Combination of One-Per-Stratum and Two-Per-Stratum Design

Xiaofei Zhang
Joint with Dr. Wayne Fuller

Center for Survey Statistics and Methodology
Iowa State university

October 19, 2016
Outline

1. Introduction
2. Intermediate design procedures
3. Design-based estimator
4. Intermediate design for autoregressive population
5. Intermediate design for area data
6. Conclusion and discussion
Sampling method
The rules and procedures by which some elements of the population are included in the sample. For example: simple random sampling, stratified sampling, systematic sampling.

Estimator
A function of sample data to calculate an estimate for parameter of interest. For example: Horvitz-Thompson estimator. A good estimator: unbiased, small variance.
Stratified sampling

- The elements of a finite population are divided into $H$ groups, indexed by $h = 1, 2, \cdots, H$, called strata.
- Within each stratum, samples are drawn independently.

Advantages

- representative sample
- efficient relative to simple random sample
- convenient to conduct
Stratified Sampling

- **Stratified sampling estimator**
  - \( \bar{y}_h = n_h^{-1} \sum_{i \in A_h} y_{hi} \) and \( A_h \) is the set of indices for sample in stratum \( h \).

- \( \bar{y}_{st} = \sum_{h=1}^{H} N_h^{-1} N_h \bar{y}_h \) and \( N \) is population size, \( N_h \) is the size of stratum \( h \).

- \( \hat{V}\{(\bar{y}_{st} - \bar{y}_N) | \mathcal{F}\} = \sum_{h=1}^{H} (N_h^{-1} N_h)^2 (1 - N_h^{-1} n_h) n_h^{-1} s_h^2 \),

- where \( s_h^2 = (n_h - 1)^{-1} \sum_{i \in A_h} (y_{hi} - \bar{y}_h)^2 \).

- No variance estimator when \( n_h = 1 \).
Figure 1: Stratified sampling, one-per-stratum

Figure 2: Stratified sampling, two-per-stratum

Figure 3: Combination of one-per-stratum and two-per-stratum
Let the population be divided into $2H$ equal sized sets then let the sets be formed into $H$ pairs of sets. Let the sets be identified by $h_j, h = 1, 2, \ldots, H, j = 1, 2$.

Let the pairs of sets be formed into groups of size $K$ and sample size 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.
Intermediate design procedures

- $K = 4$

![Figure 4: Intermediate design, $K = 4$](image)

- $K = 2$

![Figure 5: Intermediate design, $K = 2$](image)
Assume observations satisfy

\[ y_{hjt} = \mu_h + b_{hj} + e_{hjt}, \quad b_{hj} \sim (0, \sigma_b^2), \quad e_{hjt} \sim (0, \sigma_e^2). \]

Without loss of generality, let the simple random sample of size two be selected from the first stratum, then the population mean estimator is

\[
\bar{y}_{2K} = (2K)^{-1} \left[ y_{11,1} + y_{11,2} + \sum_{h=2}^{K} (y_{h1} + y_{h2}) \right].
\]
Derivation of variance

- Ignore finite population correction

\[ MSE\{\bar{y}_{2K}|b_h\} = K^{-2}[0.5\sigma_e^2 + (0.5(b_{11} - b_{12}))^2 + (K - 1)(0.5\sigma_e^2)] \]

and

\[ E\{MSE\{\bar{y}_{2K}|b_h\}\} = 0.5(K^{-1}\sigma_e^2 + K^{-2}\sigma_b^2). \]

Then the standardized variance is

\[ 2KV\{\bar{y}_{2K}\} = \sigma_e^2 + K^{-1}\sigma_b^2. \]

- The variance of the total sample mean is

\[ V\{\bar{y}_{tot}\} = 2KV\{\bar{y}_{2K}\}/2H. \]
Variance estimator

- Let the estimated between mean square be
  \[ B\hat{W}S = 0.5(K - 1)^{-1} \sum_{h=2}^{K} (y_{h1} - y_{h2})^2 \]
  and the estimated within mean square be
  \[ W\hat{M}S = 0.5(y_{111} - y_{112})^2. \]

- Then
  \[ E\{B\hat{W}S\} = \sigma_e^2 + \sigma_b^2, \quad E\{W\hat{M}S\} = \sigma_e^2. \]

- A nonnegative unbiased estimator of the standardized variance
  \[ 2K\hat{V}\{\bar{y}_{2K}\} = K^{-1}B\hat{M}S + (1 - K^{-1})W\hat{M}S. \]
Variance of the variance estimator

- Assume \( b_{hj} \) and \( e_{hjt} \) are normally distributed,
  \[
  V\{B\hat{M}S\} = 2(K - 1)^{-1}(\sigma_e^2 + \sigma_b^2)^2, \quad V\{W\hat{M}S\} = 2\sigma_e^4.
  \]

- Standardized variance of standardized estimated variance is
  \[
  KV\{2K\hat{V}(\bar{y}_2K)\} = KV\{K^{-1}B\hat{M}S + (1 - K^{-1})W\hat{M}S\} \]
  \[
  = 2K^{-1}(K - 1)^{-1}(\sigma_e^2 + \sigma_b^2)^2 + 2K(1 - K^{-1})^2\sigma_e^4.
  \]

- The variance of the variance for total sample is
  \[
  V\{\hat{V}(\bar{y}_{tot})\} = KV\{2K\hat{V}(\bar{y}_2K)\}/4H^3.
  \]
For $\sigma_b^2 / \sigma_e^2 \leq 0.58$, the standardized variance of the variance increases as $K$ increases.
Consider the population of $2H$ strata with stratum size $M$. Express the observations as

$$y_{hjt} = \bar{y}_h + B_{hj} + a_{hjt}$$

where

$$\bar{y}_h = \sum_{j=1}^{2} \sum_{t=1}^{M} y_{hjt} / (2M), \quad \bar{y}_{hj} = \sum_{t=1}^{M} y_{hjt} / M, \quad B_{hj} = \bar{y}_{hj} - \bar{y}_h$$

and

$$a_{hjt} = y_{hjt} - \bar{y}_{hj}.$$ 

The sample mean is

$$\bar{y}_{2K} = (2K)^{-1} \left[ y_{11,1} + y_{11,2} + \sum_{h=2}^{K} (y_{h1} + y_{h2}) \right].$$
Finite Population

- Ignore finite population correction

\[
MSE\{\bar{y}_{2K} | \mathcal{F}\} = \frac{1}{4K^2} [(B_{11} - B_{12})^2 + 2S_{11}^2 + 2 \sum_{h=2}^{K} S_{hj}^2],
\]

where \( S_{hj}^2 = \sum_{t=1}^{M} (y_{hjt} - \bar{y}_{hj})^2 / (M - 1) \).

- As the naming of pair one was arbitrary, it follows that

\[
MSE(\bar{y}_{2K} | \mathcal{F}) = \frac{1}{4K^2} (2S_B^2 + 2KS_w^2) = \frac{1}{2K^2} (S_B^2 + KS_w^2),
\]

where \( S_B^2 = \sum_{h=1}^{K} (B_{h1} - B_{h2})^2 / 2K = \sum_{h=1}^{K} (\bar{y}_{h1} - \bar{y}_{h2})^2 / 2K \) and
\( S_w^2 = \sum_{h=1}^{K} \sum_{j=1}^{2} S_{hj}^2 / 2K. \)
Finite Population

- The standardized variance conditional on the fixed finite population is

\[ 2KV\{\bar{y}_{2K}|\mathcal{F}\} = S^2_w + K^{-1}S^2_B. \]

Table 1: Finite Population ANOVA Table for 2K Strata

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum of squares</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata pairs</td>
<td>$K - 1$</td>
<td>$\sum_{h=1}^{K} \sum_{j=1}^{2} \sum_{t=1}^{M} (\bar{y}_h - \bar{y}_N)^2$</td>
<td></td>
</tr>
<tr>
<td>Between pairs</td>
<td>$K$</td>
<td>$\sum_{h=1}^{K} \sum_{j=1}^{2} \sum_{t=1}^{M} (\bar{y}_{hj} - \bar{y}_h)^2$</td>
<td>$MS^2_B$</td>
</tr>
<tr>
<td>Within strata</td>
<td>$N - 2K$</td>
<td>$\sum_{h=1}^{K} \sum_{j=1}^{2} \sum_{t=1}^{M} (y_{hjt} - \bar{y}_{hj})^2$</td>
<td>$S^2_w$</td>
</tr>
<tr>
<td>Within pairs</td>
<td>$N - K$</td>
<td>$\sum_{h=1}^{K} \sum_{j=1}^{2} \sum_{t=1}^{M} (y_{hjt} - \bar{y}_h)^2$</td>
<td>$S^2_p$</td>
</tr>
</tbody>
</table>
Let

\[
\hat{BWS} = 0.5(K - 1)^{-1} \sum_{h=2}^{K} (y_{h1} - y_{h2})^2, \quad \hat{WMS} = 0.5(y_{111} - y_{112})^2
\]

\[
E\{\hat{BWS}|\mathcal{F}\} = S_B^2 + S_w^2, \quad E\{\hat{WMS}|\mathcal{F}\} = S_w^2
\]

A nonnegative unbiased estimator of the variance

\[
2K\hat{V}\{\bar{y}_{2K}|\mathcal{F}\} = K^{-1}B\hat{MS} + (1 - K^{-1})W\hat{MS}.
\]
Autoregressive population

- An autoregressive process

\[ y_t = \rho y_{t-1} + e_t, \quad t = 1, \ldots, N, \]

where \( e_1, \ldots, e_N \) are IID \( N(0, \sigma^2) \).

- \[ E\{y_t\} = 0, \quad V\{y_t\} = (1 - \rho^2)^{-1}\sigma^2, \]

and

\[ Cov(y_{t+h}, y_t) = (1 - \rho^2)^{-1}\rho^{|h|}\sigma^2. \]
Expectation of the estimated variance

\[
E\{2K\hat{V}(\bar{y}_{2K} \mid \mathcal{F})\} = K^{-1} E\{B\hat{M}S\} + (1 - K^{-1}) E\{W\hat{M}S\}
\]

\[
= \frac{1}{K} \frac{\sigma^2}{(1 - \rho^2)} (1 - \frac{g_M(\rho)}{M^2}) + (1 - \frac{1}{K}) \frac{\sigma^2}{(1 - \rho^2)} (1 - \frac{2f_M(\rho)}{(M^2 - M)})
\]

where

\[
g_M(\rho) = \sum_{i=1}^{M-1} i \rho^i + \sum_{i=1}^{M-1} i \rho^{2M-i} + M \rho^M = (\rho - 2\rho^{M+1} + \rho^{2M+1})/(1 - \rho)^2,
\]

\[
f_M(\rho) = \sum_{i=1}^{M-1} (M - i) \rho^i = [(M - 1) \rho - M \rho^2 + \rho^{M+1}]/(1 - \rho)^2.
\]
Variance structure

- For $M = 3$

\[
V \left( \begin{pmatrix} y_{h1} \\ y_{h2} \end{pmatrix} \right) = \frac{\sigma^2}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \begin{pmatrix} \rho^3 & \rho^4 & \rho^5 \\ \rho^2 & \rho^3 & \rho^4 \\ \rho & \rho^2 & \rho^3 \end{pmatrix} \begin{pmatrix} \rho^4 & \rho^3 & \rho^2 \\ \rho^3 & \rho^2 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}.
\]

- $g_M(\rho)$ is the sum of entries in the square part,
- $f_M(\rho)$ is the sum of entries in the triangle part.
Variance of the estimated variance

\[ V\{2K \hat{V}(\bar{y}_2K | \mathcal{F})\} = V\{K^{-1}BMS + (1 - K^{-1})WMS\} \]

\[ = \frac{-2}{(1 - \rho^2)^2 K^2(K - 1)} \left( 1 - 2g_M(\rho)/M^2 + g_M(\rho^2)/M^2 \right) + \]

\[ \frac{2(1 - \rho^2)^2}{K^2} \left( 1 - 4f_M(\rho)/(M^2 - M) + 2f_M(\rho^2)/(M^2 - M) \right) + \]

\[ \frac{2}{1 - \rho^2} \frac{K - 2}{2K^3(K - 1)^2} f_k(\rho^4M) \times \]

\[ \left[ \frac{2}{M^2g_M(\rho^2)(1 + \rho^{-2M})^2} - \frac{8}{M^3(1 + \rho^{-2M})\rho^{2-M}} \left( \frac{1 - \rho^M}{1 - \rho} \right)^2 - \frac{1 - \rho^2M}{1 - \rho^2} + \frac{8}{M^4g_M(\rho)\rho^{-2M}} \right] + \]

\[ \frac{2}{1 - \rho^2} \frac{K - 2}{K^3} f_k(\rho^4M) \times \]

\[ \left[ \frac{2}{M^2g_M(\rho^2)(1 + \rho^{-2M})^2} - \frac{2}{M^3(1 + \rho^{-M})^2(1 + \rho^{-2M})\rho^2} \left( \frac{1 - \rho^M}{1 - \rho} \right)^2 - \frac{1 - \rho^2M}{1 - \rho^2} + \frac{4}{M^4g_M(\rho)(\rho^{-M} + \rho^{-3M})} \right] + \]

\[ \frac{1}{(1 - \rho^2)^2 K^2(K - 1)} (g_M(\rho^2) - g_M(\rho)^2/M^2)/(M^2 - 1) + \]

\[ \frac{1}{(1 - \rho^2)^2 (1 - \frac{1}{K})^2 (2f_M(\rho^2) - (2f_M(\rho))^2/(M^2 - M))/(M^2 - M - 1). \]
### Table 2: Variance relative to variance for $K = 2$ (M=100, $\sigma^2 = 1$)

<table>
<thead>
<tr>
<th>$K$</th>
<th>$\rho$</th>
<th>0.2</th>
<th>0.6</th>
<th>0.9</th>
<th>0.95</th>
<th>0.98</th>
<th>0.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.999</td>
<td>0.995</td>
<td>0.972</td>
<td>0.946</td>
<td>0.900</td>
<td>0.854</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.999</td>
<td>0.993</td>
<td>0.958</td>
<td>0.918</td>
<td>0.850</td>
<td>0.782</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.999</td>
<td>0.991</td>
<td>0.949</td>
<td>0.902</td>
<td>0.820</td>
<td>0.738</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.998</td>
<td>0.988</td>
<td>0.932</td>
<td>0.869</td>
<td>0.760</td>
<td>0.651</td>
</tr>
<tr>
<td>str1</td>
<td></td>
<td>0.998</td>
<td>0.985</td>
<td>0.915</td>
<td>0.837</td>
<td>0.700</td>
<td>0.563</td>
</tr>
</tbody>
</table>
Table 3: Variance of estimated variance relative to value for $K = 2$ (M=100, $\sigma^2 = 1$)

<table>
<thead>
<tr>
<th>K</th>
<th>$\rho$</th>
<th>0.2</th>
<th>0.6</th>
<th>0.9</th>
<th>0.95</th>
<th>0.98</th>
<th>0.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>1.495</td>
<td>1.475</td>
<td>1.366</td>
<td>1.236</td>
<td>0.996</td>
<td>0.754</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.324</td>
<td>2.287</td>
<td>2.085</td>
<td>1.841</td>
<td>1.394</td>
<td>0.943</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3.236</td>
<td>3.182</td>
<td>2.888</td>
<td>2.534</td>
<td>1.881</td>
<td>1.225</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>8.076</td>
<td>7.937</td>
<td>7.181</td>
<td>6.269</td>
<td>4.581</td>
<td>2.896</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6: Standardized variance vs standardized variance of variance
Results

- If $E(S_B^2)/E(S_w^2) \leq 0.87$, the standardized variance of the variance monotonously increases as $K$ increases.
- $E(S_B^2)/E(S_w^2) = 0.87$, for
  $(\rho, M) = (0.86, 10), (0.94, 30), (0.96, 50), (0.98, 100)$. 

![Graph showing the relationship between $\rho$ and the ratio of variances for different $M$ values.](image)
$df = \frac{2V^2(\bar{y})}{V\{\hat{V}(\bar{y})\}}$ for normal.

Table 4: Degree of freedom (M=100, $\sigma^2 = 1$)

<table>
<thead>
<tr>
<th>K</th>
<th>$\rho$</th>
<th>0.2</th>
<th>0.6</th>
<th>0.9</th>
<th>0.95</th>
<th>0.98</th>
<th>0.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49.97</td>
<td>49.59</td>
<td>47.24</td>
<td>44.81</td>
<td>40.29</td>
<td>34.21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>33.37</td>
<td>33.29</td>
<td>32.66</td>
<td>32.44</td>
<td>32.78</td>
<td>33.13</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>15.39</td>
<td>15.31</td>
<td>14.74</td>
<td>14.39</td>
<td>14.41</td>
<td>15.21</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6.16</td>
<td>6.10</td>
<td>5.72</td>
<td>5.40</td>
<td>5.08</td>
<td>5.00</td>
<td></td>
</tr>
</tbody>
</table>
• $U$: 100 $\times$ 100 equally spaced grid points in $[0, 1] \times [0, 1]$.

• Observation $z_s$ at location $s \in U$

\[ z_s = 6 + X_s + \epsilon_s, \forall s \in U \]

where $s = (s_1, s_2)$, $X_s$ is the intrinsic stationary spatial process with the exponential semi-variogram whose nugget, sill and range are 0, 1 and 0.1 and $\epsilon_s \sim N(0, \beta|6 + X_s|)$ is the measurement error.
Sample size $n = 400$.

Five method:
  - intermediate design $K = 2, 5, 10$
  - stratified one-per-stratum
  - systematic sampling
  - local pivotal method (LPM)
  - Generalized random tessellation stratified (GRTS).
LPM and GRTS

- LPM and GRTS are spatially balanced sampling designs.
- LPM (Grafström and Tillé, 2012)
  Update the inclusion probabilities to give small joint inclusion probabilities for near by units.
- GRTS (Stevens and Olsen, 2004)
  Systematic sampling after mapping two-dimensional space into one-dimensional space.
Population example
Sample example
Sample example

Systematic sampling

Intermediate sampling
Variance estimation

- **LPM and GRTS**
  Estimate variance as the variance of one-per-stratum stratified sampling.

\[
\hat{V}\{\bar{y}_{st} - \bar{y}_N}\} = \sum_{h=1}^{H} (n_h - 1) s_h^2 / [M \sum_{h=1}^{H} (n_h - 1)]
\]

- **Systematic sampling (Wolter, 2007)**

\[
\hat{V}(\bar{y}_s - \bar{y}_N) = (1/M) \sum_{j=2}^{M} (y_j - y_{j-1})^2 / 2(M - 1)
\]

- **Intermediate design**

\[
\hat{V}\{\bar{y}_{2K}|F\} = (2K)^{-1}[K^{-1} B \hat{M} S + (1 - K^{-1}) W \hat{M} S]
\]
Sample example
Results

![Diagram showing variance for different sample methods]

- MC variance
- Estimated variance
- Square root of var of var

Sample method:
- lpm
- grts
- sys
- k=2
- k=5
- k=10
- str1

Variance:
- 0.0000
- 0.0005
- 0.0010
- 0.0015
- 0.0020
Conclusion and discussion

Conclusion
- We present a sampling design method with unbiased population mean estimator and unbiased variance estimator.
- Efficiency between one-per-stratum and two-per-stratum performed well for simulated spatial population.

Extension
- Unequal probability.
- Auxiliary information.
Fuller, W., Berg, E., and Zhang, X. (2016), ”Combination of One-Per-Stratum and Two-Per-Stratum Design”, Iowa State University unpublished paper.


Thank you.