Model-Based Sampling Designs for Optimum Estimation

Sun-Woong Kim
Steven G. Heeringa
Peter W. Solenberger

Dongguk University
&
Survey Research Center
University of Michigan

Overview

- Why Superpopulation in Model-Based Inferences?
- Why Concern about Model-Based Inferences?
- Considering Population Models for Sampling Designs
- H-T Estimator
- πps Sampling Designs
- πps Sampling Designs based on Models
- Concluding Remarks

Why Superpopulation in Model-Based Inferences?

- Assume that the finite population is a random sample drawn from a larger population

- A superpopulation distribution ξ is an expression of a subjective belief or a prior knowledge

- The distribution ξ is used to make a theoretical comparison of the efficiency of estimators under alternative sampling designs $p(\cdot)$ (sampling plans)

Why Concern about Model-Based Inferences?

- The assumption of a model that is not known to hold risks large bias if the assumed model does not in fact hold
- Design-based inferences may yield larger variances, but they avoid the risk of biased inferences

(See Hansen, Madow and Tepping (1983))

Considering Population Models for Sampling Designs

- If we know enough about a population, a population model can guide useful procedures for selecting samples and give increased precision

- Sampling designs according to model assumptions should be established through an appropriate estimator such as the one of Horvitz and Thompson (1952) in order to avoid a bias

- One of the main concerns may be, given some population model assumption, what sampling design is optimal for the H-T estimator and how to find it

H-T Estimator

Let y_i be the value of the characteristic of the unit i in a finite population

- H-T estimator for the population total Y

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where $\pi_i = \sum_{i \in s} p(s)$ is the first-order inclusion probability and $p(s)$ denotes the selection probability of a sample s

- There are various forms of the variance of $\hat{Y}_{HT}$, which involves $\pi_{ij} = \sum_{i,j \in s} p(s)$, the joint probability

πps Sampling Designs

- Satisfying the condition :

$$\pi_i = np_i$$

where n is the sample size, $p_i = x_i / X$, $X = \sum x_i$ and x_i is a value correlated with the y_i

- In πps sampling designs the H-T estimator can be expressed as

$$\hat{Y}_{HT} = \frac{X}{n} \sum_{i \in s} \frac{y_i}{x_i}$$

πps Sampling Designs based on Models

- **Raj's (1956) Approach**

  - Simple Model (No Error Term): $y_i = \alpha + \beta x_i$, where α and β are constants

  - Form of variance: $Var_{\mathbf{1}}\left(\hat{Y}_{HT}\right) = \sum_{i=\mathbf{1}}^{N} \frac{y_i^{\mathbf{2}}}{\pi_i} + \mathbf{2}\sum_{i=\mathbf{1}}^{N} \sum_{j>i}^{N} \frac{\pi_{ij}}{\pi_i \pi_j} y_i y_j - Y^{\mathbf{2}}$

  - Optimization problem to minimize $Var_{\mathbf{1}}\left(\hat{Y}_{HT}\right)$ for the sample size $n = 2$ under the model:

$$\boldsymbol{Minimize} \ \sum_{i=\mathbf{1}}^{N} \sum_{j>i}^{N} \frac{\pi_{ij}}{\pi_i \pi_j}$$

$$\text{subject to } \sum_{j \neq i} \pi_{ij} = \pi_i \ , \ i = 1, \cdots, N$$

- **Alternative Optimization Problem for Raj's (1956) Approach**

  - A different form of variance:

$$Var_2\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\frac{y_i^2(1-\pi_i)}{\pi_i}+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_iy_j(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j}$$

$$=\sum_{i=1}^{N}\frac{y_i^2(1-\pi_i)}{\pi_i}+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\pi_{ij}}{\pi_i\pi_j}y_iy_j-2\sum_{i=1}^{N}\sum_{j>i}^{N}y_iy_j$$

  - The second term under the model:

$$\frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_ix_j}\pi_{ij}+\frac{X^2}{n}(n-1)\beta^2$$

- Optimization problem to minimize $Var_{\mathbf{2}}\left(\hat{Y}_{HT}\right)$ for any sample size under the model:

$$\boldsymbol{Minimize}\ \sum_{i=\mathbf{1}}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_i x_j}\pi_{ij}$$

subject to $\sum_{j\neq i}\pi_{ij}=(n-1)\pi_i,\ \ i=1,\cdots,N$

Note. If $\alpha=0$, the second term is reduced to $\frac{X^{\mathbf{2}}}{n}(n-\mathbf{1})\beta^{\mathbf{2}}$, which does not depend on the joint probabilities π_{ij}. So any πps sampling design produces the same variance under this model.

- **Optimization Problem under Superpopulation Model 1:**

$y_i = \alpha + \beta x_i + \varepsilon_i$, where $E_\xi(\varepsilon_i) = 0$, $V_\xi(\varepsilon_i) = \sigma^2 x_i^2$, and $E_\xi(\varepsilon_i \varepsilon_j) = 0$

Here E_ξ denotes the model expectation over all the finite populations that can be drawn from the superpopulation.

- Model expectation for the second term of $Var_2\left(\hat{Y}_{HT}\right)$:

$$E_\xi\left(2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\pi_{ij}}{\pi_i\pi_j}y_iy_j\right) = \frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_ix_j}E_\xi\left(y_iy_j\right)\pi_{ij}$$

$$= \frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_ix_j}\pi_{ij} + \frac{X^2}{n}(n-1)\beta^2$$

Note. This is the same as the one under the former model

- Optimization problem for minimizing the model expectation of $Var_2\left(\hat{Y}_{HT}\right)$ under model 1:

$$\textbf{\textit{Minimize}}\ \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_i x_j}\pi_{ij}$$

subject to $\sum_{j\neq i}\pi_{ij}=(n-1)\pi_i,\ \ i=1,\cdots,N$

Note. In cases of $n=2$, this will be the minimization of the sum of the weighted selection probability for each sample. That is,

$$\textbf{\textit{Minimize}}\ \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_i x_j}p(s)$$

subject to $\sum_{j\neq i}p(s)=\pi_i,\ \ i=1,\cdots,N$

- **Optimization Problem under Superpopulation Model 2:**

$y_i = \alpha + \beta x_i + \varepsilon_i$, where $E_\xi(y_i) = \alpha x_i + \beta$, $V_\xi(y_i) = \sigma^2 x_i^2$, and $Cov_\xi(y_i, y_j) = 0$

- Model expectation of the second term for $Var_2\left(\hat{Y}_{HT}\right)$ is also same with the former.

$$E_\xi\left(2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\pi_{ij}}{\pi_i\pi_j}y_iy_j\right) = \frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_ix_j}E_\xi\left(y_i\right)E_\xi\left(y_j\right)\pi_{ij}$$

$$= \frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_ix_j}\pi_{ij} + \frac{X^2}{n}(n-1)\beta^2$$

- The same optimization problem is built

Remark 1. The three models above give the same optimization problem.

Remark 2. Solving the optimization problem by using some linear programming (LP) software would provide an optimal sampling design $p(\cdot)$ to minimize the model expectation of the design-based variance.

Remark 3. If $\alpha = 0$, the model expectation of the variance is fixed without depending on the joint probabilities π_{ij}. Thus any πps sampling design is acceptable.

- **Rao and Bayless (1969) (1970)**

  - Superpopulation Model 3 (No Intercept Term):

$$y_i = \beta x_i + \varepsilon_i$$

where $E_\xi(\varepsilon_i) = 0,\ E_\xi(\varepsilon_i^2) = a x_i^g\ (a > 0,\ g \geq 0)$, and $E_\xi(\varepsilon_i \varepsilon_j) = 0$

  - Form of variance: $Var_{\mathbf{3}}\left(\hat{Y}_{HT}\right) = \sum_{i=\mathbf{1}}^{N} \sum_{j>i}^{N} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^{\mathbf{2}}$

  - Model expectation of the design-based variance $Var_{\mathbf{3}}$ of the H-T estimator under the model:

$$E_\xi(Var_{\mathbf{3}}) = \frac{aX^g}{n} \sum_{i}^{N} (\mathbf{1} - np_i)\, p_i^{g-\mathbf{1}} = V$$

Note. $E_{\xi}(Var_{\mathbf{3}})$ does not depend on the joint probabilities π_{ij}, so that all πps sampling methods with $\pi_i = np_i$ have the same expected variance.

We don't have to consider the optimization problem to minimize $E_{\xi}(Var_{\mathbf{3}})$

- **Considering Superpopulation Model with Intercept Term**

  - It is customary to introduce the intercept term into the model
  - By allowing the intercept, considerable flexibility can be gained in sampling designs

Superpopulation Model 4:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where $E_\xi(\varepsilon_i) = 0$, $E_\xi(\varepsilon_i^2) = a x_i^g \ (a > 0,\ g \geq 0)$, and $E_\xi(\varepsilon_i \varepsilon_j) = 0$

- Model expectation of Var_3 of the H-T estimator under the model:

$$E_\xi(Var_3) = V + 2\alpha \sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j - x_i\right)\left(\frac{\alpha}{x_i}+\beta\right) + \frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{1}{x_j}-\frac{1}{x_i}\right)\left(\frac{\alpha}{x_i}+\beta\right)\pi_{ij}$$

- Concerning the third term, we may consider the following optimization problem for minimizing the model expectation of Var_3 if we assume that $x_i \le x_j$ for all $i \ne j = 1,\cdots,N$

$$\textbf{\textit{Minimize}} \quad \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{1}{x_j}-\frac{1}{x_i}\right)\left(\alpha\frac{1}{x_i}+\beta\right)\pi_{ij}$$

subject to $\sum_{j \neq i} \pi_{ij} = (n-1)\pi_i, \quad i = 1, \cdots, N$

Remark 4. This optimization problem does not depend on $a(>0)$ or $g(\geq 0)$

Remark 5. Even if the assumed model does not hold (e.g. $\alpha = 0$), the πps sampling designs obtained using LP have the same model expectation of Var_3 with any other πps designs

- Model-based sampling design for minimizing expected variance estimates under superpopulation model 4:

$$\boldsymbol{\mathit{Minimize}} \quad \sum_{s\in S} E_{\xi}(v_{s,SYG})$$

where S is the collection of all possible samples and $v_{s,SYG}$ is the Sen-Yates-Grundy variance estimate given by

$$v_{s,SYG} = \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

- The model expectation of the Sen-Yates-Grundy variance estimate is expressed as

$$E_{\xi}(v_{s,SYG}) = \mathbf{2}aX^{g}\sum_{i=\mathbf{1}}^{n}\sum_{j>i}^{n}\frac{p_i^{g-\mathbf{1}}p_j}{\pi_{ij}} - \frac{\mathbf{2}}{n^{\mathbf{2}}}aX^{g}\sum_{i=\mathbf{1}}^{n}\sum_{j>i}^{n}p_i^{g-\mathbf{2}}$$

$$+\mathbf{2}\alpha X^{\mathbf{2}}\sum_{i=\mathbf{1}}^{n}\sum_{j>i}^{n}\left(\frac{(x_i x_j)/X^{\mathbf{2}}}{\pi_{ij}}\frac{x_j - x_i}{x_i\,x_j}\left(\alpha\frac{\mathbf{1}}{x_i} + \beta\right)\right)$$

$$-\frac{\mathbf{2}}{n^{\mathbf{2}}}\alpha X^{\mathbf{2}}\sum_{i=\mathbf{1}}^{n}\sum_{j>i}^{n}\frac{x_j - x_i}{x_i\,x_j}\left(\alpha\frac{\mathbf{1}}{x_i} + \beta\right)$$

- Since the second and fourth term are fixed, just consider the other terms, which are re-expressed as

$$\mathbf{2}\sum_{i=\mathbf{1}}^{n}\sum_{j>i}^{n}\frac{x_{ij}}{\pi_{ij}}$$

where $x_{ij} = ax_i^{g-1}x_j + \alpha(x_j - x_i)\left(\frac{\alpha}{x_i} + \beta\right)$

- The minimization of $\sum_{s\in S} E_\xi(v_{s,SYG})$ is equivalent to minimizing

$$\frac{2(N-2)!}{(n-2)!(N-n)!}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_{ij}}{\pi_{ij}}$$

where $x_{ij} = ax_i^{g-1}x_j + \alpha(x_j - x_i)\left(\frac{\alpha}{x_i} + \beta\right)$

- This is a nonlinear programming (NLP) problem. If $\alpha = 0$, the optimization problem is

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i^{g-1}x_j}{\pi_{ij}}$$

which depends on $g(\geq 0)$

- It is possible to achieve the desirable properties of the variance estimates by using the following constraints:

$$\sum_{j\neq i}^{N} \pi_{ij} = (n-\mathbf{1})\pi_i\,,\ i=1,\cdots,N\,.$$

$$c_{NLP}\ \pi_i\pi_j < \pi_{ij} \leq \pi_i\pi_j\,,\ \ j > i = 1,\cdots,N\,,$$

where c_{NLP} is a real number between 0 and 1

Concluding Remarks

- Raj's (1956) approach only uses a simple model, but it is useful in developing πps sampling designs

- Rao and Bayless (1969, 1970) uses the superpopulation model without the intercept term. Introducing the term into model may be desirable with respect to flexibility in sampling design. It does not result in any loss of model expectation of variance even though the model may pass through the origin

- If we have enough information about a population, the construction of the sampling designs for the minimization of model expectation of variance (or variance estimates) may be possible by using LP or NLP

- The structure of optimization problem to minimize the expected variance depends on the expression of the variance as well as model assumptions

- We may need careful simulation studies for a comparison between model-based πps sampling design and traditional πps sampling design under a diversity of superpopulation assumptions

- πps sampling designs under more complicated models such as a multiple regression model or a nonlinear model should be studied to see the relationships between models and designs