Using Auxiliary Information Under a Generic Sampling Design

Giancarlo Diana, Pier Francesco Perri

Department of Statistical Sciences, University of Padova
Department of Economics and Statistics, University of Calabria

19th International Conference on Computational Statistics

August 22-27, 2010 - Paris, France

Outline

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
  1. without reasonable motivations
  2. comparing them with estimators that are less efficient
  3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
  1. without reasonable motivations
  2. comparing them with estimators that are less efficient
  3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
  1. without reasonable motivations
  2. comparing them with estimators that are less efficient
  3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
  1. without reasonable motivations
  2. comparing them with estimators that are less efficient
  3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
  1. without reasonable motivations
  2. comparing them with estimators that are less efficient
  3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
    1. without reasonable motivations
    2. comparing them with estimators that are less efficient
    3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Auxiliary information

- Auxiliary information plays a relevant role in sampling to obtain improved design and/or more efficient estimators
- When auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* methods are widely employed in many situations
- Researchers interested in the estimation of population parameters can find a huge variety of proposals in the literature. New estimators are usually proposed by modifying the structure of existing ones but...
    1. without reasonable motivations
    2. comparing them with estimators that are less efficient
    3. overlooking that, at best, they can be equivalent to the regression estimator
- This practice has inundated the literature with papers whose theoretical and practical relevance appears rather questionable

Our aim

Motivated by Bacanli and Kadilar (BK, 2008), we intend to show how the problem of finding the best estimator for the mean of a study variable can be treated under a generic sampling design by means of a very simple class of estimators

- The class is not exhaustive and a more general discussion can be found, among others, in Diana and Perri (2007)
- The best estimator in the class is compared with BK estimators according to UPS, where inclusion probability are computed on the basis of a limited numbers of samples

Our aim

Motivated by Bacanli and Kadilar (BK, 2008), we intend to show how the problem of finding the best estimator for the mean of a study variable can be treated under a generic sampling design by means of a very simple class of estimators

- The class is not exhaustive and a more general discussion can be found, among others, in Diana and Perri (2007)
- The best estimator in the class is compared with BK estimators according to UPS, where inclusion probability are computed on the basis of a limited numbers of samples

Our aim

Motivated by Bacanli and Kadilar (BK, 2008), we intend to show how the problem of finding the best estimator for the mean of a study variable can be treated under a generic sampling design by means of a very simple class of estimators

- The class is not exhaustive and a more general discussion can be found, among others, in Diana and Perri (2007)
- The best estimator in the class is compared with BK estimators according to UPS, where inclusion probability are computed on the basis of a limited numbers of samples

Notation

- $U = \{1, 2, ..., N\}$ a finite population
- Y a study variable with unknown mean $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$
- X an auxiliary variable with $\bar{X} = N^{-1} \sum_{i=1}^{N} x_i$ known
- $p(s)$ a generic sampling design
- s a sample of size n from $p(s)$
- $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni (i,j)} p(s)$ the first and second order inclusion probabilities
- $\hat{\bar{Y}}$, $\hat{\bar{X}}$ two unbiased estimators of $\bar{Y}$, $\bar{X}$ under $p(s)$
- τ a constant that may be related to population parameters

A class of estimators for $\bar{Y}$

We introduce a very simple class of estimators for $\bar{Y}$ as

$$\hat{\bar{Y}}_{pr} = \hat{\bar{Y}} \frac{\bar{X} + \tau}{\hat{\bar{X}} + \tau}$$

Expanding it in a Taylor's series (δ-method) and retaining only terms up to the second degree, we get - for n sufficiently large - the first order approximation of the bias (B) and mean square error (MSE)

$$B(\hat{\bar{Y}}_{pr}) = \frac{1}{\bar{X} + \tau} \left[\frac{\bar{Y} Var(\hat{\bar{X}})}{\bar{X} + \tau} - Cov(\hat{\bar{X}}, \hat{\bar{Y}}) \right]$$

$$MSE(\hat{\bar{Y}}_{pr}) = Var(\hat{\bar{Y}}) + \frac{\bar{Y}^2 Var(\hat{\bar{X}})}{(\bar{X} + \tau)^2} - \frac{2\bar{Y} Cov(\hat{\bar{X}}, \hat{\bar{Y}})}{\bar{X} + \tau}$$

Optimality of the class

Minimization of $MSE(\hat{\hat{Y}}_{pr})$ is achieved for

$$\tau = \bar{X}\frac{[C(\hat{\hat{X}})^2 - C(\hat{\hat{X}}, \hat{\hat{Y}})]}{C(\hat{\hat{X}}, \hat{\hat{Y}})}$$

with $C(\hat{\hat{X}}) = \sqrt{Var(\hat{\hat{X}})}/\bar{X}$, $C(\hat{\hat{X}}, \hat{\hat{Y}}) = Cov(\hat{\hat{X}}, \hat{\hat{Y}})/\bar{X}\bar{Y}$. For this optimum choice, we get

$$minMSE(\hat{\hat{Y}}_{pr}) = Var(\hat{\hat{Y}})(1 - \rho^2_{\hat{\hat{X}}, \hat{\hat{Y}}})$$

which is the variance of the regression estimator

$$\hat{\hat{Y}}_{lr} = \hat{\hat{Y}} + \beta_{\hat{\hat{Y}}, \hat{\hat{X}}}(\bar{X} - \hat{\hat{X}})$$

Optimality of the class

Minimization of $MSE(\hat{\bar{Y}}_{pr})$ is achieved for

$$\tau = \bar{X}\frac{[C(\hat{\bar{X}})^2 - C(\hat{\bar{X}}, \hat{\bar{Y}})]}{C(\hat{\bar{X}}, \hat{\bar{Y}})}$$

with $C(\hat{\bar{X}}) = \sqrt{Var(\hat{\bar{X}})}/\bar{X}$, $C(\hat{\bar{X}}, \hat{\bar{Y}}) = Cov(\hat{\bar{X}}, \hat{\bar{Y}})/\bar{X}\bar{Y}$. For this optimum choice, we get

$$minMSE(\hat{\bar{Y}}_{pr}) = Var(\hat{\bar{Y}})(1 - \rho^2_{\hat{\bar{X}}, \hat{\bar{Y}}})$$

which is the variance of the regression estimator

$$\hat{\bar{Y}}_{lr} = \hat{\bar{Y}} + \beta_{\hat{\bar{Y}}, \hat{\bar{X}}}(\bar{X} - \hat{\bar{X}})$$

Optimality of $\hat{\bar{Y}}_{lr}$ is well-known in sampling theory but this aspect is very often overlooked. Why?

Efficiency considerations

- All the estimators belonging to the class can be only, at best, as efficient as $\hat{\bar{Y}}_{lr}$. They are equivalent to it only when $\tau = \bar{X}\frac{[C(\hat{\bar{X}})^2 - C(\hat{\bar{X}},\hat{\bar{Y}})]}{C(\hat{\bar{X}},\hat{\bar{Y}})}$

Efficiency considerations

- All the estimators belonging to the class can be only, at best, as efficient as $\hat{\bar{Y}}_{lr}$. They are equivalent to it only when $\tau = \bar{X}\frac{[C(\hat{\bar{X}})^2 - C(\hat{\bar{X}},\hat{\bar{Y}})]}{C(\hat{\bar{X}},\hat{\bar{Y}})}$
- For instance, the following estimators (in SRSWOR) are not optimum in the class

Authors	Estimators	τ
Sisodia and Dwivedi (1981)	$\hat{\bar{Y}}_{SD} = \hat{\bar{Y}}\frac{\bar{X} + C_x}{\hat{\bar{X}} + C_x}$	C_x
Singh and Kakran (1993)	$\hat{\bar{Y}}_{SK} = \hat{\bar{Y}}\frac{\bar{X} + \beta_2(x)}{\hat{\bar{X}} + \beta_2(x)}$	$\beta_2(x)$
Upadhyaya and Singh (1999)	$\hat{\bar{Y}}_{US1} = \hat{\bar{Y}}\frac{\bar{X}\beta_2(x) + C_x}{\hat{\bar{X}}\beta_2(x) + C_x}$	$C_x/\beta_2(x)$
	$\hat{\bar{Y}}_{US2} = \hat{\bar{Y}}\frac{\bar{X}C_x + \beta_2(x)}{\hat{\bar{X}}C_x + \beta_2(x)}$	$\beta_2(x)/C_x$

Bacanli-Kadilar estimators

Previous estimators have been considered by Bacanli and Kadilar (2008) under UPSWOR by replacing $\hat{\bar{Y}}$ and $\hat{\bar{X}}$ with Horvitz-Thompson estimator

$$\hat{\bar{T}}_{HT} = \frac{1}{N}\sum_{i\in s}\frac{t_i}{\pi_i}, \qquad t = x, y$$

with

$$Var(\hat{\bar{T}}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)t_i t_j, \qquad t = x, y$$

Bacanli-Kadilar estimators

Previous estimators have been considered by Bacanli and Kadilar (2008) under UPSWOR by replacing $\hat{\bar{Y}}$ and $\hat{\bar{X}}$ with Horvitz-Thompson estimator

$$\hat{\bar{T}}_{HT} = \frac{1}{N}\sum_{i\in s}\frac{t_i}{\pi_i}, \quad t = x, y$$

with

$$Var(\hat{\bar{T}}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)t_i t_j, \quad t = x, y$$

- The modified estimators have been analytically compared with the ratio estimator $\hat{\bar{Y}}_r = (\hat{\bar{Y}}_{HT}/\hat{\bar{X}}_{HT})\bar{X}$
- Numerical comparisons have been performed by using exact expressions for π_i and π_{ij} inherited from the *adaptive cluster sampling*

Some questions

1. BK estimators belong to the proposed class but they are not optimum
2. Why to compare these estimators with $\hat{\bar{Y}}_r = (\hat{\bar{Y}}_{HT}/\hat{\bar{X}}_{HT})\bar{X}$ and not with $\hat{\bar{Y}}_{lr} = \hat{\bar{Y}}_{HT} + \beta_{\hat{\bar{Y}}_{HT},\hat{\bar{X}}_{HT}}(\bar{X} - \hat{\bar{X}}_{HT})$?

   It is well-known that $MSE(\hat{\bar{Y}}_r) \geq MSE(\hat{\bar{Y}}_{lr})$

   BK estimators can not outperform $\hat{\bar{Y}}_{lr}$
3. The use of the exact expressions for π_i and π_{ij} from *adaptive cluster sampling* seems to be rather questionable

Some questions

1. BK estimators belong to the proposed class but they are not optimum
2. Why to compare these estimators with $\hat{\bar{Y}}_r = (\hat{\bar{Y}}_{HT}/\hat{\bar{X}}_{HT})\bar{X}$ and not with $\hat{\bar{Y}}_{lr} = \hat{\bar{Y}}_{HT} + \beta_{\hat{\bar{Y}}_{HT},\hat{\bar{X}}_{HT}}(\bar{X} - \hat{\bar{X}}_{HT})$?

   It is well-known that $MSE(\hat{\bar{Y}}_r) \geq MSE(\hat{\bar{Y}}_{lr})$

   BK estimators can not outperform $\hat{\bar{Y}}_{lr}$
3. The use of the exact expressions for π_i and π_{ij} from *adaptive cluster sampling* seems to be rather questionable

Some questions

1. BK estimators belong to the proposed class but they are not optimum
2. Why to compare these estimators with $\hat{\bar{Y}}_r = (\hat{\bar{Y}}_{HT}/\hat{\bar{X}}_{HT})\bar{X}$ and not with $\hat{\bar{Y}}_{lr} = \hat{\bar{Y}}_{HT} + \beta_{\hat{\bar{Y}}_{HT},\hat{\bar{X}}_{HT}}(\bar{X} - \hat{\bar{X}}_{HT})$?

   It is well-known that $MSE(\hat{\bar{Y}}_r) \geq MSE(\hat{\bar{Y}}_{lr})$
   BK estimators can not outperform $\hat{\bar{Y}}_{lr}$
3. The use of the exact expressions for π_i and π_{ij} from *adaptive cluster sampling* seems to be rather questionable

Some questions

1. BK estimators belong to the proposed class but they are not optimum
2. Why to compare these estimators with $\hat{\bar{Y}}_r = (\hat{\bar{Y}}_{HT}/\hat{\bar{X}}_{HT})\bar{X}$ and not with $\hat{\bar{Y}}_{lr} = \hat{\bar{Y}}_{HT} + \beta_{\hat{\bar{Y}}_{HT},\hat{\bar{X}}_{HT}}(\bar{X} - \hat{\bar{X}}_{HT})$?

   It is well-known that $MSE(\hat{\bar{Y}}_r) \geq MSE(\hat{\bar{Y}}_{lr})$
   BK estimators can not outperform $\hat{\bar{Y}}_{lr}$
3. The use of the exact expressions for π_i and π_{ij} from *adaptive cluster sampling* seems to be rather questionable

Possible solution: $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}
- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i / M$ and $\hat{\pi}_{ij} = M_{ij} / M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i / M$ and $\hat{\pi}_{ij} = M_{ij} / M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i/M$ and $\hat{\pi}_{ij} = M_{ij}/M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i/M$ and $\hat{\pi}_{ij} = M_{ij}/M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i/M$ and $\hat{\pi}_{ij} = M_{ij}/M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i/M$ and $\hat{\pi}_{ij} = M_{ij}/M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i/M$ and $\hat{\pi}_{ij} = M_{ij}/M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Computing π_i and π_{ij}

- The explicit derivation of π_i and π_{ij} becomes prohibitive when N and/or n increase: $\binom{N}{n}$ samples are to be investigated
- To overcome the problem, a solution can be adopted by simulating π_i and π_{ij}

Algorithm

1. each unit has a selection probability $p_i = z_i / \sum_{j=1}^{N} z_j$
2. $M < \binom{N}{n}$ samples WOR are independently drawn from U
3. M_i and M_{ij} are the number of samples that contain unit i and units (i,j)
4. estimate π_i and π_{ij} with $\hat{\pi}_i = M_i/M$ and $\hat{\pi}_{ij} = M_{ij}/M$
5. modify HT-estimator and its variance by using $\hat{\pi}_i$ and $\hat{\pi}_{ij}$

- implement in R the procedure drawing the PPS samples by `sample(U,n,replace=FALSE,prob=p)`

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\hat{Y}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i+p_j) + \frac{n-2}{N-2}\right]$

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\bar{Y}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right]$

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\bar{Y}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. EPS: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. UPS: units are selected according to Midzuno scheme: $\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i+p_j)+\frac{n-2}{N-2}\right]$

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\hat{Y}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. EPS: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. UPS: units are selected according to Midzuno scheme: $\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i+p_j) + \frac{n-2}{N-2}\right]$

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\hat{Y}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1 - p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right]$

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\hat{\bar{Y}}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1 - p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right]$

Numerical examples

- Numerical study on real data from Cochran (1977, p. 34): weekly expenditure on food (y), weekly family income (x) and number of persons per family (z)
- $N = 32$ and $n = 10, 15, 20$.
- $M = 100\,000$ samples are considered instead of $\binom{32}{n}$
- To evaluate the performance of $\hat{\pi}_i$ and $\hat{\pi}_{ij}$, we compare the efficiency of $\hat{\hat{Y}}_{lr}$ with that of BK estimators
- Two situations are considered
  1. **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  2. **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1 - p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right]$

MSE under exact and simulated π_i and π_{ij} - SRSWOR

- **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$

n	$\hat{\bar{Y}}_r$	$\hat{\bar{Y}}_{lr}$	$\hat{\bar{Y}}_{SD}$	$\hat{\bar{Y}}_{SK}$	$\hat{\bar{Y}}_{US1}$	$\hat{\bar{Y}}_{US2}$
10	6.119	6.098	6.118	6.122	6.119	6.154
	6.108	**6.063**	**6.107**	**6.112**	**6.109**	**6.148**
15	3.152	3.141	3.152	3.154	3.152	3.170
	3.148	**3.14**	**3.145**	**3.147**	**3.145**	**3.162**
20	1.669	1.663	1.669	1.670	1.669	1.678
	1.669	**1.667**	**1.668**	**1.669**	**1.669**	**1.678**

MSE under exact and simulated π_i and π_{ij} - SRSWOR

- **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$

n	$\hat{\hat{Y}}_r$	$\hat{\hat{Y}}_{lr}$	$\hat{\hat{Y}}_{SD}$	$\hat{\hat{Y}}_{SK}$	$\hat{\hat{Y}}_{US1}$	$\hat{\hat{Y}}_{US2}$
10	6.119	6.098	6.118	6.122	6.119	6.154
	6.108	**6.063**	**6.107**	**6.112**	**6.109**	**6.148**
15	3.152	3.141	3.152	3.154	3.152	3.170
	3.148	**3.14**	**3.145**	**3.147**	**3.145**	**3.162**
20	1.669	1.663	1.669	1.670	1.669	1.678
	1.669	**1.667**	**1.668**	**1.669**	**1.669**	**1.678**

- despite of the severe reduction of the cardinality of the sample space, no striking differences appear in the precision of the estimators
- small variations tend to disappear as $n \uparrow$

MSE under exact and simulated π_i and π_{ij} - SRSWOR

- **EPS**: units are selected according to SRSWOR for which π_i and π_{ij} are known in advance: $\pi_i = n/N$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$

n	$\hat{\bar{Y}}_r$	$\hat{\bar{Y}}_{lr}$	$\hat{\bar{Y}}_{SD}$	$\hat{\bar{Y}}_{SK}$	$\hat{\bar{Y}}_{US1}$	$\hat{\bar{Y}}_{US2}$
10	6.119	6.098	6.118	6.122	6.119	6.154
	6.108	**6.063**	**6.107**	**6.112**	**6.109**	**6.148**
15	3.152	3.141	3.152	3.154	3.152	3.170
	3.148	**3.14**	**3.145**	**3.147**	**3.145**	**3.162**
20	1.669	1.663	1.669	1.670	1.669	1.678
	1.669	**1.667**	**1.668**	**1.669**	**1.669**	**1.678**

- ♠ despite of the severe reduction of the cardinality of the sample space, no striking differences appear in the precision of the estimators
- ♠ small variations tend to disappear as $n \uparrow$

MSE under exact and simulated π_i and π_{ij} - Midzuno UPS

- **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i+p_j) + \frac{n-2}{N-2}\right]$

n	Method	$\hat{\bar{Y}}_r$	$\hat{\bar{Y}}_{lr}$	$\hat{\bar{Y}}_{SD}$	$\hat{\bar{Y}}_{SK}$	$\hat{\bar{Y}}_{US1}$	$\hat{\bar{Y}}_{US2}$
10	Midzuno	5.978	**5.498**	5.976	5.988	5.980	6.064
	Est. prob.	5.813	**4.945**	5.796	5.917	5.832	6.665
15	Midzuno	3.102	**2.963**	3.102	3.106	3.103	3.134
	Est. prob.	2.877	**2.065**	2.870	2.925	2.886	3.267
20	Midzuno	1.649	**1.606**	1.649	1.650	1.649	1.663
	Est. prob.	1.405	**0.886**	1.401	1.425	1.408	1.571

MSE under exact and simulated π_i and π_{ij} - Midzuno UPS

- **UPS**: units are selected according to Midzuno scheme: $\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i+p_j) + \frac{n-2}{N-2}\right]$

n	Method	$\hat{\bar{Y}}_r$	$\hat{\bar{Y}}_{lr}$	$\hat{\bar{Y}}_{SD}$	$\hat{\bar{Y}}_{SK}$	$\hat{\bar{Y}}_{US1}$	$\hat{\bar{Y}}_{US2}$
10	Midzuno	5.978	**5.498**	5.976	5.988	5.980	6.064
	Est. prob.	5.813	**4.945**	5.796	5.917	5.832	6.665
15	Midzuno	3.102	**2.963**	3.102	3.106	3.103	3.134
	Est. prob.	2.877	**2.065**	2.870	2.925	2.886	3.267
20	Midzuno	1.649	**1.606**	1.649	1.650	1.649	1.663
	Est. prob.	1.405	**0.886**	1.401	1.425	1.408	1.571

- UPS with $\hat{\pi}_i$ and $\hat{\pi}_{ij}$ offers the best solution if compared with the Midzuno scheme (and SRSWOR)
- $\hat{\bar{Y}}_{lr}$ outperforms all the other estimators whatever n
- the gain in efficiency rises as $n \uparrow$

MSE under exact and simulated π_i and π_{ij} - Midzuno UPS

- **UPS**: units are selected according to Midzuno scheme:
$\pi_i = p_i + (1-p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right]$

n	Method	$\hat{\bar{Y}}_r$	$\hat{\bar{Y}}_{lr}$	$\hat{\bar{Y}}_{SD}$	$\hat{\bar{Y}}_{SK}$	$\hat{\bar{Y}}_{US1}$	$\hat{\bar{Y}}_{US2}$
10	Midzuno	5.978	**5.498**	5.976	5.988	5.980	6.064
	Est. prob.	5.813	**4.945**	5.796	5.917	5.832	6.665
15	Midzuno	3.102	**2.963**	3.102	3.106	3.103	3.134
	Est. prob.	2.877	**2.065**	2.870	2.925	2.886	3.267
20	Midzuno	1.649	**1.606**	1.649	1.650	1.649	1.663
	Est. prob.	1.405	**0.886**	1.401	1.425	1.408	1.571

- UPS with $\hat{\pi}_i$ and $\hat{\pi}_{ij}$ offers the best solution if compared with the Midzuno scheme (and SRSWOR)
- $\hat{\bar{Y}}_{lr}$ outperforms all the other estimators whatever n
- the gain in efficiency rises as $n \uparrow$

MSE under exact and simulated π_i and π_{ij} - Midzuno UPS

- **UPS**: units are selected according to Midzuno scheme:
  $\pi_i = p_i + (1 - p_i)\frac{n-1}{N-1}$ and $\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right]$

n	Method	$\hat{\bar{Y}}_r$	$\hat{\bar{Y}}_{lr}$	$\hat{\bar{Y}}_{SD}$	$\hat{\bar{Y}}_{SK}$	$\hat{\bar{Y}}_{US1}$	$\hat{\bar{Y}}_{US2}$
10	Midzuno	5.978	**5.498**	5.976	5.988	5.980	6.064
	Est. prob.	5.813	**4.945**	5.796	5.917	5.832	6.665
15	Midzuno	3.102	**2.963**	3.102	3.106	3.103	3.134
	Est. prob.	2.877	**2.065**	2.870	2.925	2.886	3.267
20	Midzuno	1.649	**1.606**	1.649	1.650	1.649	1.663
	Est. prob.	1.405	**0.886**	1.401	1.425	1.408	1.571

- UPS with $\hat{\pi}_i$ and $\hat{\pi}_{ij}$ offers the best solution if compared with the Midzuno scheme (and SRSWOR)
- $\hat{\bar{Y}}_{lr}$ outperforms all the other estimators whatever n
- the gain in efficiency rises as $n \uparrow$

Final remarks

- Numerical comparisons based on the first order MSE seem to be not affected by n
- When using auxiliary information at the estimation stage, the optimal estimator is the regression estimator. No improvement upon it can be achieved, at least up to the first order of approximation
- The awareness of this aspect should avoid the proliferation of estimators that appear different each others but whose efficiency is known in advance
- When using UPS, the cumbersome problem of the explicit derivation of π_i and π_{ij} can be reduced by their estimation over a limited number of samples

Final remarks

- Numerical comparisons based on the first order MSE seem to be not affected by n
- When using auxiliary information at the estimation stage, the optimal estimator is the regression estimator. No improvement upon it can be achieved, at least up to the first order of approximation
- The awareness of this aspect should avoid the proliferation of estimators that appear different each others but whose efficiency is known in advance
- When using UPS, the cumbersome problem of the explicit derivation of π_i and π_{ij} can be reduced by their estimation over a limited number of samples

Final remarks

- Numerical comparisons based on the first order MSE seem to be not affected by n
- When using auxiliary information at the estimation stage, the optimal estimator is the regression estimator. No improvement upon it can be achieved, at least up to the first order of approximation
- The awareness of this aspect should avoid the proliferation of estimators that appear different each others but whose efficiency is known in advance
- When using UPS, the cumbersome problem of the explicit derivation of π_i and π_{ij} can be reduced by their estimation over a limited number of samples

Final remarks

- Numerical comparisons based on the first order MSE seem to be not affected by n
- When using auxiliary information at the estimation stage, the optimal estimator is the regression estimator. No improvement upon it can be achieved, at least up to the first order of approximation
- The awareness of this aspect should avoid the proliferation of estimators that appear different each others but whose efficiency is known in advance
- When using UPS, the cumbersome problem of the explicit derivation of π_i and π_{ij} can be reduced by their estimation over a limited number of samples

Thanks for your attention

References

BACANLI, S., KADILAR, C. (2008): Ratio estimators with unequal probability designs. *Pakistan Journal of Statistics 24 (3), 167-172*

COCHRAN, W.G. (1977): *Sampling Techniques*. John Wiley & Sons, New York

DIANA, G., PERRI, P.F. (2007): Estimation of finite population mean using multi-auxiliary information. *Metron LXV (1) 99-112*

FATTORINI, L. (2006): Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. *Biometrika 93 (10), 269-278*

SINGH, H.P., KAKRAN, M.S.(1993): A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. *Unpublished manuscript*

SISODIA, B.V.S., DWIVEDI, V.K. (1981): A modified ratio estimator using coefficient of variation of an auxiliary character. *Journal of Indian Society of Agricultural Statistics 33 (2), 13-18*

THOMPSON, S.K, SEBER, G.A.F. (1996): *Adaptive Sampling*. John Wiley & Sons, New York

UPADHYAYA, L.N., SINGH, H.P. (1999): Use a transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal 45 (5), 627-636*