A Bound for the Relative Bias of the Design Effect

Alberto Padilla
Banco de México

Abstract

Design effects are typically used to compute sample sizes or standard errors from complex surveys. In this paper, we show that the design effect estimator is biased and an upper bound for the relative bias is presented. A simulation study was conducted to assess the size of the bound for the relative bias with samples drawn from two artificially generated populations using stratified and two-stage random sampling.

Key Words: Variance of variances, Confidence interval, Coefficient of variation

1. Introduction

The design effect, *deff*, Kish (1965), is defined as the ratio of the variance of an estimator under an specific design to the variance of the estimator under simple random sampling without replacement, *srswor*. The estimator of the design effect is used for example in the computation of the sample size for complex sample designs and to build confidence intervals. In this article we will exhibit the bias of the design effect and a bound for the relative bias of the standard error of the design effect.

2. Design Effect Estimation and Bias

2.1 Definition

It is worth mentioning that all results are based on the design based approach for the planning stage of a survey. The design effect, *deff*, Kish (1965), is defined as the ratio of the variance of an estimator under an specific design different from simple random sampling, $v_{alt}(\hat{y}_{alt})$, to the variance of the estimator under simple random sampling without replacement, *srswor* :

$$deff\,(\hat{y}) = v_{alt}(\hat{y}_{alt}) / v_{srswor}(\hat{y}_{srswor})$$

The design effect estimator, *deff*, Kish (1965), is computed by plugging in the formula showed in the previous slide, estimators of the variances in both the numerator and denominator. In particular, the variance estimator under *srswor* is generally obtained by using formula:

$$\hat{v}_{srswor}(\hat{y}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-1}$$

In this formula the sample is considered as a *srswor* and it does not guarantee an unbiased estimation of the population variance under *srswor*. An example with a small population ill shed light on this point.

Remark: we will use the term relative bias to the upper bound of the ratio of the bias to the standard error.

2.1.1 Example 1. Design Effect Estimation and Bias

Table 1: Small Stratified Population

	Values	*Population mean*	
Stratum	y_{hi}	y_h	s^2_{hU}
1	{2, 3, 4.5, 2.5, 3.4}	3.08	0.91
2	{11, 14, 18}	14.33	12.33
Population		7.30	37.96

A simple random sample without replacement, *srs*, of size 2 was extracted from each statum. There are 30 possible samples under stratified random sampling, *strs*, and 70 under *srs*. The population variances for the mean under both sampling designs are, $v_{srswor}(\hat{y}) = 4.75$ and $v_{srs}(\hat{y}_{st}) = 0.395$ respectively. The design effect for this population is:

$$deff\,(\hat{y}) = v_{srs}(\hat{y}_{st}) / v_{srswor}(\hat{y}) = 0.083$$

As it was mentioned above, Kish (1965) defined the estimator of *deff* for each sample as: $\hat{deff}_K(\hat{y}_i) = \hat{v}_{strs}(\hat{y}_{st,i}) / \hat{v}_{biased,srswor}(\hat{y}_i)$, where i=1,…30, and

$$\hat{v}_{biased,srs}(\hat{y}_i) = (1-4/8)\sum_{j=1}^{4}(y_{hj}-\hat{y}_i)^2 / (4-1)4$$

In this case, $\sum_{i=1}^{30}\hat{v}_{biased,srs}(\hat{y}_i)/30 = 5.93$, which is different from $v_{srs}(\hat{y}) = 4.75$

The average over all possible samples under stratified random sampling is,

$$\sum_{i=1}^{30}\hat{deff}_K(\hat{y}_i)/30 = 0.067 \neq deff\,(\hat{y}) = 0.083$$

This result is not surprising since the estimator $\hat{deff}_K$ is a ratio estimator which is known to be biased, Cochran (1977).

This result shows the need to modify the estimator used in the denominator of $\hat{deff}_K$ with the formula proposed by Gambino (2009) to obtain an unbiased estimator:

$$\hat{v}_{unbiased,\ srs}(\hat{y}) = (1-\frac{n}{N})\frac{1}{n\,(N-1)}(\hat{y}_{sq} - \frac{\hat{y}^2 - \hat{v}_{alt}(\hat{y})}{N})$$

Where $\hat{y}_{sq,}$ $\hat{y}^2$ and $\hat{v}_{alt}(\hat{y})$ are unbiased estimators of the following population quantities: sum of squares, total squared and variance of the total under the design different from *srswor*. Hereinafter, $\hat{deff}_G$ will denote the *deff* estimator using Gambino´s correction.

Using Gambino´s correction for the estimation of $\hat{deff}_G$ we obtain:

$$\sum_{i=1}^{30}\hat{deff}_G(\hat{y}_i)/30 = 0.084 \neq deff\,(\hat{y}) = 0.083$$

This estimator remains biased, but the source of the bias stems from the use of a ratio estimator only.

2.1.2 Bounds for the Bias and Relative Bias of the Deff

Theorem: for a sample design, *alt*, different from *srs*, with variance estimators, $ms\hat{e}_{alt}$, $ms\hat{e}_{srs}$ and $ms\hat{e}_{srs} > 0$ we have:

$$E(\hat{deff}) = E\left(\frac{ms\hat{e}_{alt}}{ms\hat{e}_{srs}}\right) = \frac{v_{alt} + bias_{alt}}{v_{srs} + bias_{srs}} - \frac{\text{cov}(ms\hat{e}_{srs}, \hat{deff})}{v_{srs} + bias_{srs}}$$

Corollary 1: under a sample design *alt* different from *srs*, with unbiased estimators of the population variances of *alt* and *srs*, and using $\hat{deff}_G$, the relative bias is given by:

$$-\text{cov}(\hat{deff}_G, \hat{v}_{unbiased\,srs}) / v_{srs}$$

Remark: in this expression, $\hat{v}_{unbiased\,srs}$ is computed with Gambino´s formula (2009).

Corollary 2: a bound for the relative bias of $\hat{deff}_G$ is given by $cv(\hat{v}_{unbiased\,srs})$, i.e., the coefficient of variation of the unbiased estimators of the population variance under srs.

Remark: we are working with expression $\hat{deff}_G$, which is different from $\hat{deff}_K$. The latter expression is the quantity routinely employed in practice.

2.1.3 Example 2. Design Effect Estimation and Bias stratified random sampling

Based on Cochran (1977) example, page 137, we simulated a small population with 3 strata and 57 elements.

Table 2: Simulated Stratified Population

Strata	N_h	n_h	W_h	y_h	s^2_{hU}
1	13	9	0.22	2.33	1.62
2	18	7	0.32	1.61	0.08
3	26	6	0.46	5.04	1.18
Population	57	22			3.44

Population quantity	Value
v_{srs}	0.096
v_{strs}	0.035
$deff_K$	0.364

The bound for the relative bias, $cv(\hat{v}_{unbiased\,srs})$, computed with $\hat{v}_{unbiased\ srs}$ from the simulation, was 18.4%.

Remark: in table 2, the labels for columns 2 to 6 refer to population size, sample size, relative size, stratum mean and element variance within strata.

From the population defined in the previous slide, we simulate the extraction of 5,000 samples of size 22 under *strs*, and for each sample we computed the following estimators:

- Unbiased estimator of the variance under *strs*, $\hat{v}_{strs}$,
- Biased estimator of the variance under *srs*, using Kish definition, $\hat{v}_{biased\ srs}$,
- Unbiased estimator of the variance under *srs* using Gambino correction, $\hat{v}_{unbiased\,srs}$
- Deff estimator using Kish formula, $\hat{deff}_K$,
- Deff estimator using Gambino correction, $\hat{deff}_G$

Results for 5,000 samples for each estimator:

Estimator	Value	Bias (%)
$\hat{v}_{unbiased\,srs}$	0.0964	---
$\hat{v}_{biased\ srs}$	0.0832	-13.4%
$\hat{v}_{strs}$	0.0352	---
$\hat{deff}_K$	0.4318	18.7%
$\hat{deff}_G$	0.3724	2.4%

The bound for the relative bias, $cv(\hat{v}_{unbiased\,srs})$, computed with $\hat{v}_{unbiased\ srs}$ from the simulation, was 18.4%.

2.1.4 Example 3. Design Effect Estimation and Bias two-stage cluster sampling

We have a small population with 8 clusters or primary sampling units, PSU, and each PSU has 8 elements, *SSU*, secondary sampling units. At first stage we draw *a=3 PSU* and *b=4 SSU*, so the sample size is *n=ab=12* elements. The values within each cluster were simulated using uniform random variables. The minimum and maximum employed in the simulation are shown in the next table.

PSU	y_i	s_i^2	min & max
1	0.332	0.0058	0.2 and 0.5
2	0.444	0.0009	0.4 and 0.5
3	1.237	0.0094	1.1 and 1.4
4	0.919	0.0037	0.8 and 1.0
5	0.223	0.0064	0.1 and 0.35
6	0.610	0.0030	0.5 and 0.7
7	0.970	0.0044	0.9 and 1.1
8	0.461	0.0077	0.3 and 0.6
Population	0.650	0.1166	

The values in columns 2 and 3 refer to the within-cluster mean and variance.

Population quantity	Value
v_{srs}	0.0079
v_{clus}	0.0212
$deff_K$	2.6877

In this table v_{clus} is the population variance under two-stage random sampling. The intraclass correlation for this population is 0.95 and was computed using the result from Cochran (1977) page 291. We simulate the extraction of 3,500 samples of size 12, with a=3 PSU selected by *srs* and b=4 SSU selected by srs.

For each sample we computed the following estimators:

- Unbiased estimator of the variance under *clus*, $\hat{v}_{clus}$,
- Biased estimator of the variance under *srs*, using Kish definition, $\hat{v}_{biased\ srs}$,
- Unbiased estimator of the variance under *srs* using Gambino correction, $\hat{v}_{unbiased\ srs}$
- Deff estimator using Kish formula, $\hat{deff}_K$,
- Deff estimator using Gambino correction, $\hat{deff}_G$

The results for the 5,000 samples for each estimator are shown above:

Estimator	Average of estimators	Relative bias
$\hat{v}_{unbiased\ srs}$	0.0080	---
$\hat{v}_{biased\ srs}$	0.0066	-16.13%
$\hat{v}_{clus}$	0.0269	---
$\hat{deff}_K$	3.9037	45.24%
$\hat{deff}_G$	3.2467	20.80%

The bound for the relative bias, $cv(\hat{v}_{unbiased\ srs})$, computed with $\hat{v}_{unbiased\ srs}$ from the simulation, was 63.44%.

2.1.5 Example 4. Design Effect Estimation and Bias two-stage cluster sampling with several roh values

Using the same population of example 3 and changing elements between clusters, we repeated the simulations as in example 3 in order to obtain different values of the bound for the relative bias and to compute the formula used in practice deff≈[1+roh(b-1)] and compare it to the population deff. With this change between clusters the population mean and variance between elements was unaffected, but roh and deff changed.

roh	deff	Bound Relative bias	deff≈[1+roh(b-1)]
-0.14	0.70	0.25	0.58
-0.05	0.92	0.34	0.85
0.01	1.07	0.34	1.03
0.14	1.38	0.36	1.42
0.26	1.66	0.37	1.78
0.38	1.97	0.43	2.15
0.50	2.25	0.45	2.50
0.63	2.50	0.54	2.88
0.76	2.87	0.61	3.27
0.86	3.12	0.62	3.57
0.96	3.35	0.64	3.87

In the table, [1+roh(b-1)] is a good aproximation to deff (population value) whenever (A-1)/A and (N-1)/N are equal to unity, see Kish (1965), chapter 5. From this table, it can be seen that [1+roh(b-1)] overestimates the population deff for roh values from 0.26 to 0.96 and when roh is negative it underestimates it.

3. Conclusions

An exact expression for the bias and an upper bound to the ratio of the bias of the design effect estimator to the standard error was given. The upper bound for the bias is given by the coefficient of variation of the unbiased estimators of the variance under simple random sampling. Based on the simulations and the extensive use of the design effect in practice, it is advisable to analyse the stability of the variance estimator under simple random sampling, whenever possible. It is also advisable to work with an unbiased estimator of the variance of simple random sampling. Some more simulations are needed to assess the usefulness of formula deff≈[1+roh(b-1)], it seems that it tends to over or underestimate the true deff value.

References

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley.
Gambino, J.G. (2009). *Design effects caveat*, The American Statistician, pp. 141-145.
Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Padilla, A.M., *Una cota para el sesgo relativo del efecto del diseño*, Memorias electrónicas en extenso de la 4ª Semana Internacional de la Estadística y la Probabilidad. Julio 2011, CD ISBN: 978-607-487-324-5.
Rao, J.N.K. (1962). *On the estimation of the relative efficiency of sampling procedures*, *Annals* of the *Institute* of *Statistical* Mathematics, pp. 143-150.