Decomposing Design Effects for Stratified Sampling

Jun Liu, Vince Iannacchione, and Margie Byron
R TI International, Research Triangle Park, NC, 27709

Key Words: design effect, unequal weighting effect, clustering effect, stratification, optimal sample allocation

1. INTRODUCTION

Design effect is very often used in the design and planning of complex surveys. Very often, complex designs involve stratification and clustering, among other things. When conducting study design, the following fsituations are often encountered.

1. Need to control design effects for certain strata due to analytical reasons
2. Sample designs could be different from stratum to stratum
3. Need to estimate the overall design effect using individual stratum design effects

However, the current formula of the design effect, popularized by Kish (1965), is only available for the overall survey and can not take the advantage of available information at stratum level. In Liu, et al (2000), the unequal weighting effect is decomposed at stratum level and used for an optimal sample allocation of subsamples in a follow up wave of a longitudinal study.

In this paper, we discuss decomposing of the design effect itself into stratum level components. An hypothetical example of establishment survey, where the proposed formula is applied, is provided for illustration purposes.

2. A DECOMPOSITION FORMULA

According to Kish (1965), the design effect of a study is the ratio of the design-based variable and the variance under the simple random sampling. The design effect can be written as the multiple of two components - the component that is attributable to the variation in weights (unequal weighting effect) and the component that is attributable to the clustering of data (clustering effect). That is,

$$Deff = \frac{Var_{Design}(\bar{y})}{Var_{SRS}(\bar{y})} = UWE \times Deff_C .$$

The unequal weighting effect and the design effect attributable to the clustering are defined as (assuming the overall population variance σ^2 is the same as the stratum population variances, σ_h^2)

$$UWE = \frac{n\sum_h \sum_{ij} w_{hij}^2}{\left(\sum_h \sum_{ij} w_{hij}\right)^2}$$

and

$$Deff_C = 1 + \rho\left(\overline{m} - 1\right)$$

where ρ is the intra-clustering coefficient and $\overline{m}$ is the average cluster size.

For a stratified sample design, Liu et al (2000) derived the following decomposition for the Unequal Weighting Effect (UWE),

$$\begin{aligned} UWE &= \frac{n\sum_h \sum_{ij} w_{hij}^2}{\left(\sum_h \sum_{ij} w_{hij}\right)^2} \\ &= \sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{n}{n_h}\right) UWE_h \end{aligned}$$

where,

$$UWE_h = \frac{n_h \sum_{ij} w_{hij}^2}{\left(\sum_{ij} w_{hij}\right)^2}$$

is the stratum specific unequal weighting effect, N_h and n_h are the stratum population size and sample sizes. This formula is used to guide the subsampling in a longitudinal follow-up study where the overall unequal weighting effect needs to be balanced with that of subdomains.

The nature question, therefore, is whether a similar decomposition holds for the overall design effect.

Consider a class of linear estimators for the population mean, defined as

$$\begin{aligned} \bar{y} &= \frac{1}{N}\sum_h \sum_{ij}\left(w_{hij}\, y_{hij}\right) \\ &= \sum_h \left(\frac{N_h}{N}\right) \bar{y}_h \end{aligned}$$

where,

$$\bar{y}_h = \frac{1}{N_h}\sum_{ij}\left(w_{hij}\, y_{hij}\right)$$

is the stratum specific mean estimator.

It can be shown that the variance of the estimator can be expressed in the following two forms:

$$Var_{SRS}\left(\bar{y}\right) = \frac{\sigma^2}{n}$$

and

$$Var_{SRS}\left(\bar{y}\right) = \sum_h \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^{\ 2}}{n_h}$$

Furthermore, we assume that the stratum specific population variances σ_h^2's equal the overall population variance σ^2 .

Using these two results, with some simple algebra, one can easily show that

$$\begin{aligned} Deff &= \frac{Var_{Design}\left(\bar{y}\right)}{Var_{SRS}\left(\bar{y}\right)} \\ &= \sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{n}{n_h}\right) \frac{Var_{Design}\left(\bar{y}_h\right)}{Var_{SRS}\left(\bar{y}_h\right)} \\ &= \sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{n}{n_h}\right) Deff_h \end{aligned}$$

That is, the overall design effect Deff can be construed as a weighted sum of stratum specific design effects of $Deff_h$'s.

The stratum-specific *Deff* can be further decomposed to unequal weighting effect (UWE) and clustering effect ($Deff_c$). When there is only unequal weighting effect, it becomes the formula derived in Liu, el at (2000). When there is only clustering effect, the above becomes

$$Deff = \sum_h \left(\frac{N_h}{N}\right)^2 \left(\frac{n}{n_h}\right) \left\{1 + \rho_h\left(\overline{m}_h - 1\right)\right\}$$

The weights associated with the weighted sum are of interest for further analysis. The first component of the weights is the stratum population as a percentage of the total population squared. The second component is the inverse of the sample fraction of the stratum.

It suggested that the value of the stratum specific design effect can be mediated by the population size of the stratum relative to that of the overall population and also can be influenced by the particular sample allocation. This point will be discussed further in the next session where an illustrative example will be provided to show how the formula can be used in practice.

3. EXAMPLES

To illustrate the utility of the above formula, we have created a hypothetical example of an establishment survey of work place attitude towards people with HIV. It is reasonable to assume that employees of small establishments would tend to have similar views while employees of large establishments will have more diversified views. It therefore is important to recognize that the intra-clustering effect can vary in a wide range, depending on the size and type of establishments. To make the example simple, we assume that there will be no unequal weighting effect and the stratum specific and overall design effects will all be due to the clustering effect.

To account for this feature of the population, the frame of the establishment is divided into three categories by the number of employees: small, medium and large. The following table provides the first stage sampling frame information.

First Stage Sampling Frame and Stratification

	No of employees	No of business	Total no of employees	Rho
Large	2,000+	700	350,000	.001
Medium	200 ~ 2,000	4,000	4,000,000	.008
Small	10 ~ 199	25,000	10,000,000	.02

We further required that, for analytical reasons, the maximum design effect for domains formed by business type (size stratum) is less than 1.7 and the overall design effect is less than 1.5. The following sample allocation would provide a design that satisfies the requirements on stratum-specific design effects.

Stratum Specific Design Effect

	Average cluster size	Number of clusters	Domain design effect
large	20	40	1.46
med	3	300	1.56
small	2	660	1.70

The overall design effect for the above design, can be calculated using the formula in the above section, is 1.4, well below the requirement.

A formal optimization can be very easily set up using the design effect decomposition formula. An example of such application can be found in Liu, et al (2000), using the decomposition formula when there is only unequal weighting effect.

Another hypothetical example where the formula above can be very useful is given below:

A national household survey on health conditions by race/ethnicity.

- Sampling design: Two-stage cluster sample with geographic areas (e.g., Census Tracts or Zip Codes) as the 1st stage and households selected at the 2nd stage.
- Sample sizes: Hispanics and NH Blacks need to be over-sampled so that comparisons can be made with adequate power.
- Sampling frame: Geographic clusters are stratified by minority concentration.
- Intra-cluster correlations: As the minority concentration of a geographic cluster increases, the "similarity" of the health conditions also increases.
- Sample allocation: Assuming fixed costs, the decomposed Deff will be used to allocate the sample of households to each stratum in a way that achieves the target sample sizes with minimum design effects for each R/E domain.

Again, the domain specific design effects can be controlled through allocation of the second stage sample sizes. The overall design effect can then be estimated using the formula similar to the above example.

4. CONCLUSIONS

To summarize, we have extended the decomposition formula for the unequal weighting effect to the overall design effect for a linear estimator of the mean. We discussed the factors that influences the over design effect. We illustrated the possible application of the formula in study designs trough examples.

5. REFERENCES

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons

Liu, J. and Aragon, E. (2000). Subsampling strategies in longitudinal surveys. *Proceedings of the American Statistical Association, Section of Survey Research Methods, pp.307-312.*