MEDIAN BALANCED SAMPLING DESIGN

Yan Liu, Ernst & Young LLP and Fritz Scheuren, Urban Institute
Fritz Scheuren, 1402 Ruffner Rd., Alexandria, VA 22302, USA, scheuren@aol.com

Key Words: Balanced Sampling Design, Stratified Sampling Design, Efficiency, Ratio Estimation, Regression Estimation

1. INTRODUCTION

Suppose that a stratified random sample of size n is drawn from a population of size N in order to estimate the population mean or total. In its "classical" implementation (Cochran 1977) the sampling units are selected at random within each stratum independently of selections within other strata. Therefore, there is no correlation across strata.

Of course, the stratified sample mean in this classical setting is an unbiased estimator of the population mean based on repeated sampling; nonetheless, in small or moderate samples the actual estimate from a selected sample could severely underestimate or overestimate the population mean, especially for a widely dispersed population. An extreme example of such a situation occurs when the sampled units are all on the lower side or all on the upper side within their respective stratum. These combinations of n units in our formulation would be called "nonpreferred" and could not arise. Here is how.

Assume that a covariate is available. Our proposed median balanced sampling design is a modification of the conventional stratified sampling design that employs this covariate. The key motivation is to reduce the sample space of all possible stratified samples to an identifiable subset of "preferred" samples -- while still keeping unchanged the selection probability of each unit in the population. Unlike the conventional stratified sampling design, our design introduces a negative correlation across the strata and therefore results in a smaller variance of the mean estimator. It also decreases the probabilities of selection for "nonpreferred" combinations, so that the mean estimate from a selected sample is "not too far" from the population mean. The proposed balanced sampling design will be shown to be an improvement over the conventional stratified sampling design.

Throughout this paper, the parameter to be estimated is the population mean. The main goal is to show the advantages of median balanced sampling over conventional stratified sampling. We focus initially on the setting of two strata and sampling with replacement; then, extend the results to more strata and without replacement situations.

Organizationally, this paper is divided into 6 sections. The current Section 1 is a general introduction. In Section 2, we describe two sampling procedures under the proposed median balanced sampling design. There we also compare the variance of the mean estimator under the median balanced sampling design with the variance under the conventional stratified sampling design. In Section 3, a sample estimator of the variance under the median balanced sampling is proposed. The stability of this estimator is also compared to its counterpart in conventional stratified sampling. Section 4 presents some asymptotic results and applications. First we show an efficiency gain when the strata are set by the covariate. Then we use Taylor's series expansions to prove that the median balanced sampling results in a smaller mean square error in ratio and regression estimation. In Section 5 we look at several simple but important extensions – e.g., to three and more strata and to the without replacement setting. Section 6 is a very brief summary. Proofs are available upon request or can be found in Liu (1999).

2. MEDIAN BALANCED SAMPLING DESIGN

To start the discussion, we will operate just like in most sampling texts (e.g., Cochran 1977) by assuming that we can actually stratify on the values of y, the variable of interest. Later in Section 4 we will discuss the more realistic case when the strata are based on a covariate x or on geographic units. Here, we only focus on sampling with replacement in the two strata setting – leaving to Section 5 the general case.

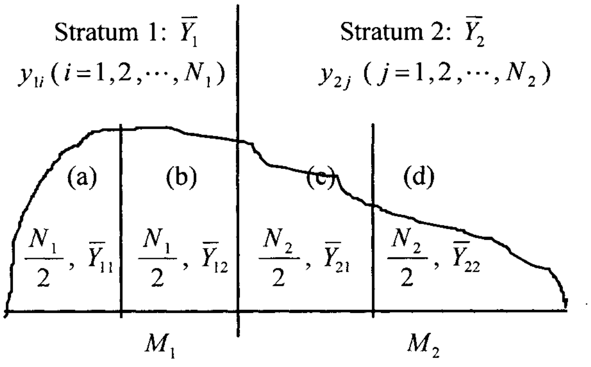
2.1 Two Sample Selection Methods

Divide the finite population of N units into 2 strata -- N_1 units in stratum 1 labeled as $y_{11}, y_{12}, \dots, y_{1N_1}$ and N_2 units in stratum 2 denoted as $y_{21}, y_{22}, \dots, y_{2N_2}$. Suppose further that N_1 and N_2 are even numbers; hence $N_1/2$ and $N_2/2$ are integers. Later, in Section 5, we will deal with the case of N_1 or N_2 odd.

Let M_1 be the median of stratum 1 and M_2 be the median of stratum 2. Without much loss of generality, assume further that no units in stratum 1 are tied to M_1 so that all N_1 units are separated by M_1 into two equal groups; in particular, that the first half of the units in stratum 1 are below M_1 and the second half of the units are above M_1. We will further presuppose that the same mild condition applies in stratum 2 with respect to its median M_2. This requirement can be relaxed, however. All that is really needed, in say the first stratum, is to divide the N_1 units into two equal groups such that the first $N_1/2$ units are not larger than the remaining $N_1/2$ units. This is true because our proposed median balanced sampling design does not depend on the actual value of the unit; it only depends on whether a unit belongs to the first half or the second half of the stratum it is in.

The proposed design requires the sample size n be even and that an equal sample size be allocated to each stratum. For example, the strata boundary can be chosen such that $N_1\sigma_1 = N_2\sigma_2$ approximately, where σ_1^2 and σ_2^2 are the variance in each stratum. Therefore the sample size of each stratum can be taken as equal under Neyman Optimum Allocation. Sometimes, of course, strata cannot be created this way, for example, when geographic units are used as strata. But the proposed design is still applicable as long as the sample size is equally allocated.

Figure 1. – Illustration of Notation in Two Strata Case


Stratum 1: $\bar{Y}_1$
$y_{1i}\,(i=1,2,\cdots,N_1)$
Stratum 2: $\bar{Y}_2$
$y_{2j}\;(j=1,2,\cdots,N_2)$
(a)
(b)
(c)
(d)
$\frac{N_1}{2}$, $\bar{Y}_{11}$
$\frac{N_1}{2}$, $\bar{Y}_{12}$
$\frac{N_2}{2}$, $\bar{Y}_{21}$
$\frac{N_2}{2}$, $\bar{Y}_{22}$
M_1
M_2

In the two strata setting just described, the population in each stratum can be equally divided by M_1 and M_2. A sample of size n is then composed of $m = n/2$ units from stratum 1 and m units from stratum 2. Thus, under the conventional stratified sampling with replacement, a sample of size $n = 2m$ is obtained by selecting m units randomly from N_1 units in stratum 1 with replacement and selecting another m units randomly from N_2 units in stratum 2 with the selections in the two strata being independent.

Figure 1 pictures the population we are sampling and in Cochran's notation the parameters in this two strata case. Notice, also, in figure 1 that for later use we have labeled the four parts of the population created by the combination of the stratum boundary and the two medians into the four cells: (a), (b), (c), and (d).

We now describe two selection methods that each employ median balanced sampling. As before the selections are all done with replacement.

Method 1: Unit Selection. -- A median balanced sample of $n = 2m$ can be obtained by first randomly selecting all m units with replacement from the N_1 units in stratum 1. The sample selection from stratum 2 would then be dependent on the outcome of the selections from stratum 1. Suppose the selection results of stratum 1 turn out to include m_1 units below the median M_1 with the other ($m - m_1$) above M_1; then we would randomly select m_1 units with replacement from stratum 2 units that are above M_2 and randomly select the remaining $(m - m_1)$ units with replacement from stratum 2 units that are below M_2. This sampling process is equivalent to first determining the value of m_1 using a binomial distribution $B(m, 1/2)$ and then randomly drawing m_1, $(m - m_1)$, $(m - m_1)$ and m_1 units independently from cells (a), (b), (c) and (d) respectively.

Method 2: Pair Selection. -- A median balanced sample of $n = 2m$ can also be obtained by drawing m pairs, one at a time. The process starts from randomly drawing a unit from the N_1 units of stratum 1, i.e., from $y_{1i}\,(i = 1,2,\cdots,N_1)$. If the selected unit is above M_1, then select a unit y_{2j} in stratum 2 from $N_2/2$ units that are below M_2. Conversely, if the selected unit is below M_1, then select a unit y_{2j} in stratum 2 from $N_2/2$ units that are above M_2. The selected two units or pair (y_{1i}, y_{2j}) can be considered as a replicate or subsample. To complete the selections, we would simply place the first pair back into the population, then draw a second pair according to the same sampling process – repeating the selection process until all m independent pairs are obtained.

An equivalent pair selection process is to first define the "preferred" combinations or pairs so that the combinations of y_{1i} and y_{2j} are either $\{(y_{1i}, y_{2j})$: $y_{1i} \in$(a) and $y_{2j} \in$(d)$\}$ or $\{(y_{1i}, y_{2j})$: $y_{1i} \in$(b) and $y_{2j} \in$(c)$\}$. There are a total of $N_1 N_2/2$ such preferred combinations. The sample would then consist of m pairs randomly selected from the $N_1 N_2/2$ preferred combinations with replacement. The remaining $N_1 N_2/2$ combinations – those with both units y_{1i} and y_{2j} above their medians or both below their medians -- are called "nonpreferred" combinations and not subject to sampling.

For each pair (y_{1i}, y_{2j}) selected from the population of $N_1 N_2/2$ preferred pairs, define a replicate estimator $\hat{\theta}_\alpha$ as

$$\hat{\theta}_\alpha = \frac{1}{N}\left(N_1 y_{1i} + N_2 y_{2j}\right), \qquad \alpha = 1, 2, \cdots, \frac{N_1 N_2}{2} \tag{2.1}$$

Each α corresponds to a preferred pair uniquely, i.e., $\alpha \leftrightarrow (i, j)$. Then the population consists of $N_1 N_2/2$ replicates and a sample from the pair selection method consists of a simple random sample of m replicates $\hat{\theta}_\alpha$ randomly selected from all $N_1 N_2/2$ replicates with replacement.

There are a total of $N_1 N_2$ pairs of (y_{1i}, y_{2j}) -- preferred and nonpreferred. If we define a replicate $\hat{t}_\alpha$ for each pair as

$$\hat{t}_\alpha = \frac{1}{N}\left(N_1 y_{1i} + N_2 y_{2j}\right), \qquad \alpha = 1, 2, \cdots, N_1 N_2 \tag{2.2}$$

then half of them are the same as $\hat{\theta}_\alpha$. A conventional stratified sample, comparable to the above pair selection, would mean selecting a simple random sample of m replicates $\hat{t}_\alpha$ from all $N_1 N_2$ pairs with replacement.

In what follows we will confine further attention just to Method 2, which we will designate as the **pair selection method.**

2.2 Sample Mean as an Unbiased Estimator

We know that under conventional stratified sampling designs (with replacement) the unit selection probabilities at each draw are

$$P(y_{1i} \text{ selected}) = 1/N_1, \quad \text{for } i = 1, 2, \ldots, N_1 \tag{2.3}$$

$$P(y_{2j} \text{ selected}) = 1/N_2, \quad \text{for } j = 1, 2, \ldots, N_2 \tag{2.4}$$

The unbiased estimator of the population mean $\bar{Y}$ is

$$\bar{y}_{st} = (N_1 \bar{y}_1 + N_2 \bar{y}_2)/N \tag{2.5}$$

where $\bar{y}_1 = \frac{1}{m}\sum_{i=1}^{m} y_{1i}$ and $\bar{y}_2 = \frac{1}{m}\sum_{j=1}^{m} y_{2j}$

Under the median balanced sampling design, the sample mean is

$$\bar{y}_{mb} = \frac{1}{N}(N_1 \bar{y}_1 + N_2 \bar{y}_2) \tag{2.6}$$

which has the same functional form as $\bar{y}_{st}$. The difference between (2.5) and (2.6) is that $\bar{y}_1$ and $\bar{y}_2$ are independent in (2.5) and they are dependent in (2.6). The unconditional selection probabilities are still the same as those in conventional stratified sampling design, as defined in (2.3) and (2.4). It is straightforward to show that for $y_{2j} < M_2$, $j = 1, 2, \ldots, N_2/2$,

$$\begin{aligned} & P(y_{2j} \text{ selected}) \\ & = P\{y_{2j} \text{ below } M_2 | \text{ First draw is above } M_1\} \\ & \quad \times P\{\text{First draw is above } M_1\} \\ & = (2/N_2) \times (1/2) = 1/N_2 . \end{aligned}$$

Similarly,

for $y_{2j} > M_2$, $j = \left(\frac{N_2}{2}+1\right), \left(\frac{N_2}{2}+2\right), \ldots, N_2$,

$P(y_{2j} \text{ selected}) = 1/N_2$.

Therefore, it follows that $\bar{y}_{mb}$ is an unbiased estimator of $\bar{Y}$.

2.3 Efficiency Gain Due to Median Balanced Sampling

We are familiar with the variance of $\bar{y}_{st}$:

$$Var(\bar{y}_{st}) = \frac{N_1^2}{N^2} Var(\bar{y}_1) + \frac{N_2^2}{N^2} Var(\bar{y}_2)$$

where

$$Var(\bar{y}_1) = \frac{1}{m}\left(\frac{1}{N_1}\sum_{i=1}^{N_1}(y_{1i} - \bar{Y}_1)^2\right)$$

and

$$Var(\bar{y}_2) = \frac{1}{m}\left(\frac{1}{N_2}\sum_{j=1}^{N_2}(y_{2j} - \bar{Y}_2)^2\right)$$

Now, it can be shown that under both selection methods of median balanced sampling,

$$Var(\bar{y}_{mb}) = Var(\bar{y}_{st}) + 2\left(N_1 N_2/N^2\right)Cov_{mb}(\bar{y}_1, \bar{y}_2) \tag{2.7}$$

where

$$\begin{aligned} & Cov_{mb}(\bar{y}_1, \bar{y}_2) \\ & = \frac{1}{2m}\left[(\bar{Y}_{11} - \bar{Y}_1)(\bar{Y}_{22} - \bar{Y}_2) + (\bar{Y}_{12} - \bar{Y}_1)(\bar{Y}_{21} - \bar{Y}_2)\right] \end{aligned}$$

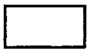
$$=\frac{1}{4m}\ (\bar{Y}_{11}-\bar{Y}_{12})(\bar{Y}_{22}-\bar{Y}_{21})<0 \qquad (2.8)$$

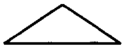
[See proof (A)]. Note the parameters $\bar{Y}_{11}$, $\bar{Y}_{12}$, $\bar{Y}_{21}$ and $\bar{Y}_{22}$ are the four cell population means pictured in figure 1.

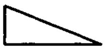
The implication from the expression (2.8) above is that $Var(\bar{y}_{mb}) < Var(\bar{y}_{st})$. So, we conclude that the median balanced sampling approach is superior to the regular stratified approach because of the negative covariance introduced between strata. Note that $Cov_{st}(\bar{y}_1,\bar{y}_2)=0$ under the conventional stratified sampling design because of the independence of sample selection between two strata.

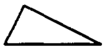
To see how much efficiency can be gained when balanced sampling is used we bring out four important special situations of potentially practical interest. These distributions (rectangular, symmetric triangular, right triangular, and triangular skewed to the right), when taken together, can often describe an empirical population distribution fairly well, despite the simplicity that each has alone.

To give a specific idea of potential efficiency gains due to balancing on the median, we set the boundary of two strata using Neyman allocation approximately and calculate the ratio $\frac{Var(\bar{y}_{mb})}{Var(\bar{y}_{st})}$. A summary of the remarkable efficiency gains in terms of $\frac{Var(\bar{y}_{mb})}{Var(\bar{y}_{st})}$ is presented under each of the distributions:


0.25


0.43


0.33


0.33 to 0.43

3. SAMPLE ESTIMATE OF VARIANCE AND ITS STABILITY

We have shown that $Var(\bar{y}_{mb})<Var(\bar{y}_{st})$. This is important in choosing a good survey design but not enough by itself. At the analysis stage where only sample data are available, the variance of $\bar{y}_{mb}$ will not be known; it must be estimated from sample data. So a stable and easy-to-calculate sample estimator is needed. To discuss this issue we will employ the notation $v(\bar{y}_{mb})$ for the sample estimator of $Var(\bar{y}_{mb})$; and $v(\bar{y}_{st})$ for the sample estimator of $Var(\bar{y}_{st})$.

Now $v(\bar{y}_{st})$ is well known from conventional stratified sampling to be

$$v(\bar{y}_{st})=\frac{N_1^{\ 2}}{N^2}v(\bar{y}_1)+\frac{N_2^{\ 2}}{N^2}v(\bar{y}_2)$$

where

$$v(\bar{y}_1)=\frac{1}{m(m-1)}\sum_{i=1}^{m}(y_{1i}-\bar{y}_1)^2$$

$$v(\bar{y}_2)=\frac{1}{m(m-1)}\sum_{j=1}^{m}(y_{2j}-\bar{y}_2)^2$$

In this section, we propose an unbiased estimator $v(\bar{y}_{mb})$ for the sample obtained by the **pair selection method** (Method 2), and show that the stability of $v(\bar{y}_{mb})$ from the balanced sampling design is at least as good as $v(\bar{y}_{st})$ from the stratified random sampling design. For another proposed unbiased estimator and its stability for the sample obtained by **the unit selection method** (Method 1), the same conclusion can be reached and the details can be found in Liu (1999).

Suppose the sample is obtained by the pair selection method, then the sample consists of m replicates $\hat{\theta}_\alpha$ defined by expression (2.1), $\alpha \leftrightarrow (i,j)$ and $\alpha=1,2,\cdots,m$. Hence

$$\bar{y}_{mb}=\frac{1}{m}\sum_{\substack{\alpha=1\\ \alpha\leftrightarrow(i,j)}}^{m}\hat{\theta}_\alpha=\frac{1}{N}\left(N_1\bar{y}_1+N_2\bar{y}_2\right)$$

The proposed variance estimator of $\bar{y}_{mb}$ is

$$v(\bar{y}_{mb})=\frac{1}{m(m-1)}\sum_{\alpha=1}^{m}\ (\hat{\theta}_\alpha-\bar{y}_{mb})^2$$

$$=\frac{1}{N^2}\left(N_1^{\ 2}v(\bar{y}_1)+N_2^{\ 2}v(\bar{y}_2)+2N_1N_2\ \mathrm{cov}_{mb}(\bar{y}_1,\bar{y}_2)\right)$$

where

$$\mathrm{cov}_{mb}(\bar{y}_1,\bar{y}_2)\ =\ \frac{1}{m(m-1)}\ \sum_{\substack{\alpha=1\\ \alpha\leftrightarrow(i,j)}}^{m}(y_{1i}-\bar{y}_1)(y_{2j}-\bar{y}_2)$$

Comparing $v(\bar{y}_{st})$ and $v(\bar{y}_{mb})$ the only differences are the covariance terms, with 0 for $v(\bar{y}_{st})$, a nonzero term $\mathrm{cov}_{mb}(\bar{y}_1,\bar{y}_2)$ for the paired selection method. Furthermore, both $v(\bar{y}_{st})$ and $v(\bar{y}_{mb})$ are known to be unbiased.

The functional form of $Var\{v(\bar{y}_{mb})\}$ and $Var\{v(\bar{y}_{st})\}$ can be derived. But it is complicated to compare $Var\{v(\bar{y}_{mb})\}$ and $Var\{v(\bar{y}_{st})\}$ through their analytic forms. So we will next show a better stability of the proposed variance estimators by looking at the ratio of $\frac{Var\{v(\bar{y}_{mb})\}}{Var\{v(\bar{y}_{st})\}}$ under the four theoretical distributions. The number of replicates m has a minor impact on the ratio, so we show our results for two specific values of m. The following summary of the simulation results shows a smaller $Var\{v(\bar{y}_{mb})\}$ compared to $Var\{v(\bar{y}_{st})\}$:

	▭	△	◺	△
m=15	0.20	0.35	0.23	0.23 to 0.35
m=500	0.22	0.37	0.24	0.24 to 0.37

From above results along with the results in Section 2.3, the median balanced sampling, in special cases at least, has not only a higher precision at the design stage, but also its sample estimator of the variance has a better stability at the estimation stage.

4. SOME RESULTS AND APPLICATIONS IN THE LARGE SAMPLE SITUATION

In practice, the limiting results when the sample size $n=2m \rightarrow \infty$ are often used as approximations when the sample size is large. For example, confidence intervals are based on the normal distribution. The accuracy of an asymptotic approximation is difficult to evaluate, but measures, such as asymptotic variance can be compared. In this connection, central limit theorems are a powerful tool in deriving properties of estimators without knowing the distribution of the original data. Perhaps most importantly, when the sample size is reasonably large, limiting results can be applied in the situations that exact results are impossible or too complicated to be of practical use.

4.1 Gain in Efficiency When Stratifying on a Covariate

So far we have assumed that the strata are set using the values of y -- the variable of interest. We have also supposed that the sample selection also depends on the values of y. In other words, that the preferred combinations are determined by at least knowing which units of y are below or above strata medians.

As we have shown if we could stratify by the values of y, the median balanced sample has a smaller variance, regardless of the sample size. In practice, of course, we rarely know the values of y in the population; instead, at best, we may have knowledge of values of a covariate x -- a variable we hope is closely related to y.

In this section we explore the benefits on estimating $\bar{Y}$ with a median balanced design based on x. It is clear that we cannot specify the preferred combinations of y through $(x_1, x_2, \cdots, x_N)$; we can only specify the preferred combinations for x itself. But if x and y are closely related, we expect that a median balanced sampling design based on x to produce a better estimator of $\bar{Y}$ than the regular stratified random sampling design does – *provided the sample size is large enough.* In examining this assertion, we use the fact that the relationship between sample means of y and x is asymptotically bivariate normal to show that $Var(\bar{y}_{mb}) \leq Var(\bar{y}_{st})$ [proof (B)].

If the strata are set by geographic units, we can still balance on the covariate x as long as the sample size is the same in each stratum. The conclusion that $Var(\bar{y}_{mb}) \leq Var(\bar{y}_{st})$ still holds under large sample assumption.

4.2 Applications to Ratio and Regression Estimation

Naturally, we expected that the median balanced sampling would result in a smaller bias and a smaller variance for these estimators; and, indeed, this turns out to be the case, under fairly general conditions. To examine the performance of ratio and regression estimators under the alternative sampling designs being considered, we will be using Taylor's series expansions, with all the necessary regularity conditions implied.

The combined ratio estimators of $R = \bar{Y}/\bar{X}$ and the population mean $\bar{Y}$ from classical stratified random sampling are

$$\hat{R}_{st} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \quad \text{and} \quad \hat{\bar{Y}}_{R-st} = \frac{\bar{y}_{st}}{\bar{x}_{st}}\bar{X} = \hat{R}_{st}\bar{X}$$

respectively.

The corresponding estimators from median balanced sampling are

$$\hat{R}_{mb} = \frac{\bar{y}_{mb}}{\bar{x}_{mb}} \quad \text{and} \quad \hat{\bar{Y}}_{R-mb} = \frac{\bar{y}_{mb}}{\bar{x}_{mb}}\bar{X} = \hat{R}_{mb}\bar{X}$$

The combined ratio estimator is in general biased for both the sample designs under consideration. To quantify the bias, at least approximately, we employ a Taylor series approach as is commonly done in ratio and regression estimation. The first order approximation is believed to yield satisfactory results in large complex survey.

When stratifying and median balancing are done on x, to a first order approximation [proof (C)],

$$MSE(\hat{R}_{mb}) < MSE(\hat{R}_{st})$$

or equivalently,

$$MSE(\hat{\bar{Y}}_{R-mb}) < MSE(\hat{\bar{Y}}_{R-st}).$$

The combined linear regression estimators of the population mean $\overline{Y}$ from stratified sampling and balanced sampling are

$$\overline{y}_{lr-st} = \overline{y}_{st} + \hat{\beta}\ (\overline{X} - \overline{x}_{st})$$

$$\overline{y}_{lr-mb} = \overline{y}_{mb} + \hat{\beta}\ (\overline{X} - \overline{x}_{mb})$$

where $\hat{\beta}$ is the estimate of regression coefficient from the sample. Here, the subscript *lr* stands for linear regression. $\overline{y}_{lr-st}$ and $\overline{y}_{lr-mb}$ are not unbiased estimators of $\overline{Y}$, but the biases under both designs are negligible. Their first order approximations of MSE [Proof (D)] are

$$MSE(\overline{y}_{lr-st}) = Var(\overline{y}_{st}) + \beta^2 Var(\overline{x}_{st}) - 2\beta Cov(\overline{x}_{st}, \overline{y}_{st})$$

$$MSE(\overline{y}_{lr-mb}) = Var(\overline{y}_{mb}) + \beta^2 Var(\overline{x}_{mb}) - 2\beta Cov(\overline{x}_{mb}, \overline{y}_{mb})$$

The comparison of above two equations leads to that $MSE(\overline{y}_{lr-mb}) < MSE(\overline{y}_{lr-st})$.

5. EXTENTIONS

We have avoided certain details so far to derive the results under the setting of (a) selection with replacement, (b) population size N a convenient multiple of the number of strata, and (c) two strata. As we will see, all these constraints can be relaxed.

Sampling Without Replacement -- In practice, sampling without replacement is often used. The estimated variance under without replacement sampling is always smaller than that under with replacement by introducing some form of finite population correction (pfc). However, the estimated variance for with replacement is often used even in the without replacement sampling situation because of its neater functional form. The effect is a slight overestimation when the sampling proportion is low and the fpc can be ignored. But fpc can be an important factor in some circumstances and so without replacement median balanced design must be taken up directly. It can be derived that under sampling without replacement (see Liu (1999)),

$$Var(\overline{y}_{mb}) = Var(\overline{y}_{st}) + \frac{2N_1N_2}{N^2} Cov_{mb}(\overline{y}_1, \overline{y}_2) + e$$

where

$$Cov_{mb}(\overline{y}_1, \overline{y}_2) = \frac{1}{4m}\frac{N_1 - m}{N_1 - 1}(\overline{Y}_{11} - \overline{Y}_{12})(\overline{Y}_{22} - \overline{Y}_{21}) < 0$$

and

$$e = o(1/N) \leq 0$$

Therefore, $Var(\overline{y}_{mb}) < Var(\overline{y}_{st})$.

Population Size Not Conveniently Divisible -- When population size N is odd, a minor adjustment of the sampling process ensures the same first and second order selection probabilities. One operationally straightforward adjustment is provided here. The strata boundary can be chosen such that N_1 is odd and N_2 is even – with equal sample sizes m in both strata. The sampling process follows what is described in Section 2.1 with one exception. If the middle unit in stratum 1 is selected, the matching unit from stratum 2 will be selected from all N_2 units without balancing on the median of stratum 2 as we would do for non-middle units. The variance of the sample mean is very close to the general result in Section 2.3. See Liu (1999) for details.

More than Two Strata -- In generalizing the results obtained so far to more than two strata, we first consider the three strata case. Similar to the two strata case, divide the finite population into 3 strata such that a sample of size n (assume n is an integer multiple of 3) allocated to each of the three strata with $m = n/3$ units. Assume N_1, N_2 and N_3, the population sizes of 3 strata, be even numbers. Define z_l as

$$z_l = \frac{N_1 y_{1i} + N_2 y_{2j}}{N_1 + N_2}, \qquad l = 1, 2, \cdots, N_1 N_2$$

and let M_z be the median of the z_l. We first select units from strata 1 and 2 using balanced sampling method, then conditional on whether the corresponding value of z_l is above or below M_z, we select units from stratum 3.

This selection pattern can be used for more than three strata. When the number of strata is even, we can use the two strata balancing technique for every two adjacent strata.

6. BRIEF SUMMARY

In summary, we believe we have provided a better estimator than $\overline{y}_{st}$ of the population mean through the use of our proposed balanced sampling designs and the corresponding proposed estimators such as $\overline{y}_{mb}$ and $v(\overline{y}_{mb})$. Median balanced sampling design thus can be seen as generally an improvement over conventional stratified sampling design. We have had considerable experience in its use and are pleased with how well practice bears out theory.

Proofs and references are available upon request.