International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Clustering Component of Design Effect

Nidal Mohamed Mustafa Abd Elsalam*

Department of Statistics, Faculty of Science, Tabuk University, Kingdom of Saudi Arabia
Email: nidalmm2@gmil.com

Abstract

This work is concerned with expressions for the clustering component of the design effect in terms of parameters that are expected to affect the design effect in cluster sampling, for equal cluster sizes and unequal probability sampling (PPS). This involved investigating the pattern of design effect when such parameters change using factorial combination method. It is shown that the expressions for the design effect helped to reveal the impact on the design effect of clusters means, clusters variances, clusters homogeneity, sample stages, and sample sizes. It is also shown that the pattern of design effect changes significantly with change in those parameters.

Keywords: Cluster Sampling; Population Design Effect (Deff); intra-cluster correlation coefficient; factorial combination.

1. Introduction

The design effect continues to be a valuable tool to guide a sampler both in designing a survey and analyzing the results, as noted by the authors in [4] and [2]. The design effect is defined as the ratio of the variance of an estimate for a given survey design to the variance of the estimate for a simple random sample of the same size, as noted by the author in [5].

---

* Corresponding author.
E-mail address: nidalmm2@gmil.com.

In practice, the design effect must be handled with great care as it usually represents the combined effect of a number of components, of which clustering is one, which are often interdependent, as stated by author in [7].

This paper extends the previous work of the authors in [4] and in [6], on the design effect in cluster sampling in which the clustering component of the design effect was expressed only in terms of the intra-cluster correlation coefficient (ρ) and sub-sample size (b) within the primary selected units (PSU'S) in the formula:

$D^2(\bar{y}_{cl}) = 1 + (b-1)\rho$

Where $D^2(\bar{y}_{cl})$ refers to the population design effect of the mean in cluster sampling. While other parameters such as clusters means, clusters variances, sample stages, sample sizes that might affect the pattern of the design effect were not considered before.

The aims of the paper are to investigate the clustering component of the design effect and to examine how the design effect changes with varying population and sampling characteristics. To achieve these ends, expressions for the design effect in terms of parameters of interest are derived. Parameters affecting the pattern of the design effect are combined using factorial combination method, making it possible to give the expected magnitude of the design effect for a survey designed with a certain structure.

The study shows that clustering has a significant impact on the overall design effect; causing it to increase. In addition, the parameters of interest significantly affect the way the design effect varies. The biggest effect is shown in case of clusters variances, represented in within cluster's variances and between cluster's variances, compared to the effects of other parameters.

2. Materials and Methods

Starting with Expressions for the Design Effect in Cluster Sampling and concentrating on the mean as our estimate, the design effect is usually written in the form:

$$Deff = \frac{V(\bar{\bar{y}})}{V(\bar{y})} \qquad (1)$$

Where, the numerator stands for the variance of the mean in cluster sampling and the denominator represents the variance in case of simple random sample. Considering the case of equal cluster sizes, the expressions for the design effect are derived for each of the single-stage, two-stage, and three-stage, respectively as follows:

2.1 Single-stage sampling

$$Deff = M\left(\frac{1}{1+\frac{S_w^2}{S_b^2}}\right) \qquad (2)$$

Formula (2) is equivalent to the familiar formula found in standard sampling texts, but it has the property of expressing Deff in terms of within cluster variance (S_w^2) and between cluster variance (S_b^2). Hence, it enables us to see directly the effect of changes in these quantities.

Let:

$$A = \left(\frac{1}{1 + \frac{S_w^2}{S_b^2}} \right) \qquad (3)$$

Leading to:

$$Deff = AM \qquad (4)$$

Since (3) differs from (4) only in the latter being multiplied by a constant, it suffices for purposes of examining the change in design effect when S_w^2 & S_b^2 change, to concentrate on A. Results for different cluster sizes can then be easily obtained by multiplying by the required M.

2.2 Two-stage sampling

There are N first-stage units (FSU) or clusters, each of size M. A SRS of size n FSU's is selected. From each selected FSU, a SRS of m second-stage units (SSU's) is selected. Whereas S_1^2 represents variance of FSU's means, and S_2^2 is the variance of SSU's within FSU's. The design effect in case of two-stage sampling can now be calculated using various combinations of n, m, N, M, S_1^2 and S_2^2 through formulae number (5) below:

$$Deff = \frac{V_{cl}(\bar{\bar{y}})}{V(\bar{\bar{y}})} = \frac{(nm)\left[\left(\frac{N-n}{N} \right) \frac{S_1^2}{n} + \left(\frac{M-m}{M} \right) \frac{S_2^2}{nm} \right]}{\left(\frac{NM-nm}{NM} \right) \left[\frac{N(M-1)S_2^2 + M(N-1)S_1^2}{(NM-1)} \right]} \qquad (5)$$

2.3 Three-stage sampling

Here we assume that the population contains N FSU's, each with M SSU's, each of which has R third-stage units (TSU's). A SRS of n FSU's is selected. From each selected FSU, a SRS of m SSU's is selected, and from each selected SSU, a SRS of r TSU's is selected. S_3^2 is the variance among TSU's within SSU's within FSU's. The population design effect in three-stage sampling can be easily now derived by substituting for the values of n, m, r, N, M, R, S_1^2 , S_2^2 and S_3^2 as it is shown in formulae (6)

$$Deff = \frac{V(\bar{\bar{y}})}{V(\bar{y})} = \frac{\left[(1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_2^2}{nm} + (1-f_3)\frac{S_3^2}{nmr}\right]}{\frac{1}{nmr}\left(\frac{NMR-nmr}{NMR}\right)\left[\frac{NM(R-1)S_3^2 + RN(M-1)S_2^2 + MR(N-1)S_1^2}{(NMR-1)}\right]} \quad (6)$$

2.4 Single-stage sampling (PPS)

Regarding the case of probabilities proportional to size (PPS), the derived expressions for the design effect in case of single-stage and two-stage sampling are respectively as follows:

$$Deff* = \frac{(M_0-1)nM*[\sum_i^N M_i(\bar{Y}_i - \bar{\bar{Y}})^2]}{nM_0(1-\frac{nM*}{M_0})[\sum_i^N (M_i-1)S_i^2 + \sum_i^N M_i(\bar{Y}_i - \bar{\bar{Y}})^2]} \quad (7)$$

Where: M* = M_0/N is the average cluster size

The derived variance of the mean of SRS of nM* elements from a population of M_0 elements is;

$$V(\bar{y}) = \frac{(1-f)}{nM*(M_0-1)}[\sum_i^N (M_i-1)S_i^2 + \sum_i^N M_i(\bar{Y}_i - \bar{\bar{Y}})^2]$$

with

$$f = \frac{nM*}{M_0}$$

Now, selecting the ith unit with probability [Pi = Mi/M_0], and assuming sampling is with replacement:

$$\bar{\bar{y}}_{pps} = \frac{\sum_I^n \bar{Y}_i}{n}$$

Is an unbiased estimate of $\bar{\bar{Y}}$, with variance;

$$V(\bar{\bar{y}}_{pps}) = \frac{1}{nM_0}\sum_i^n M_i(\bar{Y}_i - \bar{\bar{Y}})^2$$

This depends on the variability of the $\overline{Y_i}$ and where;

$$\overline{\overline{Y}} = \frac{\sum_{i}^{N} M_i \overline{Y}_i}{M_0}$$

$$\overline{Y} = \frac{\sum_{i}^{N} M_i \overline{Y}_i}{N} = \frac{Y}{N}$$

Since sample size is a random variable as stated above, we get an approximation of Deff in formulae (7), that differs slightly from that of author [5], by comparing $V(\overline{\overline{y}}_{pps})$ to the variance of a SRS of a size equal to the expected sample size. Where the star on the design effect shows that this is not exactly kish's design effect.

2.5 Two-stage sampling (PPS)

In this section, we assume N clusters (FSU's) the ith of size M_i. A sample of size n FSU's is selected. From each selected FSU, SSU's are selected. There are three cases here as follows:

CASE 1

In this case, a sample of FSU's is selected with probability proportional to size. From each selected FSU, an equal probability of selection method (epsem) is employed to select a sample of b second-stage units (SSU's). The overall sample can be shown to be an (epsem) sample and the design effect is of the following form as noted by the author in [4]:

$$Deff = 1 + (b-1)\rho \qquad (8)$$

Where; ρ represents the intra-cluster correlation coefficient. Thus equation (8) is a useful model for the design effect from clustering for a variety of (epsem) sample designs, with respect to modification of interpretation of ρ.

CASE 2

In this case, probability proportional to estimated size (PPES) and sub sampling rates are applied in the sampled primary sampling units (PSU's), to give overall (epsem) design. This usually results in variable sub sample sizes. Assuming that the variation in sub sample sizes is not large, equation (8) can be used as an approximation with b being replaced by average sub sample size as follows:

$$Deff = 1 + (\overline{b} - 1)\rho \qquad (9)$$

Where; $\bar{b}$ is the average sub sample size

CASE 3

When the variation in sub sample sizes per PSU is substantial, the approximation of (9) becomes inadequate. The author in [6] extended the above approximation to deal with unequal sample sizes by replacing $\bar{b}$ by a weighted average sub sample size. Thus, the design effect due to clustering with unequal cluster sizes can be written as:

$$Deff = 1 + (b' - 1)\rho \qquad (10)$$

Where; $b' = \dfrac{\sum_i b_i^2}{\sum_i b_i}$, a weighted average sub sample size

3. Results

Previously, the clustering component of the design effect was expressed only in terms of the intra-cluster correlation coefficient (ρ) and sub-sample size (b) within the primary selected units (PSU'S), depending on Kish's design effect, as stated by author [5]. In this study, attention was paid to the expressions for the design effect in cluster sampling, in terms of clusters means, clusters variances, clusters sizes, together with clusters homogeneity, and sample sizes. In addition, the pattern of the design effect under varying circumstances was also examined. The obtained results are summarized and discussed as follows:

- The study showed that the clustering component has a major impact on the design effect and always causes it to increase. The parameters of interest significantly affect the design effect if their different combinations are applied in the derived expressions.
- Investigating the effect of variability among cluster sizes (M_i), cluster means (Y_i), within clusters variances (S_i^2) and sample sizes (n), and after taking and combining several different levels of the variables, let us take a look at table (1) in the appendix, which represents a factorial model analysis. It shows the four factors, each at two levels, that are combined in a full factorial model, that takes the form;

$Y_{ijks} = \mu + M_i + Y_j + n_k + v_s + (MY)_{ij} + (Mn)_{ik} + (Mv)_{is} + (Yn)_{jk} + (Yv)_{js} + (nv)_{ks} + (MYn)_{ijk} + (MYv)_{ijs} + (Ynv)_{jks} + (Mnv)_{iks} + e_{ijks}$

i=1, 2; j=1, 2; k=1, 2; s=1, 2

It can be seen that all main effects of factors, M (M_i variance), Y (Y_i variance), n, v (S_i^2), have statistically significant effect on the design effect. Some of the two-factor and three-factor interactions have proved that the interaction between those two or three factors contributes significantly to the variation in the design effect i.e. all those having p-value less than 0.05, the rest of the interactions do not contribute to or have no effect on the Deff, for example, ($M_i * n_k$) , ($M_i * v_s$) and so on.

The coefficients and p-value for the main effects and interactions are shown in table (1a) The interpretation of the coefficients for level 1 of M_i is that the effect on the design effect averaged over the other level (level 2) is to increase the Deff by (3.4256). Regarding the interpretation of interactions, for example, level 1 of M_i and level 1 of Y_i leads to decrease the Deff by (3.4292), and the interpretation goes on like this.

- Comparing the effect on the design effect of the variability in sample sizes in equations (9) & (10), when ρ represents different levels leads to table (2) below which shows the Deff for different levels of $\bar{b}$, b' & ρ, in which the intra-class correlation coefficient ρ, is taken at four levels, namely 0.05, 0.1, 0.5 and 0.9, while $\bar{b}$ values and their corresponding b' values are taken at three levels, representing low, medium and high. And as expected, the design effect increases as ρ increases. Equation (10) yields also design effect values that are higher than those of equation (9), as a result of using a weighted average sub sample sizes. Thus, the more the values of the average sub sample size $\bar{b}$, the weighted average sub sample size b' & the intra-cluster correlation coefficient ρ, the higher the levels of the design effect.

Table 2: Design effects for two-stage cluster sampling with probabilities proportional to size for (b-) & (b')

b- / b'		ρ =0.05		ρ =0.1		ρ =0.5		ρ =0.9	
10	13	1.45	1.6	1 .9	2.2	5.5	7.0	9.1	11.8
20	26	1.95	2.25	2.9	3.5	10.5	13.5	18.1	23.5
50	65	3.45	4.2	5.9	7.4	25.5	33.0	45.1	58.6

- In case of single-stage cluster sampling for equal cluster sizes, table (3) below demonstrates the fact that the design effect increases as the ratio of within cluster variance to between cluster variance decreases.

Table 3: Design Effects due to single-stage cluster sampling (equal cluster sizes) for some values of S^2_w, S^2_b

S^2_w / S^2_b	A
0.001	0.999001
0.051	0.951470
0.151	0.868810
2.000	0.333333
10.000	0.090909
50.000	0.019608
100.000	0.009101

- Regarding two-stage cluster sampling, table (4) shows various combinations of S_1^2 & S_2^2 values that are denoted by [(1.1)-(1.6)] and used to calculate the design effect.

Table 4: Combinations 1.1 to 1.6 of S^2_1, S^2_2, for population of size 5000

Combinations	S^2_1	S^2_2
1.1	900	1000
1.2	1000	900
1.3	15	15
1.4	15	10
1.5	10	100
1.6	100	10

- Accordingly table (5) confirms that the design effect increases as the cluster sizes increase and the between cluster variability increases.

Table 5: Design Effects due to two- stage cluster sampling (equal cluster sizes) for S^2_1, S^2_2 for combinations (1.1) to (1.6)

N	M	n	m	Deff 1.1	Deff 1.2	Deff 1.3	Deff 1.4	Deff 1.5	Deff 1.6
500	10	100	5	2.533	2.734	2.2473	3.0125	1.000	4.143
250	20	50	15	7.039	7.768	6.0124	8.783	1.2649	12.964
200	25	25	15	7.114	7.840	6.0923	8.853	1.7470	13.043
100	50	20	25	10.884	12.136	9.3703	13.745	2.5739	20.472
50	100	15	30	11.503	12.701	9.8280	14.381	2.8275	21.454
40	125	10	50	20.411	22.625	17.3196	25.730	4.4304	38.843
20	250	5	70	27.861	30.945	23.5685	35.289	5.8827	53.847
10	500	4	100	32.939	36.696	27.7465	42.026	6.7921	65.304
5	1000	2	300	107.533	120.782	89.4921	139.860	19.6902	227.341

- For three-stage cluster sampling, using combinations of S^2_1, S^2_2 and S^2_3 of table (6), the design effect increases, as can be seen from table (7), as the third-stage units sizes increase but the variance among the third-stage units is the smallest.

Table 6: Combinations (2.1 to 4.6) of S^2_1, S^2_2 and S^2_3 for population of size 5000

Combinations	S^2_1	S^2_2	S^2_3
2.1	800	900	1000
2.2	800	1000	900
2.3	1000	900	800
2.4	1000	800	900
2.5	900	800	1000
2.6	900	1000	800
3.1	8	10	15
3.2	8	15	10
3.3	10	15	8
3.4	10	8	15
3.5	15	10	8
3.6	15	8	10
4.1	10	100	150
4.2	10	150	100
4.3	100	150	10
4.4	100	10	150
4.5	150	100	10
4.6	150	10	100

Table 7: Design Effects due to Three - stage cluster sampling (equal cluster sizes) for S^2_1, S^2_2 and S^2_3 for combinations [(2.1 to 2.6), (3.1 to 3.6), (4.1 to 4.6)] with different values of N, M, n, m, R & r

N	M	R	n	m	r	Deff 2.1	Deff 2.2	Deff 2.3	Deff 2.4	Deff 2.5	Deff 2.6
50	30	10	10	10	5	14.37	14.46	17.46	17.40	15.85	16.00
45	30	10	10	10	5	14.08	14.16	17.10	17.03	15.52	15.67
40	30	10	10	10	5	13.70	13.79	16.63	16.57	15.10	15.25
35	30	10	10	10	5	13.21	13.30	16.03	15.96	14.55	14.70
30	20	8	10	10	5	13.06	13.11	15.88	15.83	14.42	14.53
25	20	8	10	10	5	12.18	12.24	14.79	14.74	13.43	13.54
20	20	8	10	10	5	10.76	10.83	13.06	12.97	11.84	11.97
15	20	5	5	10	3	8.81	8.21	9.91	9.92	9.07	9.06

10	20	5	5	10	3	6.80	6.82	8.18	8.18	7.49	7.50
5	20	5	3	10	3	6.02	6.04	7.25	7.24	6.61	6.65

N	M	R	n	m	r	Deff 3.1	Deff 3.2	Deff 3.3	Deff 3.4	Deff 3.5	Deff 3.6
50	30	10	10	10	5	12.04	12.38	14.90	14.47	20.88	20.79
45	30	10	10	10	5	11.79	12.14	14.61	14.17	20.44	20.33
40	30	10	10	10	5	11.48	11.83	14.23	13.78	19.87	19.76
35	30	10	10	10	5	11.08	11.44	13.74	13.28	19.14	19.03
30	20	8	10	10	5	10.93	11.18	13.49	13.18	19.01	18.95
25	20	8	10	10	5	10.20	10.47	12.61	12.27	17.67	17.60
20	20	8	10	10	5	9.04	9.34	11.19	10.80	15.53	15.44
15	20	5	5	10	3	6.94	6.97	8.36	8.37	11.82	11.85
10	20	5	5	10	3	5.77	5.85	6.96	6.90	9.72	9.73
5	20	5	3	10	3	5.10	5.20	6.19	6.09	8.65	8.64

N	M	R	n	m	r	Deff 4.1	Deff 4.2	Deff 4.3	Deff 4.4	Deff 4.5	Deff 4.6
50	30	10	10	10	5	3.46	4.01	18.49	17.51	25.95	25.51
45	30	10	10	10	5	3.43	3.98	18.12	17.12	25.40	24.94
40	30	10	10	10	5	3.39	3.94	17.65	16.63	24.70	24.22
35	30	10	10	10	5	3.32	3.89	17.04	15.98	23.79	23.28
30	20	8	10	10	5	3.03	3.47	16.71	16.05	23.57	23.35
25	20	8	10	10	5	2.94	3.40	15.61	14.87	21.92	21.63
20	20	8	10	10	5	2.80	3.28	13.85	12.97	19.26	18.86
15	20	5	5	10	3	2.03	2.22	10.15	10.34	14.41	14.78
10	20	5	5	10	3	1.89	2.11	8.45	8.43	11.87	12.07
5	20	5	3	10	3	1.80	2.03	7.55	7.41	10.66	10.78

- Regarding probability proportional to size, in single-stage cluster sampling, the design effect increases as the ratio of within cluster variance to between cluster variance decreases. This is so for both low and high levels of within clusters' variances, S_i^2. And for the two-stage cluster sampling, the design effect increases as the intra-cluster correlation coefficient increases ρ, meanwhile the average and weighted average sub sample sizes, $\bar{b}$ and b', increase.

4. Conclusion

An attempt has been made to study the clustering component of design effect, and its behavior under various circumstances. This was through deriving expressions for the design effect in terms of the parameters affecting its pattern.

The constraints of this study are revealed in the fact that the case of equal cluster sizes is the only case that is taken into consideration. Further studies are required to investigate the clustering component of design effect under the conditions of different clusters' sizes. Moreover, the obtained results are a consequence of the cases of single, two, and three stages of cluster sampling. Thus, it is hoped that these results may be extended to cover more than the previous cases, and will help throw more light on the way the design effect changes as the above factors change. Sample designers should, at the design stage find some guidelines in these findings that can help them determining the expected precision of their intended sampling design. So the need arises for future work which may include exploring and examining the estimation and precision of the design effect in its different cases.

References

[1] David A. Lacher, Lester R. Curtin, Jeffery P. Hughes "Why large Design Effects Can Occur In Complex Sample Designs". Internet: http://www.cdc.gov/nchs/about/major/ nhanes/currentnhanes.htm , Feb. 22, 2015 [September 2004].

[2] Hans Peterson. And Pedro Luis D.N.S "Analysis of Design Effects for Surveys In Developing Countries", in *Proc. Household Sample Surveys In Developing & Transition Countries*, 2005, pp.123-143.

[3] Holt, D. H. Discussion of the paper by Verma, V., C. Scott and C.O' Muircheartaigh: "Sample Designs and Sampling Errors for the World Fertility Survey". *Journal of the Royal Statistical Society, Series* , vol. 143, pp. 468-469, 1980.

[4] Kalton, G. Michael Brick J., Thanh Le^ "Estimating components of design effects for use in sample designs", in *Proc. Household sample surveys in Developing and Transition Countries***,** 2005, pp. 95-121**.**

[5] Kish, L. "Methods for design effects". *Journal of Official Statistics*, vol. 11, pp. 55-77, 1995.

[6] Lynn, P. and Gabler, S.. "Approximations to b* In the Prediction of Design Effects Due To Clustering", *Journal of Survey Methodology*, Vol. 31, pp. 101-104.

[7] Siegfried Galber, Sabine Hader and Lahiri Partha ."A model based justification of Kish's formula for design effects for clustering & weighting". Internet: http://www.iser.essex.ac.uk/pubs/workpaps/, Feb. 22, 2015 [Dec. 2005]**.**

Appendix

Table I: General Linear Model: Deff versus Mi, Yj, nk, Vs

Factor	Type Levels	Values
Mi	fixed	2 1 2
Yj	fixed	2 1 2
nk	fixed	2 1 2
Vs	fixed	2 1 2

Table II: Analysis of Variance for Deff, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Mi	1	187.75	187.75	187.75	560.24	0.027
Yj	1	3761.14	3761.14	3761.14	1.1E+04	0.006
nk	1	209.58	209.58	209.58	625.39	0.025
Vs	1	1376.76	1376.76	1376.76	4108.19	0.010
Mi*Yj	1	188.15	188.15	188.15	561.42	0.027
Mi*nk	1	22.53	22.53	22.53	67.23	0.077
Mi*Vs	1	2.25	2.25	2.25	6.71	0.235
Yj*nk	1	209.02	209.02	209.02	623.72	0.025
Yj*Vs	1	1368.54	1368.54	1368.54	4083.66	0.010
nk*Vs	1	53.09	53.09	53.09	158.41	0.050
Mi*Yj*nk	1	22.48	22.48	22.48	67.09	0.077
Mi*Yj*Vs	1	2.21	2.21	2.21	6.61	0.236
Mi*nk*Vs	1	0.33	0.33	0.33	0.99	0.502
Yj*nk*Vs	1	52.86	52.86	52.86	157.72	0.051
Error	1	0.34	0.34	0.34		
Total	15	7457.04				

Table III: Coefficients & P-value for the main effects & interactions

Term	Coef	SE Coef	T	P
Constant	15.3659	0.1447	106.17	0.006
Mi	3.4256	0.1447	23.67	0.027
Yj	-15.3320	0.1447	-105.94	0.006
nk	-3.6192	0.1447	-25.01	0.025
Vs	9.2762	0.1447	64.10	0.010
Mi*Yj	-3.4292	0.1447	-23.69	0.027
Mi*nk	-1.1867	0.1447	-8.20	0.077
Mi*Vs	-0.3749	0.1447	-2.59	0.235
Yj*nk	3.6144	0.1447	24.97	0.025
Yj*Vs	-9.2484	0.1447	-63.90	0.010
nk*Vs	-1.8215	0.1447	-12.59	0.050
Mi*Yj*nk	1.1854	0.1447	8.19	0.077
Mi*Yj*Vs	0.3720	0.1447	2.57	0.236
Mi*nk*Vs	0.1437	0.1447	0.99	0.502
Yj*nk*Vs	1.8176	0.1447	12.56	0.051

S = 13.22 R-Sq = 74.2% R-Sq(adj) = 64.9