ESTIMATING INTRACLUSTER HOMOGENEITY IN MULTISTAGE SAMPLES

Anne F. Clemmer and William D. Kalsbeek, University of North Carolina at Chapel Hill

INTRODUCTION

Data collection costs must be considered when a sample survey is designed. When costs are considered, invariably a cluster sample will be chosen over a simple random sample. Cluster sampling reduces data collection costs by taking advantage of the fact that units of the population are often found in close geographic proximity. If costs were the only consideration, then the survey would be conducted in a few immense clusters. However, the important item to consider is the number of independent observations per cluster (Sudman, 1976; Kish, Groves, and Krotki, 1976). Intracluster homogeneity is a measure of the homogeneity of responses within a cluster. Less information is collected in a cluster with a large measure of homogeneity than in one with a small measure. Different variables provide different measures of intracluster homogeneity within the same cluster. For example, a cluster may be regionally homogeneous but heterogeneous with respect to age.

Intracluster homogeneity is primarily used in survey design to determine the total sample size. Kish (1965) defined a measure of intracluster homogeneity which he called roh or rate of homogeneity as

$$\text{roh}\ (\bar{y}) = \frac{\text{deff}\ (\bar{y}) - 1}{\bar{b} - 1} \qquad (1.1)$$

where deff $(\bar{y})$ is the design effect for $\bar{y}$ or ratio of the variance of the design under consideration to the variance of a simple random sample of the same total size and $\bar{b}$ is the average sample cluster size. The concept of roh emerges from rho or intraclass correlation which is defined only for the special case of two-stage sampling of equal clusters with simple random sampling at each stage. An estimate of roh is often available from a previous survey or a pretest. This estimate of roh may then be used to impute a value of roh for another variable as well as a design effect and standard error for the new variable (Kish, Groves, and Krotki, 1976). The standard error and design effect for the new variable are then used to determine the total sample size for the new variable.

The importance of intracluster homogeneity has been observed by several authors. Kish and Frankel (1974) note that clustering induces large and positive correlations between element values with a resultant increase in variance. They report increases in variance for both linear and nonlinear statistics. Holt, Smith, and Winter (1980) and Nathan and Holt (1980) have investigated the effect of intracluster homogeneity on variance of regression coefficients in complex surveys. Ordinary least squares estimates of variance were found in both papers to be underestimates. It was speculated that this situation occurred because the particulars of the design such as intracluster homogeneity and stratification were not taken into account. More recently, Hansen, Madow, and Tepping (1983), in their comparison of inference from model-dependent and probability-sampling surveys, stress the relevance of design in sample surveys. They suggest that intracluster homogeneity is particularly important and that failure to recognize such effects may lead to underestimates of variance and understatement of confidence intervals. Lastly, the importance of including clustering effects in tests of hypotheses from complex survey data has been researched by Fellegi (1980) and Rao and Scott (1981). Both papers show that under a complex sample design, the usual chi-squared test statistics for goodness of fit and independence are asymptotically distributed as the weighted sum of independent chi-squared variables where the weights are functions of the design effects. The design effect is a function of cluster homogeneity as shown in equation 1.1.

We have seen that roh is both useful and important. It would be helpful to have some reasonable method for estimating roh. Equation 1.1 is generally used. However, we speculate that when roh is large, the following formula may perform better:

$$\text{roh}\ (p) = 1 - \frac{1}{H} \Sigma_h^H \frac{\Sigma_i^{a_h} b_{hi} \check{P}_{hi} (1 - \check{P}_{hi})}{a_h \bar{b}_h \check{P}_h (1 - \check{P}_h)} \qquad (1.2)$$

where H is the number of strata, a_h is the number of selected PSUs in stratum-h, b_{hi} is the number of sample elements in PSU-i of stratum-h, $\check{P}_{hi}$ is the attribute proportion in PSU-i of stratum-h, b_h is the average sample cluster size in stratum-h, and $\check{P}_h$ the sample attribute proportion for stratum -h.

Equation 1.1 is called the design effect method and equation 1.2, the proportion variation method because the expression was derived by realizing that roh is approximately the proportion of the total variation not accounted for within clusters. The objective of this research is to determine which of these two methods actually comes closer to measuring true intracluster homogeneity.

One example where the proportion variation method may work better is in calculating roh for degree of urbanization. If a PSU is rural, then the sample attribute proportion for cluster-i in stratum-h, $\check{p}_{hi}$, is zero; conversely, in urban PSUs, $\check{p}_{hi}$ is one and $(1 - \check{p}_{hi})$ is zero. Thus, for each case, the summation in equation 1.2 is zero and roh is one. Roh was calculated for degree of urbanization using the design effect method and found to be 0.3 which is not near one; thus, some credence was lent to the speculation that the proportion variation method may be superior.

METHODOLOGY

To investigate whether the design effect or proportion variation method is superior for estimating the rate of homogeneity, ten independently replicated samples were selected from the National Medical Care Expenditure Survey (NMCES) household file which thus serves as the population to which inference will be made. Independent replication enabled the variance to be calculated for sample estimates and thus, to test hypotheses at known probability levels. NMCES was sponsored by the National Center for Health Services Research (NCHSR) with support from the National Center for Health Statistics (NCHS). Its objectives were to analyze how Americans use health care services and to determine the patterns and characteristics of health expenditures and health insurance.

The NMCES sample is composed of 40,320 individuals in approximately 14,000 households and 142 primary sampling units (PSUs) or clusters. The sampling design was documented elsewhere (Cohen and Kalsbeek, 1981). To control the workload associated with this research, it was decided to create a study population from the NMCES sample. An equal cluster population of 50 individuals per cluster was agreed upon. The population for study in this dissertation is thus a two-stage, equal cluster population composed of 142 clusters and 50 individuals per cluster.

As mentioned earlier, ten independently replicated samples were selected from the study population so that variances of sample estimates could be calculated for hypothesis testing. These replicated samples or replicates were selected to be representative of the study population as well as being independent and approximately self-weighting. The sample size per replicate was somewhat arbitrarily set at 500 so the total of ten replicates would not yield an unusually large number of observations. Each replicate consisted of two stages with 20 clusters selected from the total of 142 at the first stage and 25 individuals from the total of 50 at the second. Thus, each replicate consisted of 500 individuals (20 PSUs X 25 individuals/PSU) with an equal number of individuals selected per cluster.

The study population contained about 200 analysis variables and 25 domain variables. Analysis variables are reporting variables and domain variables are those for which a specific sample size is planned in the survey (Kish, 1965). For example, region and sex are common domain variables. A subset of these analysis and domain variables was identified for intensive study.

The goal in the selection of analysis variables was to represent a variety of levels of roh. Four ranges were determined and are listed in Table 1 along with the selected analysis variables in each category. Continuous and categorical analysis variables are present both for the low and high range of roh. Only one category was selected for each of the categorical variables; the response of "good" was chosen for perceived health status and the response of "yes" for whether insured by a Health Maintenance Organization or not.

Domain variables were selected to include both cross-class and segregated types, as well as small and large domain size categories. Selected domain variables are enumerated in Table 2. Segregated domains are those in which only one domain category is present in a cluster; region and urban or rural status are examples of segregated domain variables. Cross-class domains are those where several domain categories are present in each cluster; for instance, age, sex, and income.

Sample rates of homogeneity were calculated by the design effect and proportion variation estimation methods for each of the 52 different categories formed by the four analysis variables in each of the 13 domain categories (see Tables 1 and 2). The sample rohs calculated by each estimation method were determined separately for each of the ten samples. Additionally, the true rate of homogeneity was calculated from the study population for each of the 52 categories. A relative deviation was determined for each estimation method in each of the 52 categories and ten samples by comparing the sample roh with the true value and dividing by the true value. The average differences in the relative deviations for each estimation method were compared for each of the 52 domain categories. The testing procedure is detailed in the following paragraphs.

A relative deviation was calculated for each estimation method in each of the 52 categories and ten samples as

$$RD_{ijk} = \frac{r_{ijk} - \rho_i}{\rho_i},$$

where i = domain category from 1 to 52;
j = estimation method, where 1 is design effect and 2 is proportion variation;
k = sample from 1 to 10;
r_{ijk} = sample rate of homogeneity for ith domain category using jth estimation method from kth sample;
ρ_i = true rate of homogeneity for domain category -i.

A difference in the absolute values of the relative deviations for each estimation method was calculated for each domain category (i) and sample (k) as

$$d_{i.k} = |RD_{i1k}| - |RD_{i2k}| .$$

Absolute values were used to determine which of the two relative deviations was closer to zero (i.e., the difference of absolute values) or which of the two methods actually came closer to estimating the true rate of homogeneity.

We then calculated an average difference for each domain category across the ten samples as

$$d_{i..} = \frac{\Sigma_k^{10} d_{i.k}}{10}.$$

For a cross-class domain, we conjectured that the design effect method is superior to the proportion variation method. (i.e., $|RD_{i1k}|$ was closer to zero than $|RD_{i2k}|$ and $d_{i.k}$ was, thus, negative.) We, therefore, wanted to test the hypotheses

$$H_0: \Delta_i.. = 0$$

$$H_1: \Delta_i.. < 0$$

where $\Delta_i..$ is the true average difference in absolute values of the relative deviations for domain category -i. It can be shown that if the null hypothesis is true

$$t = \frac{\bar{d}_{i..}}{\text{ste}(\bar{d}_{i..})}$$

follows a t distribution with 9 degrees of freedom where

$$\text{ste}\ (\bar{d}_{i..}) = \sqrt{\frac{\Sigma_k^{10}\ d^2_{i.k} - \frac{\left[\Sigma_k^{10}\ \ d_{i.k}\right]^2}{10}}{9(10)}}$$

and $d_{i.k} = |RD_{i1k}| - |RD_{i2k}|$. We rejected the null hypothesis if t was less than the t distribution critical value associated with a Type I error rate of α and 9 degrees of freedom. In a similar manner, we speculated that the proportion variation method is superior to the design effect method for segregated domains. We then wanted to test the hypotheses

$$H_0: \Delta_i.. = 0$$

$$H_1: \Delta_i.. > 0.$$

The test statistic and critical value were computed exactly as before except we rejected the null hypothesis if the test statistic was greater than the critical value. We thus had 52 tests which would determine for each of the domain categories which estimation method was better.

DISCUSSION

Sampling rates of homogeneity are summarized by analysis variable, estimation method, and type of domain in Table 3. Of particular interest are the four domain size categories; size 1 is a small domain containing about five percent of the population, and size 4 is a large domain composing approximately 50 percent of the population.

We note from Table 3 that in general the mean of roh for the proportion variation method is larger than for the design effect method implying as predicted that the proportion variation method may better estimate large rates of homogeneity. Secondly, we see that standard deviations of roh for the proportion variation method are smaller than for the design effect method; thus, the proportion variation method generates a more compact distribution than the design effect method. And lastly, for each estimation method, we observe that standard deviations are larger for small domains than large ones implying both methods are more variable for estimating roh in small domains than large ones.

In Table 4, the results of the paired data t test are presented for the two categorical analysis variables. The type and size of domain are specified as well as the preferred method. The preferred method is based on the outcome of the t test. Significance is noted at the five percent level. We see, as hypothesized, that the proportion variation method is definitely preferred for segregated domains; and the design effect method is generally the method of choice for cross-class domains.

In Table 5, we have the same comparison for the two continuous variables. The method of choice is unclear for segregated domains. For the first continuous variable, the design effect method is significantly preferred in one case and the proportion variation method in the other, although not significantly. The same outcome again occurs among segregated domains for the second continuous variable. For cross-class domains, however, a pattern is beginning to emerge. It appears that the design effect method is preferred for larger domains and the proportion variation method for smaller ones. Re-examining Table 4, we also notice this same pattern for cross-class domains.

Thus, we may conclude from these preliminary results that first, for segregated domains, the proportion variation method does appear to be preferred to the design effect method for estimating rates of homogeneity among categorical variables. The method of choice is not clear for continuous variables. Secondly, for cross-class domains, it appears that the design effect method is preferable for estimating roh in large domains, whereas the proportion variation method may be the method of choice for small domains.

ACKNOWLEDGMENTS

The research associated with this paper was funded by the Veterans Administration Medical Center, Durham, North Carolina 27705. The authors would also like to thank Dana Quade for his helpful comments and Ruth Drum for her careful typing.

REFERENCES

Cohen, S. B. and Kalsbeek, W. D. (1981), NMCES Estimation and Sampling Variances in the Household Survey, National Center for Health Services Research, Hyattsville, Maryland.

Fellegi, I. P. (1980), Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples, Journal of the American Statistical Association, 75, 261-268.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953), Sample Survey Methods and Theory, Vol. I and II, New York: John Wiley and Sons, Inc.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys, Journal of the American Statistical Association. 78, 776-807.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980), Regression Analysis of Data from Complex Surveys, Journal of the Royal Statistical Society A, 143, 474-487.

Kish, L. (1965), Survey Sampling, New York: John Wiley and Sons, Inc.

Kish, L. and Frankel, M. R. (1974), Inference From Complex Samples, Journal of the Royal Statistical Society B, 36, 1-37.

Kish, L., Groves, R. M., and Krotki, K. P. (1976), Sampling Errors for Fertility Surveys, Occasional Paper No. 17. London: International Statistical Institute, World Fertility Survey.

Nathan, G. and Holt, D. (1980), The Effect of Survey Design on Regression Analysis, Journal of the Royal Statistical Society B, 42, 377-386.

Rao, J. N. K. and Scott, A. J. (1981), The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, Journal of the American Statistical Association, 76, 221-230.

Sudman, S. (1976), Applied Sampling, New York: Academic Press.

Table 1. Selected analysis variables

Category	Roh	Description	Type
Very small	0.001	Average dollars for all hospital admissions	Continuous
Small	0.021	Perceived health status "good"	Categorical
Medium	0.070	Average time in minutes to get to usual source of care	Continuous
Large	0.179	Insured by Health Maintenance Organization "Yes"	Categorical

Table 2. Selected domain variables

Description	Categories and percentages	Total cat.	Type
Race	White non-Hispanic (65%), Black non-Hispanic (13%), Hispanic and other (5%)	3	Cross-class
Age	0-18 (32%), 19-59 (52%), 60-64 (4%), 65-102(12%)	4	Cross-class
Disabled veteran status	Disabled male vet (1%), Other male vet (11%), Other male 20+(18%), All other (70%)	4	Cross-class
Urbanization	Urban, rural	2	Segregated
		13	

Table 3. Comparison of rates of homogeneity by estimation method and analysis variable

Categorical analysis variable:
Perceived Health Status "Good"

		Design Effect		Proportion Variation	
DOMAIN	No. Est.	Mean	Std. Dev.	Mean	Std. Dev.
Total	10	0.059	0.057	0.067	0.037
Urbanization	20	0.128	0.163	0.064	0.040
Size 1	25	0.132	3.999	0.327	0.193
Size 2	30	0.262	0.515	0.305	0.126
Size 3	20	0.165	0.150	0.240	0.080
Size 4	30	0.065	0.065	0.096	0.042

Categorical analysis variable:
Insured by Health Maintenance Organization "Yes"

Total	10	0.152	0.112	0.128	0.065
Urbanization	11	0.039	0.192	0.130	0.066
Size 1	6	0.629	3.393	0.489	0.258
Size 2	16	0.122	0.345	0.244	0.173
Size 3	18	0.090	0.261	0.209	0.176
Size 4	29	0.101	0.142	0.128	0.124

Continuous analysis variable:
Average dollars for all hospital admissions

Total	10	-0.001	0.014	-0.041	0.001
Urbanization	20	0.001	0.053	-0.041	0.001
Size 1	19	-0.072	1.846	0.008	0.053
Size 2	30	-0.019	0.165	-0.033	0.010
Size 3	20	0.012	0.070	-0.037	0.003
Size 4	30	-0.005	0.020	-0.044	0.004

Continuous analysis variable:
Average time in minutes to get to usual source of care

Total	10	0.061	0.028	-0.041	0.001
Urbanization	20	0.137	0.133	-0.041	0.001
Size 1	25	0.844	2.378	0.037	0.050
Size 2	30	0.248	0.582	-0.014	0.022
Size 3	20	0.119	0.163	-0.024	0.013
Size 4	30	0.068	0.045	-0.042	0.005

Table 4. Comparison of segregated and cross-class domains for categorical analysis variables.

Perceived health status "Good"

TYPE	SIZE	CATEGORY		Preferred Method
SEG		Urban-ization:	Urban	PV*
SEG			Rural	PV*
CROSS	1	Race:	Hisp & Other	PV*
CROSS	2		Black	PV*
CROSS	4		White	DE
CROSS	1	Age:	60-64	PV
CROSS	2		65+	DE
CROSS	3		0-18	DE
CROSS	4		19-59	DE*
CROSS	1	Veteran Status:	Dis. Male Vet.	PV
CROSS	2		Other Male Vet.	DE*
CROSS	3		Other Male 20+	DE*
CROSS	4		All Others	DE

Insured by Health Maintenance Organization "Yes"

SEG		Urban-ization:	Urban	PV
SEG			Rural	-
CROSS	1	Race:	Hisp & Other	PV
CROSS	2		Black	PV
CROSS	4		White	DE
CROSS	1	Age:	60-64	-
CROSS	2		65+	DE*
CROSS	3		0-18	DE
CROSS	4		19-59	PV
CROSS	1	Veteran Status:	Dis. Male Vet.	-
CROSS	2		Other Male Vet.	DE*
CROSS	3		Other Male 20+	DE
CROSS	4		All Others	PV

*Denotes significance at the five percent level.

Table 5. Comparison of segregated and cross-class domains for continuous analysis variables.

Average dollars for all hospital admissions

TYPE	SIZE	CATEGORY		Preferred Method
SEG		Urban-ization:	Urban	DE*
SEG			Rural	PV
CROSS	1	Race:	Hisp & Other	PV*
CROSS	2		Black	PV*
CROSS	4		White	DE*
CROSS	1	Age:	60-64	PV*
CROSS	2		65+	PV
CROSS	3		0-18	PV
CROSS	4		19-59	DE*
CROSS	1	Veteran Status:	Dis. Male Vet.	DE
CROSS	2		Other Male Vet.	PV
CROSS	3		Other Male 20+	PV*
CROSS	4		All Others	DE*

Average time in minutes to get to usual source of care

SEG		Urban-ization:	Urban	PV
SEG			Rural	DE*
CROSS	1	Race:	Hisp & Other	PV*
CROSS	2		Black	PV*
CROSS	4		White	DE*
CROSS	1	Age:	60-64	PV*
CROSS	2		65+	PV
CROSS	3		0-18	PV*
CROSS	4		19-59	DE*
CROSS	1	Veteran Status:	Dis. Male Vet.	DE
CROSS	2		Other Male Vet.	PV
CROSS	3		Other Male 20+	PV
CROSS	4		All Others	DE*

*Denotes significance at the five percent level.