Journal of the Korean Statistical Society (2005), **34**: *3, pp 185–195*

A GENERALIZATION OF THE INTRACLASS CORRELATION IN CLUSTER SAMPLING

KYU-SEONG KIM[1]

ABSTRACT

This article is concerned with the intraclass correlation in survey sampling. From a design-based viewpoint the intraclass correlation is generalized to a finite population with unequal sized clusters. Under simple random cluster sampling the intraclass correlation is given in an explicit form, which is a generalization of the usual one. The range of it is found and the design effect is expressed by means of it. An example is given to compare the intraclass correlation with the homogeneity measure numerically, which shows that two measures are not the same except some limited cases.

AMS 2000 subject classifications. Primary 62D05.
Keywords. Design-based intraclass correlation, Design effect, Homogeneity measure.

1. INTRODUCTION

The efficiency of cluster sampling in a finite population is affected by two sources of variation : a selection method and the population structure. When a sample of clusters is chosen by simple random sampling the variance of the cluster sample mean could be represented by the sample size as well as the measure of population structure. In this situation two measures for the population structure have been commonly used : one is the intraclass correlation and the other one is the homogeneity measure.

The intraclass correlation is a measure of the correlation between two elements in the same cluster. In a fixed finite population with equal sized clusters let y_{ij} is the value of the jth element in the ith cluster then the intraclass correlation of y_{ij} and $y_{ij'}$ is defined as

Received July 2004; accepted August 2005.
[1]Department of Statistics, University of Seoul, Seoul 130-743, Korea (e-mail: kskim@uos.ac.kr)

$$\rho_0 = \frac{\sum_{i=1}^{N}\sum_{j\neq j'}^{M_0}(y_{ij}-\bar{\bar{Y}})(y_{ij'}-\bar{\bar{Y}})/(M_0-1)}{(M-1)S_y^2} \tag{1.1}$$

where $\bar{\bar{Y}}$ is the overall mean of the survey variate y_{ij}, $S_y^2 = \sum_i \sum_j (y_{ij}-\bar{\bar{Y}})^2/(M-1)$ the population variance of y_{ij}s, N the number of clusters, M_0 the cluster size and $M = NM_0$ the population size. (*e.g.*, Kish, 1965; Cochran, 1977; Sändal *et al.*, 1992). Unfortunately, as commented by Särndal *et al.*(1992), it is defined only for populations with equal sized clusters, so it is necessary to extend the intraclass correlation to a population with unequal sized cluster since they are more common than equal sized clusters in most practical situations.

The homogeneity measure is an alternative to represent the population structure, which, unlike the intraclass correlation, is suitable for both cases, equal and unequal sized clusters(*e.g.* Hansen *et al.*, 1953; Särndal, *et al.*, 1992) :

$$\delta = 1 - \frac{S_W^2}{S_y^2}. \tag{1.2}$$

where $S_W^2 = \sum_i \sum_j (y_{ij} - \bar{\bar{Y}}_i)^2/(M-N)$ and $\bar{\bar{Y}}_i$ is the ith cluster mean. Under the finite population with equal sized clusters, two measures are almost the same. The difference is given by

$$\delta - \rho_0 = \frac{1}{M-1}\frac{S_W^2}{S_y^2},$$

which becomes zero when M is large. For these reasons the homogeneity measure is frequently regarded as a generalization of the intraclass correlation even under the population with unequal sized clusters (e.g., Clark and Steel, 2002). But such a generalization may not be true when the cluster sizes are quite different. In the next section, we will investigate the difference of the two measures.

This article deals with the intraclass correlation from the design-based viewpoint in cluster sampling. In section 2, we generalize the definition of the intraclass correlation to the population with unequal sized clusters, which will be compared with the homogeneity measure to explain the difference of the two. In section 3, we derive a generalized intraclass correlation under simple random cluster sampling and find the range of the values. In addition we express the design effect in terms of the intraclass correlation as well as the effect of unequal sized clusters. Also it is compared with the homogeneity measure theoretically. An example is given to examine the difference of the two numerically in section 4. The final section includes a summary and conclusion.

2. GENERALIZATION OF THE INTRACLASS CORRELATION

2.1. Generalization

We consider a finite population whose M elements are grouped into N clusters and whose ith cluster has M_i elements so that $\sum_{i=1}^{N} M_i = M$. Denote by y_{ij} the value of jth element of ith cluster. In design-based approach of survey sampling the population values are regarded as fixed values, so the population structure is dependent on survey values as well as the size of population and clusters. In such a clustered population the intraclass correlation is originally defined as

$$\rho = \frac{E\{(y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}})\}}{E(y_{ij} - \bar{\bar{Y}})^2}, \quad j \neq j' \tag{2.1}$$

In order to calculate the design-based expectation in the above formula a specific sampling design is presented. For general derivation let $p(ij)$ be a selection probability of element j in cluster i and $p(ij,\ ij')$ a selection probability of elements $j\&j'$ in cluster i. Then the intraclass correlation could be represented as follows,

$$\rho = \frac{\sum_i \sum_{j \neq j'} (y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}}) p(ij,\ ij')}{\sum_i \sum_j (y_{ij} - \bar{\bar{Y}})^2 p(ij)} \tag{2.2}$$

where $\sum_i \sum_j p(ij) = \sum_i \sum_{j \neq j'} p(ij, ij') = 1$.

According to a sampling design, a specific form of ρ will be driven by calculating selection probabilities. As an example we consider a two-stage sampling scheme, in which a cluster is selected by probability proportional to size (PPS), M_i, sampling and then elements are selected by simple random sampling from the selected clusters. In this case we have $p(ij) = (M_i/M)(1/M_i) = 1/M$ and $p(ij,\ ij') = (M_i/M)(1/((M_i - 1)M_i)) = 1/M(M_i - 1)$. Then the intraclass correlation given in (4) can be reduced to the following,

$$\rho_{pps} = \frac{\sum_i \{\sum_{j \neq j'} (y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}})/(M_i - 1)\}}{(M-1)S_y^2}. \tag{2.3}$$

2.2. Homogeneity Measure

Some researchers including a referee for this article mentioned that the homogeneity measure could be regarded as an extension of the intraclass correlation to the population with unequal sized clusters. It can be true in some cases, but at the same time it may not be true in other cases. It is because the construction of

the two measures have different basis. While the homogeneity measure is based only on the population structure, in which sampling scheme is not included, the intraclass correlation is on the basis of both the population structure and a sampling scheme. The formula (2.1) is an unique definition most statisticians agree with but a specific formula given in (2.2) has various forms according to sampling schemes. As a result in some situations the intraclass correlation defined in (2.2) could be similar to the homogeneity measure, but in other situations they may be quite different. In this subsection, we examine the difference of the two measures.

After some algebra we have another formula of the homogeneity measure as follows,

$$\delta = \frac{1}{S_y^2(M-N)}\{\sum_i (\frac{1}{M_i} - \frac{N-1}{M-1}) \sum_j (y_{ij} - \bar{\bar{Y}})^2 + \sum_i \frac{1}{M_i} \sum_{j \neq j'} (y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}})\} \tag{2.4}$$

Comparing the above formula with the formula (2.3), we come to know that two formulas could be similar when cluster sizes are approximately equal to M_0 and the cluster sizes are quite large so that $M_i \approx M_i - 1$, in which the first term in the brace will be zero and the other corresponding terms are very similar. But when either cluster sizes are quite different or the sampling scheme is not probability proportional to size sampling, two measure may not be close. It means that the homogeneity measure should be used as an extension of the intraclass correlation in limited cases such as PPS sampling.

3. Intraclass Correlation in Cluster Sampling

3.1. A Generalized Intraclass Correlation

The intraclass correlation in (1.1) is appropriate for the population with equal sized clusters. We now generalize this formula to the population with unequal sized clusters. We consider the simple cluster sampling from such a population. We choose n clusters from N clusters by simple random sampling and then we survey all elements in the selected cluster. Then the probabilities $p(ij)$ and $p(ij, ij')$ are constants for all elements and pairs of elements. Since sum of selection probabilities for all cases should be equal to be one, we easily have $p(ij) = 1/M$ and $p(ij, ij') = 1/\sum_i M_i(M_i - 1)$. The following is a generalization of the intraclass correlation under cluster sampling from a population with unequal sized clusters.

Definition 3.1. *Under simple random cluster sampling from a population*

with unequal sized clusters, the intraclass correlation as in (2.2) is reduced as

$$\rho_w = \frac{\sum_i \sum_{j \neq j'} (y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}}) / \sum_i M_i(M_i - 1)}{(M-1)S_y^2/M}. \tag{3.1}$$

Clearly, the above expression is equal to the expression in (1.1) when all clusters are equal, i.e., $M_i = M_0$.

Now we try to find the range of ρ_w, which is slightly different from that of equal sized case because it is affected by the difference of cluster sizes. The following gives a simple answer.

THEOREM 3.1. *The range of the intraclass correlation ρ_w as in (3.1) is given by*

$$-\frac{1}{\bar{M}(1 + C_M^2) - 1} \leq \rho_w \leq 1 \tag{3.2}$$

where $\bar{M} = \sum_{i=1}^N M_i/N$ is the average of cluster sizes and $C_M^2 = \sum_i \sum_i (M_i - \bar{M})^2/(N\bar{M}^2)$ is relative variance of cluster sizes.

PROOF. After clustering we can decompose the overall variance, S_y^2, over the population into between variance, S_B^2, and within variance, S_W^2, such that $S_y^2 = S_B^2 + (M-N)S_W^2/(M-1)$, where $S_B^2 = \sum_i M_i(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2/(M-1)$ and $\bar{\bar{Y}}_i = \sum_j y_{ij}/M_i$ is the ith cluster mean. Then the values of ρ_w should lie between two extreme cases : one is $S_B^2 = 0$ and the other one is $S_W^2 = 0$. The former is the case of perfect heterogeneity and the latter is the case of perfect homogeneity in clustering.

When the clusters are perfectly homogeneous, *i.e.*, $y_{ij} = \bar{\bar{Y}}_i$ for all i, the numerator of ρ_w in (2.1) is equal to the denominator from which it follows $\rho_w = 1$. For the case of perfect heterogeneity we consider the following equation,

$$\sum_i [\sum_j (y_{ij} - \bar{\bar{Y}})]^2 = \sum_i \sum_j (y_{ij} - \bar{\bar{Y}})^2 + \sum_i \sum_{j \neq j'} (y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}})$$

from which each term can be calculated as follows :

$$(i) \quad \sum_i [\sum_j (y_{ij} - \bar{\bar{Y}})]^2 = \sum_i M_i(M_i - 1)(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2 + (M-1)S_B^2$$

$$(ii) \quad \sum_i \sum_j (y_{ij} - \bar{\bar{Y}})^2 = (M-1)S_B^2 + (M-N)S_W^2$$

$$(iii) \quad \sum_i \sum_{j \neq j'} (y_{ij} - \bar{\bar{Y}})(y_{ij'} - \bar{\bar{Y}}) = \rho_w (M-1)S_y^2 \sum_i M_i(M_i - 1)/M$$

Using these equations we have an alternative representation of the intraclass correlation as follows,

$$\rho_w = \frac{M}{(M-1)S_y^2}\left\{\frac{\sum_i M_i(M_i-1)(\bar{\bar{Y}}_i-\bar{\bar{Y}})^2}{\sum_i M_i(M_i-1)} - \frac{(M-N)S_W^2}{\sum_i M_i(M_i-1)}\right\} \tag{3.3}$$

Now putting $\bar{\bar{Y}}_i = \bar{\bar{Y}}$ into the above equation we get $\rho_w = -1/(\bar{M}(1+C_M^2)-1)$ as the lower bound of the intraclass correlation. □

This result is in agreement with the case of equal cluster size in which $C_M^2 = 0$, so that $-1/(\bar{M}-1) \leq \rho_0 \leq 1$. Theorem 2 says that the range of intraclass correlation when the cluster sizes are not equal is slightly smaller than that of the equal sized cluster case.

3.2. The Design Effect under Simple Cluster Sampling

The design effect is commonly used to represent the efficiency of a complex sampling design compared with simple random sampling design with the same sample size. It is well known that the design effect under simple random cluster sampling is given by .

$$\text{Deff} = 1 + \rho_0(\bar{M}-1) \tag{3.4}$$

It is true when the cluster sizes are all the same. However, it may not be true when the cluster sizes are not equal, because the unequal size of clusters gives a sort of effect to the variance of simple random cluster mean. The following theorem shows how does the different sizes give an effect to the variance formula and then to the design effect.

THEOREM 3.2. *We consider simple random cluster sampling and assume that N is so large that $1/N$ can be negligible. Then the design effect is represented as*

$$\text{Deff} = 1 + \rho_w(\bar{M}(1+C_M^2)-1) + \bar{M}\frac{C_M^2 + 2C_{MY}}{C_y^2} \tag{3.5}$$

where $C_y^2 = S_y^2/\bar{\bar{Y}}^2$ is the relative variance of the value y_{ij}s and

$$C_{MY} = \sum_i \left(\frac{M_i}{M}\right)\frac{(M_i-\bar{M})(\bar{\bar{Y}}_i-\bar{\bar{Y}})}{\bar{M}\bar{\bar{Y}}},$$

which is the relative covariance of the cluster size M_i and the cluster mean $\bar{\bar{Y}}_i$ weighted by relative cluster size M_i/M.

PROOF. Let $Y_i = \sum_j y_{ij}$ be the ith cluster total and s_c be a sample of clusters with size n_c chosen from N clusters by simple random sampling, then the variance of a usual unbiased estimator for the population mean, $\bar{\bar{Y}}_c = \sum_{i \in s_c} Y_i/(\bar{M}n_c)$, is given by

$$V_{CL} = \frac{1}{\bar{M}^2} \frac{S_{cy}^2}{n_c} \tag{3.6}$$

where $S_{cy}^2 = \sum_i (Y_i - \bar{Y})^2/(N-1)$ and $\bar{Y} = \sum_i Y_i/N$. The population variance of the cluster totals, S_{cy}^2, can be easily shown that

$$S_{cy}^2 = \bar{M} S_y^2 [1 + \rho_w(\bar{M}(C_M^2 + 1) - 1)] + \bar{\bar{Y}}^2 \bar{M}^2 (C_M^2 + 2C_{MY}) \tag{3.7}$$

Now let m be the expected number of elements with n_c clusters, then $m = n_c\bar{M}$. So, under simple random sampling with the same size of sample, the variance of sample mean becomes $V_{SRS} = S_y^2/(n_c\bar{M})$. Hence the design effect, which is the ratio of V_{CL} to V_{SRS}, is given by the result as desired. □

REMARK 3.1. *If $M_i = \bar{M}$, then the design effect becomes*

$$\text{Deff} = 1 + \rho_0(\bar{M} - 1),$$

because $\rho_w = \rho_0$ and $C_M^2 = C_{MY} = 0$.

The similar results as in Theorem 3 can be found in Clark & Steel (2002) and Särndal *et al.*(1992). Clark & Steel derived the design effect under two stage sampling in which simple random sampling is employed in each stage. But their variance formula was expressed by means of sum of squares between cluster totals instead of the intraclass correlation. In Särndal, *et al.* (1992), the design effect is represented by the homogeneity measure defined as in (1.2).

3.3. Homogeneity Measure

In order to investigate the difference between the intraclass correlation and the homogeneity measure in the population with unequal sized clusters, it is necessary for two measures to be represented by means of the variation of cluster

sizes and the correlation between the cluster sizes and the cluster means. After some algebra, it can be easily shown that

$$\rho_w - \delta = \frac{M}{(M-1)S_y^2}\left\{\sum_i (w_i - w_i')(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2 + \frac{(M-N)S_W^2/M}{(\bar{M}-1)(1+(\bar{M}-1)/\bar{M}C_M^2)}\right\} \tag{3.8}$$

where $w_i = M_i(M_i-1)/\sum_i M_i(M_i-1)$ and $w_i' = M_i/\sum_i M_i$ are weights attached to the y variate. The first term in the brace of the right hand side in the above equation will be greater than zero when the cluster sizes are positively correlated with the cluster mean and be negative if not. Furthermore, the value is almost zero if the cluster sizes and the cluster means are almost uncorrelated. The second term in the brace is always non-negative, and the value become much larger as the relative variance of the cluster sizes C_M^2 is much greater. So the difference is directly affected the variation of the cluster sizes and the correlation between the cluster sizes and the cluster means. In an example in the following section, it will be shown that the difference could not be small in a real situation.

4. AN EXAMPLE

An empirical study was conducted to examine the effect of unequal cluster sizes to the intraclass correlation as well as the design effect and also to compare the intraclass correlation with the homogeneity measure numerically.

First, we considered an approximate intraclass correlation calculated as if cluster sizes are all equal such as

$$\rho_0 = 1 - \frac{\bar{M}}{\bar{M}-1}\frac{(M-N)S_W^2}{(M-1)S_y^2}$$

which becomes the exact value under the same size of clusters, but an approximate value under unequal sizes of clusters. The relative difference, RD,

$$RD(\%) = \frac{\rho_0 - \rho_w}{\rho_w} \times 100$$

is considered to measure the amount of percents when the intraclass correlation of unequal sizes of clusters is approximated as if the sizes of clusters are the same.

Next, we separated the design effect into two parts :

$$D_1(\%) = \frac{1}{\text{Deff}}\{1 + \rho_w(\bar{M}(1+C_M^2) - 1)\} \times 100$$

and

$$D_2(\%) = \frac{1}{\text{Deff}} \bar{M} \frac{(C_M^2 + 2C_{MY})}{C_y^2} \times 100,$$

in which D_2 purely occurs from the unequal sizes of clusters. Comparing two terms, we could find how much percents does the design effect affected by the unequal sizes of clusters.

Finally, in order to compare the intraclass correlation with the homogeneity measure we calculated the difference of the two quantities and decomposed it into two parts such as $\rho_w - \delta = E_1 + E_2$ where

$$E_1 = \frac{M}{(M-1)S_y^2} \sum_i (w_i - w_i')(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2$$

and

$$E_2 = \frac{(M-N)S_W^2}{(M-1)S_y^2} \left\{ (\bar{M} - 1)(1 + \frac{\bar{M} - 1}{\bar{M} C_M^2}) \right\}$$

The first term E_1 occurs from the correlation between the cluster sizes and the cluster means, and the second term from the variation of cluster sizes.

A part of 2000 Korean agricultural census data is used for our study as a dataset, which consists in 29,509 clusters with different sizes. The average value of clusters is 38.80 and the coefficient of variation is about 0.98. Among a lot of survey variables in the dataset, six key variables are chosen as follows :

- Rice1 - area of rice field of one's own
- Rice2 - area of rice field borrowed for farming
- Farm1 - area of farm of one's own
- Farm2 -area of farm borrowed for farming
- Crop - crop area
- Sales - sales of agricultural products

Table 4.1 shows the values of the intraclass correlation and the design effect for six variables. The values of six intraclass correlations are below 0.2, and the approximate intraclass correlations have similar range. However, the values of relative difference between two correlations take negative values up to 45%, from which it can be noted that the exact intraclass correlation is greater than the approximate intraclass correlation for six variables and it is because of the effect of unequal sizes of clusters. Since they are not small, if the approximate intraclass correlation is used as if it is exact intraclass correlation in the situation of unequal clusters, it can be misleading. The effect of unequal sizes of clusters is

TABLE 4.1 *The intraclass correlation and the design effect*

	Intraclass correlation			Design effect		
Variables	ρ_w	ρ_0	RD	Deff	D_1	D_2
Rice1	0.170	0.200	-17.6 %	28.371	48.7 %	51.2 %
Rice2	0.087	0.099	-13.7 %	13.140	57.9 %	42.0 %
Farm1	0.147	0.203	-38.0 %	18.811	64.5 %	35.4 %
Farm2	0.068	0.099	-45.5 %	7.406	82.9 %	17.0 %
Crop	0.158	0.182	-15.1 %	29.140	44.4 %	55.5 %
Sales	0.100	0.127	-27.0 %	19.920	43.0 %	56.9 %

TABLE 4.2 *The intraclass correlation and the homogeneity measure*

Variables	ρ_w	δ	$\rho_w - \delta$	E_1	E_2
Rice1	0.170	0.200	-0.030	-0.041	0.010
Rice2	0.087	0.099	-0.011	-0.023	0.012
Farm1	0.147	0.203	-0.055	-0.066	0.010
Farm2	0.068	0.099	-0.031	-0.043	0.011
Crop	0.158	0.182	-0.024	-0.034	0.010
Sales	0.100	0.127	-0.026	-0.038	0.011

also found in the design effect. D_2 in the Table 4.1 take values from 17% to 56%, which comes from the unequal sizes of clusters. Also we can find much higher values of the design effect ranged from 7.4 to 29.1, which is because the cluster sizes are larger on average in single cluster sampling. Practically, however, the design effects could be lowered by subsampling from each selected cluster.

Table 4.2 gives some interesting results about the effect of unequal cluster sizes in comparing the intraclass correlation with the homogeneity measure. It is noted that the homogeneity measure δ is almost same with ρ_0 by the definition. The values of difference between the quantities for six variables are all negative, in which the first term E_1s are all negative and the second term E_2s are all positive. Negative E_1 means that M_is are more strongly correlated with $(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2$ than $M_i(M_i - 1)$. In fact the correlation between M_i and $(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2$ is greater negatively than between $M_i(M_i - 1)$ and $(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2$. Table 4.2 says that the intraclass correlation is not equal to the homogeneity measure when the cluster sizes are not the same, and the intraclass correlation is less than the homogeneity measure if the cluster sizes are correlated negatively with the cluster means. On the other hand, the cluster sizes are postively correlated with the cluster means, then the intraclass correlation will be greater than the homogeneity measure.

5. CONCLUSION

In this article we investigated the intraclass correlation from the design-based viewpoint in cluster sampling. We generalized the definition of the intraclass correlation to the population with unequal sized clusters and then compared it with the homogeneity measure theoretically. A generalized intraclass correlation was also derived under simple random cluster sampling. The range of the values was found and the design effect was expressed by means of the newly defined intraclass correlation. An example was given to examine the difference of the two measures numerically.

As shown in section 3, the intraclass correlation and the homogeneity measure are different measures, because the former is constructed on the basis of the population structure as well as a sampling scheme, but the latter is dependent only on the population structure. As a natural consequence, two measures come to be similar in some limited situations. In other situations, we need more attention in using two measures.

REFERENCES

CLARK, R.G. AND STEEL, D.G. (2002). "The effect of using household as a sampling unit", *International Statistical Review*,**70**, 289–314.

COCHRAN, W.G. (1977). *Sampling techniques*, Wiley.

HANSEN, M., HURWITZ, W. AND MADOW, W.(1953). *Sample survey methods and theory*, **vol. 1** and **2**, Wiley.

KISH, L.(1965). *Survey sampling*, Wiley.

SÄRNDAL, C.E., SWENSSON, B. AND WRETMAN, J.(1992). *Model assisted survey sampling*, Springer-Verlag.