2[\frac{1}{n}(1-\alpha)S_1^2 + \frac{1}{(n+n^*)}\alpha S_2^2] + N^2 n\alpha(1-\alpha)[\frac{\overline{Y}_1}{n} - \frac{\overline{Y}_2}{(n+n^*)}]^2$$

3. Variance of an Estimate of Population Proportion

If we are estimating a population proportion $P = (1-\alpha)P_1 + \alpha P_2$ where P_1 and P_2 are population proportions of some characteristic of interest in the domains and the sample proportion defined as

$$\hat{P} = \frac{\hat{Y}}{N}$$

then, the variance of $\hat{P}$ is

$$V(\hat{P}) = \frac{(1-\alpha)P_1(1-P_1)}{n} + \frac{\alpha P_2(1-P_2)}{n+n^*} + n\alpha(1-\alpha)[\frac{P_1}{n} - \frac{P_2}{(n+n^*)}]^2 .$$

Another expression for the variance is

$$V(\hat{P}) = \frac{(1-\alpha)P_1(1-P_1)}{n} + \frac{\alpha P_2(1-P_2)}{n+n^*} + \frac{\alpha(1-\alpha)}{n}[P_1 - \frac{P_2 n}{(n+n^*)^2}] \qquad (4)$$

For computing the variance of the estimate under simple random sampling we will assume that the total sample size is $(n+n_2^*)$

If we select a simple random sample of $(n+n_2^*)$ units and compute the sample proportion as

$$p = \frac{y}{(n+n_2^*)}$$

where y is the sample total, then $V(p) = \frac{P(1-P)}{n+n_2^*}$.

This variance can be written in terms of the proportions P_1 and P_2 is

$$V(\hat{P}) = \frac{(1-\alpha)P_1(1-P_1)}{n+n_2^*} + \frac{\alpha P_2(1-P_2)}{n+n_2^*} + \frac{\alpha(1-\alpha)}{n+n_2^*}(P_1-P_2)^2 \quad (5).$$

We can compare (4) and (5) to assess the impact of oversampling on the variance of the overall estimate.

4. Ratio Estimator

In this section we show the variance of the ratio estimator used in estimating the population proportion.

We are estimating he population proportion using

$$\hat{R} = \frac{\hat{Y}}{\hat{N}}.$$

$$\hat{R} = \frac{\hat{Y}}{\hat{N}} = \frac{\frac{N}{n}(n_1 p_1) + \frac{N}{(n+n^*)}(n_2 p_2 + n_2^* p_2^*)}{\frac{N}{n}(n_1) + \frac{N}{(n+n^*)}(n_2 + n_2^*)} \quad .$$

It can be shown that the variance of $\hat{R}$ is

$$V(\hat{R}) = \frac{(1-\alpha)P_1(1-P_1)}{n+n^*} + \frac{\alpha P_2(1-P_2)}{n+n^*} + n\alpha(1-\alpha)[(\frac{P_1}{n} - \frac{P_2}{n+n^*}) - (\frac{P}{n})(\frac{n^*}{n+n^*})]^2.$$

5. Example

Assume that the subgroup from which we want an oversample is 20% of the total population. This makes $\alpha = 0.20$. Let the initial sample size be 100. We expect a sample of 20 in the subgroup and 80 will be from the rest of the population. Assume that we want an oversample of 10 additional units from the subgroup identified for oversampling. This means we need to select an additional sample of 50 sampling

units from the rest of the population. That is $n = 100$ and $n^* = 50$. We assume two proportions for the main group just to illustrate the increase in variance because of oversampling as compared to no oversampling. For the purpose of computing the variance under SRS assuming no oversampling we will use the total sample size as 100+10 =110. Table 1 gives the variance under oversampling using the estimator given above and also the variance under simple random sampling for various values of the population proportion in the subgroup. The design effects which reflect oversampling are also shown.

Table 1: Variances and Design Effects

Main Group	Subgroup	Overall	Oversample	SRS 110	Design Effect
0.5	0.75	0.55	0.002250	0.002250	1.00
	0.50	0.50	0.002378	0.002273	1.05
	0.40	0.48	0.002407	0.002269	1.06
	0.30	0.46	0.002424	0.002258	1.07
	0.20	0.44	0.002428	0.002244	1.08
	0.10	0.42	0.002420	0.002215	1.09
0.3	0.9	0.42	0.001944	0.002215	0.88
	0.8	0.40	0.001980	0.002182	0.91
	0.7	0.38	0.002004	0.002142	0.94
	0.6	0.36	0.002016	0.002096	0.96
	0.5	0.34	0.002015	0.002040	0.99

6. Conclusion

If in a complex sampling design in which there is oversampling of units from a subpopulation, we want assess the impact of stratification, clustering and unequal weights but not oversampling, then the variance shown here under simple random sampling but with oversampling can be used. The design effect reflects the increase in variance but without taking oversampling into account. As seen from this table, the design effects because of oversampling are small and sometimes less than one.

References:

Park I. and Lee. H. (2004). "Design Effects for the Weighted Mean and Total Estimation under Complex Survey Sampling." Survey Methodology, Vol.32, No.2, pp. 183-193

Kish, L., "Methods for Design Effects" Journal of Official Statistics 11: 55-77.