Modeling Cluster Design Effects When Cluster Sizes Vary

James R. Chromy
RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709

Abstract

The intracluster correlation coefficient is often used to model the design effect associated with cluster sampling. Typically, the average cluster size is inserted into the design effect formula even though the cluster size will vary in practice. Cluster size variation is particularly a problem when the design effect for domain estimates is desired and the domain of interest is not evenly distributed over all sample clusters. Should an alternate way of computing the intracluster correlation be developed or is some other simple solution available to resolve this problem? This paper presents an alternate, but simple, model for accounting for both clustering and the variation in cluster sizes.

Key Words: Sample design, variance models, cluster sampling

1. Overview

In advance of sample design and selection appropriate cost and variance models are required in order to develop an efficient design that meets all (or most) of the stated objectives. The objectives are usually translated into required survey estimates and upper bounds on the standard errors associated with those estimates.

Model-based thinking is required to develop models that can predict the standard errors that should result from a particular design. This is in contrast to robust estimation procedures used to actually estimate the standard errors when the survey data are finally available.

Even though good models may have several components, this paper focuses on the design effect associated with clustering. In particular, it addresses the clustering effect when the cluster sizes vary. Total cluster size can vary if approximate size measures are used and the final allocation to the cluster is set to satisfy specified sampling rates. Unit nonresponse within clusters can also contribute to unequal sized clusters. Most surveys support not only aggregate population estimates, but also estimates pertaining to a number of specified domains. Aggregate population cluster sizes can be controlled to some extent and may be fairly stable. Domain cluster sizes cannot be so easily controlled and may have a large variance.

This paper explores the use of the coefficient of variation (CV) of the sample cluster size as a way of modeling the cluster design effect. In order to focus on clustering effect, other aspects of the overall design effect are ignored and variance models are simplified whenever possible.

2. Definitions and Notation

Modeling the clustering effect on the variance is simplest when cluster sizes are equal and become less simple when incorporating variable cluster sizes.

2.1 Equal Cluster Sizes

The design effect is just the ratio of the variance obtained when recognizing the clustering effect (and other features of the sample design) to the base variance, V_o, that would have been obtained under a simple one-stage, with replacement, equal probability design.

$$Deff = V_{design}/V_o \tag{1}$$

Often the base variance is defined as arising from simple random sampling and would require incorporating a finite population correction factor into the base variance. Usually the finite population correction is ignored or assumed to be close to 1.0 as long as the overall sampling rate is low. With these assumptions, $V_o = \sigma^2/n$.

A model of the design-based variance can be written in terms of the variance components and sample size parameters for a clustered sample with equal-sized clusters as

$$V_{design} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} \tag{2}$$

where σ_1^2 and σ_2^2 are the first and second stage variance components, n_1 is the number of clusters, and n_2 is the number of elements in each cluster. For simplicity the design-based variance is shown for with replacement sampling at both stages. The cluster size, n_2, is treated as fixed in the standard setup.

The definition of the intracluster correlation coefficient, or ICC, follows Kish's (1965, pp. 161-162) definition which would be an approximation by some strict interpretations, but is most useful for getting back to variance component representation of the design-based variance. This form of the intracluster correlation coefficient is:

$$\rho = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} \text{ and } \sigma^2 = \sigma_1^2 + \sigma_2^2 \tag{3}$$

The design-based variance can then be expressed as:

$$V_{design} = \frac{\sigma^2}{n}[1 + \rho(n_2 - 1)] \text{ where } n = n_1 n_2. \tag{4}$$

Using this form of the design variance and inserting it into equation (1) yields:

$$Deff = 1 + \rho(n_2 - 1) \tag{5}$$

Note that equation (5) is a commonly-used presentation of the cluster design effect as 1 plus the ICC times the quantity of the cluster size minus one.

[1]

2.2 What Happens If Cluster Sizes Vary?

A common practice is to su
bstitute the mean cluster size into equation (5):

$$Deff = 1 + \rho(\bar{n}_2 - 1) \tag{6}$$

[1] A more rigorous form of the intracluster correlation coefficient allows $-1/(n_2 - 1) \leq \rho \leq 1$.

where $n_{2i} =$ number of elements in cluster i (not equal for all i), and the mean is calculated as

$$\bar{n}_2 = \sum_{i=1}^{n_1} n_{2i}/n_1 \, . \tag{7}$$

This approach is considered acceptable if the cluster sizes do not vary greatly. In general, this practice will underestimate the clustering design effect. Some practitioners have considered an alternative definition of the ICC to resolve the problem. I would advocate against that approach.

Once a structure is fixed, recognize the variance components and the ICC as population parameters. By structure, I mean the way the population and the frame are partitioned by strata, by clusters within strata, and by elements within clusters. Sometimes the structure is apparent in the population as with schools and students within schools. Other times it is part of the design process as with strata and clusters defined by contiguous land areas. In either case, I prefer to think of the variance components defined on this structure as population parameters. Since the ICC is derived from the variance components, it is also a population parameter.

The sample design then addresses the approach to selecting the cluster sample and achieving the planned sample size targets at each level. Cluster size variation can be both planned and unplanned. Domain cluster sizes are often the result of a screening process and application of fixed sampling rates causing them to vary much more wildly than the overall cluster sizes.

Holt (1980) put forward an alternative calculation of the mean cluster size based on weighting over all ultimate elements in the sample. Park *et al* (2003) incorporated Holt's approach in a more general model which factors in design effects for stratification, clustering, and unequal weighting. Chromy and Myers (2001) also considered a more complex model incorporating several factors into the overall design effect and used the CV methodology discussed in this paper at both the area cluster and the household level. Eldridge, Ashby, and Kerry (2006) develop a number of models for the design effect including one based on CV methodology discussed here; interestingly, their paper is in an epidemiology journal, not a statistics or survey methodology journal.

Holt's representation just uses the weighted average, n'_2 , in the equation for the clustering design effect:

$$V_{design} = \frac{\sigma^2}{n}[1 + \rho(n'_2 - 1)] \tag{8}$$

Holt's n'_2 can be thought of as an average over the all n ultimate sample elements or as a weighted average computed at the cluster level:

$$n'_2 = \sum_{i=1}^{n_1} w_i n_{2i} \ \text{ and } \ w_i = n_{2i}/(n_1 \bar{n}_2) \tag{9}$$

With a little algebra, Holt's weighted mean cluster size can also be expressed as function of the unweighted mean of cluster size and the coefficient of variation of cluster size. First, rewrite Holt's weighted mean as:

$$n'_2 = \frac{1}{n} \sum_{i=1}^{n_1} n_{2i}^2 \tag{10}$$

Then, if we treat the target cluster size, $\bar{n}_2$, as known

$$\hat{\sigma}_{n_2}^2 = \frac{\sum_{i=1}^{n_1} n_{2i}^2 - n_1 \bar{n}_2^2}{n_1} \tag{11}$$

and

$$n'_2 = \frac{\hat{\sigma}^2_{n_2} + \bar{n}^2_2}{\bar{n}_2} = \bar{n}_2\left(1 + CV^2_{n_2}\right) . \quad (12)$$

If we estimate the cluster size, $\bar{n}_2$, this solution is still a good approximation. This form was also developed in the paper by Eldridge, Ashby, and Kerry for application to cluster randomized trials. It can easily be inserted into the formula for the clustering design effect:

$$Deff = 1 + \rho\left[\bar{n}_2\left(1 + CV^2_{n_2}\right) - 1\right]. \quad (13)$$

The average cluster size is usually a clear objective of a survey design and, as a target value, would be known during planning. The coefficient of variation for cluster size can be based on similar surveys, prior rounds of the same survey, or on reasonable assumptions about the expected distribution of cluster sizes.

This general approach can be compared to the use of the CV of weights to model the unequal weighting effect. Similar algebraic steps are used to express the unequal weighting effect in terms of the coefficient of variation of the weights.

3. A Simple Example

To illustrate that the approach works at least in one extreme case, consider a case where the variation in cluster size is by design. This could arise in making regional estimates where a fourth of the sample clusters and a fourth of the total sample are allocated to one region. Table 1 shows stratum 1 as the one coinciding with the domain of interest.

Table 1. Stratified Sample Example

Stratum	Number of Clusters	Average Cluster Size	Domain Sample Size	Squared CV of Cluster Size
1	100	12	1,200	0
2, 3, and 4	300	0	0	0
Total	400	3	1,200	3.0

If we consider stratum 1 by itself, the average cluster size is 12 and the CV is 0. The clustering design effect is then just

$$Deff = 1 + \rho(12 - 1).$$

If we were not aware of the stratification or chose to ignore it, we could compute the mean cluster size, the variance of the cluster size, and the squared coefficient of variation of cluster size over all 400 clusters are $\bar{n}_2 = 3,\ \hat{\sigma}^2_{n_2} = 27, and\ CV^2_{n_2} = 3$. Inserting these results in equation (13) would yield the numerical equivalent to analyzing only stratum 1 data.

$$Deff = 1 + \rho[3(1 + 3) - 1].$$

This example illustrates that the design effect using the CV of cluster size works when we have an alternative way of computing the design effect.

4. When Should We Worry About Variable Cluster Size?

When should we be concerned about using the design effect model that assumes equal cluster sizes? If the ICC is low (near zero) and the CV of cluster size is low (near zero), the model utilizing average cluster size will be almost right. In fact if either condition holds, ICC low or CV of cluster size low, the average cluster size model (equation (6)) of the cluster design effect will be close to correct.

Table 2 compares the modeled design effect when ignoring cluster size variability by using equation (6) (column with the CV of cluster size =0) with design effects which account for cluster size variability (columns with CVs of cluster size ranging from 0.1 to 2.0) using equation (13). Intracluster correlation coefficients (ICCs) of 0.01 and 0.05 represents low clustering effects; ICCs of 0.10 and 0.20 represent much larger clustering effect, but ones that are quite common in school surveys where cluster sizes are also likely to be large. The chosen CVs also represent low and high potential values. Higher CVs will generally apply to small domains that would have low average cluster size even though the total cluster size is much higher as shown in Table 6 (Chromy and Myers, 2001).

The ICC = 0.05 and CV=0.3 combination shows that only modest increases in design effect occur by taking account of cluster variability with average clusters sizes of 1 and 3; much larger, but still reasonable, increases occur with larger cluster sizes. For CVs greater than 0.5, the equation (6) values are serious under projections of the clustering design effect for larger average cluster sizes.

Table 2. Model Design Effect Comparisons (Rounded to 2 Decimal Places)

	Average Cluster	CV of Cluster Size								
ICC	Size	0	0.1	0.2	0.3	0.4	0.5	1	1.5	2
0.01	1	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.02	1.04
0.05	1	1.00	1.00	1.00	1.00	1.01	1.01	1.05	1.11	1.20
0.10	1	1.00	1.00	1.00	1.01	1.02	1.03	1.10	1.23	1.40
0.20	1	1.00	1.00	1.01	1.02	1.03	1.05	1.20	1.45	1.80
0.01	3	1.02	1.02	1.02	1.02	1.02	1.03	1.05	1.09	1.14
0.05	3	1.10	1.10	1.11	1.11	1.12	1.14	1.25	1.44	1.70
0.10	3	1.20	1.20	1.21	1.23	1.25	1.28	1.50	1.88	2.40
0.20	3	1.40	1.41	1.42	1.45	1.50	1.55	2.00	2.75	3.80
0.01	10	1.09	1.09	1.09	1.10	1.11	1.12	1.19	1.32	1.49
0.05	10	1.45	1.46	1.47	1.50	1.53	1.58	1.95	2.58	3.45
0.10	10	1.90	1.91	1.94	1.99	2.06	2.15	2.90	4.15	5.90
0.20	10	2.80	2.82	2.88	2.98	3.12	3.30	4.80	7.30	10.80
0.01	25	1.24	1.24	1.25	1.26	1.28	1.30	1.49	1.80	2.24
0.05	25	2.20	2.21	2.25	2.31	2.40	2.51	3.45	5.01	7.20
0.10	25	3.40	3.43	3.50	3.63	3.80	4.03	5.90	9.03	13.40
0.20	25	5.80	5.85	6.00	6.25	6.60	7.05	10.80	17.05	25.80

5. Average Cluster Sizes Smaller Than One

It is not unusual for special domain sample sizes to be quite small and less than one per cluster. Small average cluster sizes can also occur when screening for a person who meets certain specifications (e.g., females 15 to 35 years old). Suppose the sample design allows selecting at most one eligible person per cluster (*e.g.,* a household) and only a fraction, *p<1*, of selected persons meet the eligibility requirements. Applying equation (6) would indicate that the cluster sampling design would have a design effect less than 1.

$$1 + \rho(p - 1) < 1.$$

Now consider equation (13) with $\bar{n}_2 = p,\ \hat{\sigma}^2_{n_2} = p(1 - p), and\ CV^2_{n_2} = (1 - p)/p.$

$$1 + \rho\left[p\left(1 + \tfrac{1-p}{p}\right) - 1\right] = 1.$$

In practice, the realized domain sample size itself is a random variable when applying a screening approach to a fixed total sample. Controlling the cluster sample size will usually result in an unequal weighting effect, but that is not the topic of this paper. The sample design may provide some allowance for the expected variation in the realized sample size of each target domain and design effects due to unequal weighting.

6. Conclusions and Recommendations

A few points summarize the lessons learned from examining alternative clustering design effect models:

1. Tracking both the ICC and CV of cluster size in existing surveys can be very helpful in designing future surveys.
2. For a given frame structure, variance components and the ICC should be treated as population parameters.
3. In most surveys, cluster size variability can only partially be controlled.
4. Expect some cluster size variability and incorporate it into any model of clustering design effects during the planning process.
5. Use the CV of cluster size as a convenient and somewhat portable measure for cluster size variability.

References

Chromy, James R., and Lawrence E. Myers. 2001. "Variance Models Applicable to the NHSDA." *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Eldridge, Sandra M., Deborah Ashby, and Sally Kerry. 2006. "Sample Size for Cluster Randomized Trials: Effect of Coefficient of Variation of Cluster Size and Analysis Method." *International Journal of Epidemiology* no. 35:1292-1300.

Holt, D. 1980. "Discussion of Paper by Verma, Scott, and O'Muircheartaigh." *Journal of the Royal Statistical Society, A* no. 143 (4):468-469.

Park, Inho, Marianne Winglee, Jay Clark, Keith Rust, Andrea Sedlak, and David Morganstein. 2003. "Design Effects and Survey Planning." *Proceedings of the Section on Survey Research Methods, American Statistical Association*:3179-3185.