Survey Designs with Small Stratum Sample Sizes

Wayne A. Fuller[*] Emily Berg[†] Xiaofei Zhang[‡]

Abstract

One-per-stratum and Two-per-stratum designs are common survey sampling designs. The one-per-stratum design generally has the smallest variance, but no unbiased estimator of variance is available. We present an intermediate design which is a combination of the one-per-stratum and two-per-stratum designs, and for which an unbiased variance estimator is available. A closely related design is a one-per-stratum design with supplemental observations for variance estimation. We compare the variance and variance of estimated variance for versions of the intermediate design and for the supplement design.

Key Words: Survey design, stratified sampling , design variance.

1. Introduction

Stratified sampling is a component of the majority of survey samples. In particular, stratification is used for populations ordered such that elements close together are more alike than units far apart. When strata are formed on the basis of the ordered population and proportional allocation is used, typically the variance of an estimated mean will decrease as the number of strata increases. Therefore, one-per-stratum and two-per-stratum designs are common for ordered populations.

Cochran (1946) was one of the first to compare designs for an ordered population. He showed that systematic sampling was superior to stratified sampling for certain types of populations. Other authors studying design for ordered populations include Fuller (1970), Bellhouse (1984), Breidt (1995) and Papagorgiou and Karakostas (1998).

No design-unbiased estimator of variance exists for the one-per-stratum design, or for systematic sampling. We propose a design closely related to one-per-stratum for which design-unbiased variance estimation is possible.

2. Intermediate design: components of variance

2.1 Sampling procedure

Let a population of size N be divided into n equal sized sets. Let n be even and let the sets be formed into pairs. Let the sets be identified by $hj, h = 1, 2, \cdots, n/2 = H, j = 1, 2$. The stratified sample with one unit per stratum is created by selecting one unit at random from each of the n sets. The stratified sample with two units per stratum is created by selecting two units at random from the subpopulation formed by a pair of the original sets. If unequal probabilities are to be used in selection, we assume the probabilities sum to one for each set of the original n sets. We also assume that the joint probabilities for the two-per-stratum design are nonnegative such that variance estimation is possible. In the standard equal-probability two-per-stratum design the expected fraction of the samples that have an element in each of the two sets of the pair $(h1, h2)$ is one half.

[*] Iowa State University, 2438 Osborn Dr, Ames, IA 50011

[†] Iowa State University, 2438 Osborn Dr, Ames, IA 50011

[‡] Iowa State University, 2438 Osborn Dr, Ames, IA 50011

To consider an intermediate design, let the pairs of sets be formed into groups of K pairs. The total number of elements to be selected from a group is $2K$. Let one of the pairs be chosen at random and let one set of the pair of sets, chosen at random, be assigned two sample elements, and the other set of the pair assigned zero elements. One element is selected from each set of the pair for the remaining $K-1$ pairs of sets.

2.2 Design variance

Assume an equal probability design. Then the sample mean for the proposed sample, for a group of $2K$ sets is

$$\begin{aligned}\bar{y}_{2K} &= (2K)^{-1}[\sum_{t=1}^{2} y_{11t} + \sum_{h=2}^{K}\sum_{j=1}^{2} y_{hjt}] \\ &= K^{-1}[\bar{y}_{11.} + \sum_{h=2}^{K} \bar{y}_{h..}],\end{aligned} \tag{1}$$

where y_{hjt} is observation t on set hj, $\bar{y}_{h..}$ is the mean of the two observations for pair h, $\bar{y}_{11.}$ is the mean of the two observations in set one of pair one, and with no loss of generality, we assume the two units are selected in set one of the two sets of pair one. By the method of selection, every unit has the same probability of selection, and

$$E\{\bar{y}_{2K}\} = \bar{y}_N,$$

where $\bar{y}_N$ is the population mean.

To study properties of the design, define

$$y_{hjt} = \mu_h + b_{hj} + e_{hjt}, \tag{2}$$

where $\mu_h + b_{hj}$ is the population mean for set hj and the pair population mean is $\mu_h + 0.5(b_{h1} + b_{h2})$. Let

$$\begin{aligned} V\{b_{hj}\} &= \sigma^2_{b,h}, & E\{\sigma^2_{b,h}\} &= \sigma^2_b, \\ V\{e_{hjt}\} &= \sigma^2_{e,hj}, & E\{\sigma^2_{e,hj}\} &= \sigma^2_e. \end{aligned}$$

We ignore the finite population correction and assume equal probabilities of selection. The MSE of the sample mean conditional on the selection of set 11 to receive two units is

$$MSE\{\bar{y}_{2K}|b_{12} \notin A\} = (2K)^{-2}[2\sigma^2_{e,11} + (b_{11} - b_{12})^2 + \sum_{h=2}^{K}\sum_{j=1}^{2} \sigma^2_{e,hj}],$$

where $b_{12} \notin A$ denotes no observations in set 12. The unconditional MSE for a sample of size $2K$ is the average of $MSE\{\bar{y}_{2K}|b_{12} \notin A\}$ over all possible assignments of pairs and points within pairs. Thus,

$$E[MSE\{\bar{y}_{2K}\}] = 0.5(K^{-1}\sigma^2_e + K^{-2}\sigma^2_b), \tag{3}$$

We will also call the $E\{MSE\}$ of (3) $V\{\bar{y}_2K\}$. For a sample of size n with one two-zero-observation pair of sets in every $2K$ sets,

$$V\{\bar{y}_n\} = n^{-1}(\sigma^2_e + K^{-1}\sigma^2_b),$$

where $n = 2Kr$ and r is the number of replications of $2K$ sets. The standardized variance is

$$nV\{\bar{y}_n\} = \sigma_e^2 + K^{-1}\sigma_b^2. \tag{4}$$

For $K = 2$, the variance of (3) is that of the two-per-stratum design, except for finite population terms. As K is increased, the variance of the design moves toward the variance of the one-per-stratum design.

2.3 Variance estimation

To construct an estimator of variance let

$$\widehat{bws} = 0.5(K-1)^{-1}\sum_{h=2}^{K}(y_{h1} - y_{h2})^2 \tag{5}$$

and

$$\widehat{wms} = 0.5(y_{111} - y_{112})^2, \tag{6}$$

where, as before, we let the pair with both elements in a single set be the pair with $h = 1$. We have

$$E\{\widehat{bws}\} = \sigma_e^2 + \sigma_b^2 \tag{7}$$

and

$$E\{\widehat{wms}\} = \sigma_e^2. \tag{8}$$

It follows that

$$\hat{V}\{\bar{y}_n\} = r^{-1}\hat{V}\{\bar{y}_{2K}\}$$

where

$$\hat{V}\{\bar{y}_{2K}\} = (2K)^{-1}[K^{-1}\widehat{bms} + (1 - K^{-1})\widehat{wms}]. \tag{9}$$

is a nonnegative unbiased estimator of the variance of the estimated mean. An unbiased estimator of variance for a finite population is

$$\hat{V}\{\bar{y}_{2K}\} = (2K)^{-1}[K^{-1}\widehat{bms} + [1 - (KM)^{-1}(K + M)]\widehat{wms}]$$

To evaluate the effect of K on the efficiency of variance estimators, assume b_{hj} and e_{hjt} are normally and independently distributed. Then,

$$V\{\widehat{bms}\} = 2(K-1)^{-1}(\sigma_e^2 + \sigma_b^2)^2, \tag{10}$$

$$V\{\widehat{wms}\} = 2\sigma_e^4, \tag{11}$$

and

$$\begin{aligned} V\{2K\hat{V}(\bar{y}_{2K})\} &= KV\{K^{-1}\widehat{bms} + (1 - K^{-1})\widehat{wms}\} \\ &= 2K^{-2}(K-1)^{-1}(\sigma_e^2 + \sigma_b^2)^2 + 2(1 - K^{-1})^2\sigma_e^4. \end{aligned} \tag{12}$$

The standardized variance of standardized estimated variance is

$$KV\{2K\hat{V}(\bar{y}_{2K})\} = 2K^{-1}(K-1)^{-1}(\sigma_e^2 + \sigma_b^2)^2 + 2K(1 - K^{-1})^2\sigma_e^4. \tag{13}$$

Table 1: Example Variances for Components-of-Variance Normal Distribution Model.

	$(\sigma_b^2, \sigma_b^2) = (1, 1)$			$(\sigma_b^2, \sigma_b^2) = (0.5, 1)$		
K	$2KV\{\bar{y}_{2K}\}$	$KV\{2K\hat{V}(\bar{y}_{2K})\}$	df^*	$2KV\{\bar{y}_{2K}\}$	$KV\{2K\hat{V}(\bar{y}_{2K})\}$	$df*$
2	1.50	5.00	54.00	1.25	3.25	57.6
3	1.33	4.00	53.3	1.17	3.42	48.0
4	1.25	5.17	36.3	1.12	4.87	30.9
5	1.20	6.40	25.4	1.10	6.62	21.9
10	1.10	16.29	7.5	1.05	16.25	8.1

*Approximate degrees of freedom of variance estimator for $n = 120$.

Examples of the standardized variances and standardized variance of variance are given in Table 1. The trade-off between variance of the estimated mean and variance of the estimated variance is clear in the table. The use of $K = 4$ gives efficiency for the mean half way between one-per-stratum and two-per-stratum and gives an estimator of variance with reasonable variance.

Remark. Let $r_v = \sigma_b^2/\sigma_e^2$. Then the standardized variance of variance for $K = 2$ and $\sigma_e^2 = 1$, is

$$V\{\hat{V}|K = 2\} = (1 + r_v)^2 + 1$$

and the same quantity for $K = 3$ is

$$V\{\hat{V}|K = 3\} = (1/3)(1 + r_v)^2 + (8/3).$$

The two quantities are equal for $r_v = 0.58$. For ratios of σ_b^2 to σ_e^2 less than 0.58 the standardized variance of the variance increases as K increases. □□

Table 1 contains the approximate degrees of freedom for the variance estimator for a sample of $n = 120$. We define the approximate degrees of freedom for a sample of size $2K$ by

$$df_K = 2[V\{\hat{V}(\bar{y}_{2K})\}]^{-1}[V\{\bar{y}_{2K}\}]^2. \tag{14}$$

A sample of size n has $n/(2K)$ sets of size $2K$. The degrees of freedom in the table is $60K^{-1}df_K$. The degrees of freedom for $K = 2$ is not equal to $0.5n$ because the observations in a stratum of size two sets are not independent.

The variance can be estimated using replication methods. The squared difference for elements in different sets satisfies

$$E\{(y_{h1t} - y_{h2t})^2\} = 2(\sigma_e^2 + \sigma_b^2)$$

and the squared difference for elements in the same set, $t \neq r$,

$$E\{(y_{h1t} - y_{h1r})^2\} = 2\sigma_e^2.$$

Thus, for example, for a jackknife replicate created by deleting one of the elements in set $h1$, and doubling the weight of the other

$$E\{(\bar{y}^{(111)} - \bar{y})^2\} = E\{[(2K)^{-1}(y_{111} - y_{112})]^2\} = 0.5K^{-2}\sigma_e^2. \tag{15}$$

Likewise, for a stratum with an observation in each set of the two sets,

$$E\{(\bar{y}^{(h11)} - \bar{y})^2\} = E\{[(2K)^{-1}(y_{h11} - y_{h21})]^2\} = 0.5K^{-2}(\sigma_e^2 + \sigma_b^2). \qquad (16)$$

If one creates K replicates of this type, a jackknife estimator of the variance is

$$(K-1)^{-1}\sum_{h=2}^{K}(\bar{y}^{(h11)} - \bar{y})^2 + K(1-K^{-1})(\bar{y}^{(111)} - \bar{y})^2. \qquad (17)$$

The definition of the replicates can be changed to simplify the expression.

3. Intermediate design: autoregressive

Let the finite population of y-values be generated by the stationary autoregressive process

$$\begin{aligned} y_t &= \mu + \rho(y_{t-1} - \mu) + u_t, \\ u_t &\sim NI(0, \sigma_u^2). \end{aligned}$$

In much of the discussion we will let $\mu = 0$, with no loss of generality. We have

$$V\{y_t\} = (1-\rho^2)^{-1}\sigma_u^2 =: \sigma_y^2$$

and

$$C\{y_t, y_{t+h}\} = (1-\rho^2)^{-1}\rho^{|h|}\sigma_u^2.$$

The variance of a mean of M observations is

$$\begin{aligned} V\{\bar{y}_M\} &= \sigma_y^2\{M^{-1}(1-\rho)^{-1} - 2M^{-2}(1-\rho)^{-2}(\rho - \rho^{M=1})\} \qquad (18) \\ &= M^{-1}\sigma_y^2(1-\rho)^{-1}(1+\rho) + O(M^{-2}). \end{aligned}$$

The within stratum mean square for a stratum of size M is

$$\begin{aligned} S_{w,M}^2 &= M(M-1)^{-1}[\sigma_y^2 - V\{\bar{y}_M\}] \qquad (19) \\ &= M(M-1)^{-1}\sigma_y^2\{1 - (1-\rho)^{-2}[M^{-1}(1-\rho^2) - 2M^{-2}(\rho - \rho^{M+1})]\}. \end{aligned}$$

We define the between component for pairs of strata

$$\sigma_b^2 = 2(S_{w,2M}^2 - S_{w,M}^2).$$

An equivalent definition of σ_b^2 follows from

$$V\{\bar{y}_{2M}\} = V\{\bar{y}_M\} + 0.5\sigma_b^2.$$

See equation (4). The within and between components are given for several values of ρ for $M = 100$ in Table 2. The expected values of standardized variances for finite populations that are realizations of a first order autoregressive process are gives in Table 3. The stratum population size is 100 and the variance includes the finite population correction. The one-per-stratum variance in the last column of Table 3 is σ_w^2 of Table 2 multiplied by 0.99. For ρ less than 0.80, the variance of the mean for one-per-stratum is greater than 95% of the variance of the mean for two-per-stratum. For such ρ, moving from one-per-stratum to the intermediate design with K on the order of four, results in less than four percent loss of efficiency. For ρ larger than 0.8, there is a noticeable increase in variance of the mean as one moves from one-per-stratum to two-per-stratum.

Table 2: Within and between components for observations on an autoregressive process divided into sets of size $M = 100$.

ρ	0.2	0.5	0.8	0.9	0.95	0.99	0.995
σ_w^2	1.036	1.307	2.565	4.402	7.102	13.464	14.994
σ_b^2	0.006	0.026	0.207	0.814	2.769	16.792	23.396

Table 3: Standardized variance of estimated mean for autoregressive process. (Sampling rate = 0.01).

ρ	k				One-per-
	2	3	4	5	stratum
0.00	0.99	0.99	0.99	0.99	0.99
0.50	1.32	1.32	1.31	1.31	1.31
0.80	2.67	2.63	2.61	2.59	2.56
0.90	4.81	4.67	4.56	4.48	4.40
0.95	8.49	8.03	7.66	7.38	7.10
0.99	21.86	19.06	16.82	15.14	13.46
0.995	26.61	22.74	19.64	17.32	14.99

Table 4: Degrees of freedom for estimated variance of estimation mean for autoregressive process (Sample size = 100).

ρ	k			
	2	3	4	5
0.00	50.0	33.3	15.6	6.2
0.50	49.8	33.3	15.4	6.1
0.80	48.8	33.0	15.1	6.0
0.90	47.2	32.7	14.7	5.7
0.95	44.8	32.4	14.4	5.4
0.99	36.8	33.0	14.8	5.0
0.995	34.2	33.1	15.2	5.0

The variance of the estimated variance of $\bar{y}_{2K}$ is not a simple expression because of the correlation of observations of an autoregressive process. An approximation for the degrees of freedom of a variance estimator is defined in equation (14). Table 4 contains that approximation for the parameter combinations of Table 3. The loss of degrees of freedom as one increases K is fairly uniform across the values of ρ.

4. Supplemental observations

We now consider an alternative method of modifying the one-per-stratum design to obtain an unbiased estimator of variance.

As before, we consider a population ordered on an auxiliary variable. Let the population be divided into groups, where m_K is the number to be selected from each group. The group is divided into $m_K 1$ strata (sets). One set in the group is chosen at random and two elements are chosen in the selected set. One element is chosen in each of the remaining $m_K - 1$ sets to complete the sample. We call the design with supplemental observations a supplement design.

Clearly, the sample size, m_K , need not be even, but for comparison purposes, we give properties for $m_K = 2K$. The stratified estimator for a group with sample size $m_K = 2K$ is

$$\hat{\mu} = \sum_{h=1}^{(} 2K - 1)^{-1} \bar{y}_h, \tag{20}$$

where

$$\begin{aligned} \bar{y}_h &= 0.5 \sum_{j=1}^{n_k} y_{hj} && \text{if } \ n_h = 2 \\ &= y_{h1} && \text{if } \ n_h = 1 \end{aligned}$$

and n_h is the number selected in stratum h. The variance is

$$V\{\hat{\mu}\} = (2K - 1)^{-2}(2K - 1.5)\sigma_{e2}^2, \tag{21}$$

where σ_{e2}^2 is the within stratum variance. The variance estimator is

$$\hat{V}\{\hat{\mu}\} = (2K - 1)^{-2}(2K - 1.5)s^2, \tag{22}$$

where

$$s^2 = 0.5(y_{h1} - y_{h2})^2.$$

The variance of $\hat{V}\{\hat{\mu}\}$ is

$$V\{\hat{V}\{\hat{\mu}\}\} = [(2K - 2)^{-1}(2K - 1.5)]^2 V(s^2). \tag{23}$$

The sum of the squared weights defining the sample mean is larger for the supplemented design than for an intermediate design of the same sample size. Hence, the variance of the supplemented design is larger than that of the intermediate design of the same sample size for populations with small between components of variance. For many populations, the within component of variance is larger for the supplemented design because the strata are larger. On the other hand, the between component of variance contributes to the variance for the intermediate design.

Table 5: Relative Variance1 for alternative designs under the autoregressive model (M = 100).

ρ	$V(\hat{\mu})/V(\bar{y}_{2K})$			$V\{V(\hat{\mu})\}/V\{V(\bar{y}_{2K})\}$		
	(3,6)	(4,8)	(5,10)	(3,6)	(4,8)	(5,10)
0	1.082	1.063	1.050	2.294	1.897	1.664
0.2	1.081	1.062	1.050	2.294	1.898	1.664
0.5	1.078	1.060	1.048	2.292	1.897	1.663
0.8	1.067	1.052	1.042	2.277	1.889	1.655
0.9	1.049	1.038	1.031	2.256	1.879	1.645
0.95	1.013	1.010	1.008	2.223	1.872	1.637
0.99	0.875	0.895	0.909	2.012	1.835	1.628
0.995	0.832	0.857	0.877	1.551	1.813	1.626

The relative properties of the intermediate design and the supplemented one-per-stratum design for the first order autoregressive process are given in Table 5. The supplemented design for a $(K, 2K)$ specification has one supplement element in each group of $2K - 1$ strata.

The intermediate design has smaller variance of the mean for autoregressive coefficients less than 0.95, which would include many populations encountered in practice. The variance of the estimated variance is larger for the supplemented design for all coefficients in the table. Thus, for most populations encountered in practice the intermediate design is preferred to the supplemented design.

Acknowledgement

This research was partially supported by Research Agreements 68-7482-14-502 and 68-7482-17-009 with the Natural Resources Conservation Service of the United States Department of Agriculture.

REFERENCES

Bellhouse, D. R. (1984), “A review of optimal designs in survey sampling,” *Journal of Statistics*, 12, 53-65.

Breidt, F.J. (1995), “Markov chain designs for one-per-stratum sampling,” *Survey Methodology*, 21, 63-70.

Cochran, W.G. (1946), “Relative accuracy of systematic and stratified random samples for a certain class of populations,” *Annals of Mathematical Statistics*, 17, 164-177.

Fuller, W.A. (1970), “Sampling with random stratum boundaries,” *Journal of the Royal Statistical*, B32, 209-226.

Fuller, W.A. (2009), *Sampling Statistics*, Wiley, New York.

Papageorgiou, I. and Karakostas, K.X. (1998), “On optimal sampling designs for autocorrelated finite populations,” *Biometrika*, 85, 482-486.

Wolter, K.M. (2007),*Introduction to Variance Estimation* (2nd ed.), Springer, New York.