Survey Designs with Small Stratum Sample Sizes

Xiaofei Zhang
Joint with Dr. Wayne Fuller

Center for Survey Statistics and Methodology
Iowa State University

March 24, 2017
Outline

1. Introduction (Review of last presentation)
2. Intermediate Design
3. Supplement Design
4. Comparison of Alternative Designs for Autoregressive Population
5. Application to CEAP Data
6. Conclusion and discussion
Sampling Design

- 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 estimate a 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 $s_h^2 = (n_h - 1)^{-1} \sum_{i \in A_h} (y_{hi} - \bar{y}_h)^2$. 
Measurable Designs

- Measurable: $\pi_i > 0$ for all $i$ and $\pi_{ik} > 0$ for all $i, k$.
- For measurable designs, it is possible to construct a design-unbiased estimator of the variance of a design linear estimator (Fuller, 2009, p.11).
- Non-measurable designs: one-per-stratum stratified sampling and systematic sampling.
- Solutions to variance estimation in non measurable designs:
  - Collapsed strata,
  - Model-based estimation (Fuller, 2009, sec. 5.3),
  - Design-based solution, sampling with random stratum boundaries (Fuller, 1969).
Figure 1: Stratified sampling, one-per-stratum

Figure 2: Stratified sampling, two-per-stratum
Intermediate Design Procedures

- **Sampling frame structure: sets, pairs and groups:**
  - 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 $h_j, h = 1, 2, \cdots, H, j = 1, 2$.
  - Let the pairs of sets be formed into groups of size $K$ and sample size is $2K$.

- **Sampling Procedures:**
  - 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 3: Intermediate design, $K = 4$]

- $K = 2$

![Figure 4: Intermediate design, $K = 2$]
Assume observations satisfy

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

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

\[ \bar{y}_{2K} = (2K)^{-1} \left[ y_{11,1} + y_{11,2} + \sum_{h=2}^{K} (y_{h1} + y_{h2}) \right]. \]
Ignore finite population correction, the standardized variance is

\[ 2K \bar{V}\{\bar{y}_{2K}\} = \sigma^2_e + K^{-1}\sigma^2_b. \]

Let the estimated between mean square be

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

and the estimated within mean square be

\[ \hat{WMS} = 0.5(y_{111} - y_{112})^2. \]

A nonnegative unbiased estimator of the standardized variance

\[ 2K \hat{V}\{\bar{y}_{2K}\} = K^{-1} B\hat{M}S + (1 - K^{-1})W\hat{M}S. \]
Assume \( y_{hjt} \) are normally distributed, standardized variance of standardized estimated variance is

\[
KV \{ 2K \hat{V}(\bar{y}_{2K}) \} = KV \{ K^{-1} B \hat{M} S + (1 - K^{-1}) W \hat{M} S \} \\
= 2K^{-1}(K - 1)^{-1}(BMS)^2 + 2K(1 - K^{-1})^2 W MS^4
\]

The variance of the variance for total sample is

\[
V \{ \hat{V}(\bar{y}_{tot}) \} = KV \{ 2K \hat{V}(\bar{y}_{2K}) \} / 4H^3.
\]
Figure 5: Intermediate design, $K = 2$

Figure 6: Supplement design, $n = 4$, $D = 1$
Sampling frame structure:
- $D$: number of supplements.
- Let the population be divided into $n - D$ sets.

Sampling Procedures:
- Choose one set at random in each string of $(n - D)/D$ sets (choose $D$ sets in total),
- Choose two elements in each selected sets,
- Choose one element in each of the remaining $n - 2D$ sets.
Figure 7: Supplement design, $n = 6, D = 1$

Figure 8: Supplement design, $n = 6, D = 2$
Supplement Design Estimators

- Mean estimator:
  \[
  \hat{\mu} = \sum_{h=1}^{n-D} (n - D)^{-1} \bar{y}_h,
  \]
  where
  \[
  \bar{y}_h = \begin{cases} 
  0.5 \sum_{j=1}^{2} y_{hj} & \text{if } n_h = 2 \\
  y_{h1} & \text{if } n_h = 1 
  \end{cases}
  \]

- Variance:
  \[
  V\{\hat{\mu}\} = (n - D)^{-2}[n - 2D + 0.5D]\sigma_e^2,
  \]
  where \(\sigma_e^2\) is the within stratum variance.
Variance estimator:

\[ \hat{V}\{\hat{\mu}\} = (n - D)^{-2}[n - 1.5D]s^2, \]

where

\[ s^2 = D^{-1}0.5 \sum_{h}(y_{h1} - y_{h2})^2. \]

If the observations are normally distributed,

\[ V\{\hat{V}(\hat{\mu})\} = [(n - 2D)^{-2}(n - 1.5D)]^2D^{-1}2\sigma_{e2}^4. \]
Example: infinite population, values on the line $y = x$ for $x \in [0, 8]$.

Table 1: Properties of Stratified Samples for $n = 8$ and Linear Population

<table>
<thead>
<tr>
<th>$n - D$</th>
<th>$D$</th>
<th>$96V(\hat{\mu})$</th>
<th>$45V(s^2)$</th>
<th>$360V{\hat{V}(\hat{\mu})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0</td>
<td>1.0000</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.3861</td>
<td>0.7464</td>
<td>0.7921</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1.9752</td>
<td>0.6914</td>
<td>0.7682</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2.8672</td>
<td>0.9557</td>
<td>1.0704</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.0000</td>
<td>1.7500</td>
<td>1.7500</td>
</tr>
</tbody>
</table>
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

\[ \text{Cov}(y_{t+h}, y_t) = (1 - \rho^2)^{-1} \rho^{|h|} \sigma^2. \]
Population size $N$, sample size $n = 2H$,
Divide the population into $2H$ equal sized sets, size of set $m = \frac{N}{2H}$.
Form the pairs of sets into groups of size $K$.
Intermediate design:

$$2KV_{int}\{\bar{y}_{2K}\} = S_w^2 + K^{-1}S_b^2.$$ 

Supplement design:
- $D$ is the number of supplements in each group,
- divide each group into $2K - D$ strata.

$$V_{sup}\{\hat{\mu}\} = (2K - D)^{-2}(2K - 1.5D)S_1^2$$
Comparison of Alternative Designs

Table 2: Relative variances for intermediate design and supplement design \((m = 100, \sigma^2 = 1)\)

<table>
<thead>
<tr>
<th>(\rho)</th>
<th>(K)</th>
<th>(V(\hat{\mu})/V(\bar{y}_{2K})) D=1</th>
<th>(V(\hat{\mu})/V(\bar{y}_{2K})) D=2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>1.08</td>
<td>1.03</td>
<td>1.13</td>
</tr>
<tr>
<td>0.8</td>
<td>1.07</td>
<td>1.02</td>
<td>1.13</td>
</tr>
<tr>
<td>0.95</td>
<td>1.01</td>
<td>1.00</td>
<td>1.12</td>
</tr>
<tr>
<td>0.995</td>
<td>0.83</td>
<td>0.93</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Xiaofei Zhang (CSSM)
Table 3: Relative variances of estimated variance for intermediate design and supplement design ($m = 100, \sigma^2 = 1$)

<table>
<thead>
<tr>
<th>(\rho)</th>
<th>(K)</th>
<th>(D=1)</th>
<th>(V{\hat{V}(\hat{\mu})}/V{\hat{V}(\bar{y}_{2K})})</th>
<th>(D=2)</th>
<th>(V{\hat{V}(\hat{\mu})}/V{\hat{V}(\bar{y}_{2K})})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>2.29</td>
<td>1.27</td>
<td>3</td>
<td>1.25</td>
</tr>
<tr>
<td>0.8</td>
<td>3</td>
<td>2.28</td>
<td>1.26</td>
<td>3</td>
<td>1.27</td>
</tr>
<tr>
<td>0.95</td>
<td>3</td>
<td>2.22</td>
<td>1.23</td>
<td>3</td>
<td>1.34</td>
</tr>
<tr>
<td>0.995</td>
<td>3</td>
<td>1.91</td>
<td>1.20</td>
<td>3</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.27</td>
<td>1.67</td>
<td>10</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.26</td>
<td>0.66</td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.23</td>
<td>0.66</td>
<td></td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.20</td>
<td>0.69</td>
<td></td>
<td>0.69</td>
</tr>
</tbody>
</table>
Assume $\frac{\bar{y}}{\sqrt{V(\bar{y})}} \sim t_{df}$, $V(\bar{y}) \sim \chi_n^2$, then $df = \frac{2V^2(\bar{y})}{V\{\hat{V}(\bar{y})\}}$.

Table 4: Degree of freedoms for intermediate design and supplement design ($m = 100$, $\sigma^2 = 1$, $n = 120$)

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>K</th>
<th>Intermediate design</th>
<th>Supplement design(D=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>39.204</td>
<td>7.250</td>
<td>20.000</td>
</tr>
<tr>
<td>0.8</td>
<td>38.871</td>
<td>7.010</td>
<td>19.444</td>
</tr>
<tr>
<td>0.95</td>
<td>38.236</td>
<td>6.361</td>
<td>17.658</td>
</tr>
<tr>
<td>0.995</td>
<td>39.236</td>
<td>5.898</td>
<td>14.200</td>
</tr>
</tbody>
</table>
Conservation Effects Assessment Project (CEAP).

Variables of interest: acres for each land use (cultivated cropland, non-cultivated cropland, pastureland and other uses).

Population size $N = 612$ segments, sample size $n = 36$.

Table 5: Missouri (2010)

<table>
<thead>
<tr>
<th>Segment ID</th>
<th>Geo-order</th>
<th>Cultivated Cropland</th>
<th>Noncultivated Cropland</th>
<th>Pastureland</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>577</td>
<td>1713</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>37.15</td>
<td>-89.54</td>
</tr>
<tr>
<td>290</td>
<td>1822</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>37.22</td>
<td>-94.59</td>
</tr>
<tr>
<td>291</td>
<td>2018</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>37.25</td>
<td>-94.57</td>
</tr>
<tr>
<td>289</td>
<td>2019</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>37.26</td>
<td>-94.61</td>
</tr>
<tr>
<td>294</td>
<td>2023</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>37.32</td>
<td>-94.49</td>
</tr>
</tbody>
</table>
Data Description

Scatterplot

Geo-order

states_map$lat

36 37 38 39 40

36 37 38 39 40

states_map$long

-96 -94 -92 -90

-96 -94 -92 -90

Latitude

37.5 38.5 39.5 40.5

37.5 38.5 39.5 40.5

Longitude

-95 -93 -91

-95 -93 -91
Stratification

Order 1

Order 2

Order 3

Order 4
Sampling Methods

- Five methods:
  - Systematic sampling
  - Local pivotal method (LPM)
  - Stratified sampling: one-per-stratum and two-per-stratum
  - Intermediate design: $K = 3, 6$
  - Supplement design: $D = 6, 3$. 
Sample Example

intermediate (G=9)

supplement (D=9)

intermediate (G=3)

supplement (D=3)
Variance Estimation

- One-per-stratum, using collapsed strata.
  \[
  \hat{V}(\bar{y}_s - \bar{y}_N) = \left(\frac{2}{M}\right) \sum_{j=1}^{M/2} (y_{2j-1} - y_{2j})^2 / 2/M
  \]

- LPM, 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)
  \[
  \hat{V}(\bar{y}_s - \bar{y}_N) = \left(\frac{1}{M}\right) \sum_{j=2}^{M} (y_j - y_{j-1})^2 / 2(M - 1)
  \]
### Table 6: Variance of mean (latitude) \( \times 10^{-4} \)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Sys</th>
<th>1-per</th>
<th>2-per</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.9</td>
<td>0.4</td>
<td>1.3</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>2.6</td>
<td>9.7</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>4.2</td>
<td>12.6</td>
</tr>
<tr>
<td>4</td>
<td>291.2</td>
<td>9.9</td>
<td>71.7</td>
</tr>
</tbody>
</table>
### Table 7: Variance of mean (cultivated cropland) ($\times 10^{-2}$)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Sys</th>
<th>1-per</th>
<th>2-per</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.51</td>
<td>2.84</td>
<td>2.82</td>
</tr>
<tr>
<td>2</td>
<td>4.01</td>
<td>2.81</td>
<td>2.83</td>
</tr>
<tr>
<td>3</td>
<td>1.53</td>
<td>2.79</td>
<td>2.81</td>
</tr>
<tr>
<td>4</td>
<td>2.49</td>
<td>2.70</td>
<td>2.91</td>
</tr>
</tbody>
</table>
Table 8: Variance of mean (cultivated cropland) ($\times 10^{-2}$)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Sys</th>
<th>Lpm</th>
<th>Int 3</th>
<th>Sup 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.51</td>
<td>2.85</td>
<td>2.83</td>
<td>3.06</td>
</tr>
<tr>
<td>2</td>
<td>4.01</td>
<td>2.82</td>
<td>2.83</td>
<td>3.14</td>
</tr>
<tr>
<td>4</td>
<td>2.49</td>
<td>2.91</td>
<td>2.84</td>
<td>3.12</td>
</tr>
</tbody>
</table>
### Table 9: Relative bias of estimated variance (cultivated cropland)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Sys</th>
<th>Lpm</th>
<th>1-per</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.8</td>
<td>-1.0</td>
<td>4.2</td>
</tr>
<tr>
<td>2</td>
<td>30.2</td>
<td>-0.9</td>
<td>5.6</td>
</tr>
<tr>
<td>3</td>
<td>14.2</td>
<td>-0.9</td>
<td>4.1</td>
</tr>
<tr>
<td>4</td>
<td>-0.5</td>
<td>-0.8</td>
<td>13.0</td>
</tr>
</tbody>
</table>
## Table 10: Standard error of estimated variance (latitude)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>sys</th>
<th>lpm</th>
<th>1-per</th>
<th>2-per</th>
<th>int 3</th>
<th>sup 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.11</td>
<td>0.28</td>
<td>0.75</td>
<td>0.63</td>
<td>0.39</td>
<td>0.60</td>
</tr>
<tr>
<td>2</td>
<td>1.16</td>
<td>2.08</td>
<td>4.78</td>
<td>4.45</td>
<td>2.56</td>
<td>3.11</td>
</tr>
<tr>
<td>3</td>
<td>3.96</td>
<td>4.03</td>
<td>6.56</td>
<td>5.82</td>
<td>4.09</td>
<td>6.25</td>
</tr>
<tr>
<td>4</td>
<td>3.41</td>
<td>5.48</td>
<td>18.96</td>
<td>27.04</td>
<td>12.80</td>
<td>74.79</td>
</tr>
</tbody>
</table>
### Table 11: Standard error of estimated variance (cultivated cropland)

<table>
<thead>
<tr>
<th>Procedure</th>
<th>sys</th>
<th>lpm</th>
<th>1-per</th>
<th>2-per</th>
<th>int 3</th>
<th>sup 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.47</td>
<td>1.04</td>
<td>0.87</td>
<td>0.82</td>
<td>0.97</td>
<td>1.55</td>
</tr>
<tr>
<td>2</td>
<td>0.72</td>
<td>1.08</td>
<td>0.86</td>
<td>0.81</td>
<td>0.99</td>
<td>1.57</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>1.07</td>
<td>0.87</td>
<td>0.81</td>
<td>0.97</td>
<td>1.53</td>
</tr>
<tr>
<td>4</td>
<td>0.44</td>
<td>1.06</td>
<td>0.90</td>
<td>0.83</td>
<td>0.98</td>
<td>1.55</td>
</tr>
</tbody>
</table>
Table 12: $S_w^2$ and $\sigma_b^2$ (N=612, n=36)

<table>
<thead>
<tr>
<th></th>
<th>Cultivated Cropland</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>order1</td>
<td>$S_w^2$</td>
<td>1.086</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>$\sigma_b^2$</td>
<td>-0.013</td>
<td>0.006</td>
</tr>
<tr>
<td>order2</td>
<td>$S_w^2$</td>
<td>1.073</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\sigma_b^2$</td>
<td>0.020</td>
<td>0.052</td>
</tr>
<tr>
<td>order4</td>
<td>$S_w^2$</td>
<td>1.033</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>$\sigma_b^2$</td>
<td>0.155</td>
<td>0.459</td>
</tr>
</tbody>
</table>
Conclusion

- We present intermediate design and supplement design. Unbiased variance estimators are derived for both designs.
- Efficiency of intermediate design performed better than supplement design for the variance of mean and variance of the estimated variance.
- Efficiency of intermediate design performed well for population with large between mean square.

Extension

- Unequal probability.
- Auxiliary information.
- General population size.
- Forming area strata.
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.