Compounded Design for Sample Survey

Wang Guoming
Research Institute of Statistical Sciences
National Bureau of Statistics
People's Republic of China
Email Address: wanggm@stats.gov.cn

Most survey involve not one but several study variables. In general, the cost collecting the data on these variables are different, so we can classify these variables into k categories by the unit cost. To simplify the discussion, let's suppose that there are $H+L$ variables: $\mathbf{y}_{h(\text{h=1,...,H})}$ and $\mathbf{x}_{l(\text{l=1,...,L})}$, high cost on $\mathbf{y}_h$ and low cost on $\mathbf{x}_l$, and the cost function is linear $C = C_0 + (C_H + C_L)n$, where, C_H is the unit cost to collect data on $\mathbf{y}_h$ and C_L the unit cost on $\mathbf{x}_l$. Now, we have two choices under total survey budget C: **a**. interview n units to collect data on all variables, or, **b**. interview $n_1(< n)$ units to collect data on all variables and $n_2(= (n-n_1)\frac{C_L+C_H}{C_L})$ units to collect data on $\mathbf{x}_l$ only. Which is the better?

For simple random cases, if we are interested in the population means, the estimators of choice **a** are $\begin{cases} \hat{\mathbf{x}}_l^a = \sum_{i=1}^{n} \frac{x_{l,i}}{n} \\ \hat{\mathbf{Y}}_h^a = \sum_{i=1}^{n} \frac{y_{h,i}}{n} \end{cases}$, those of choice **b** are $\begin{cases} \hat{\mathbf{x}}_l^b = \sum_{i=1}^{n_1+n_2} \frac{x_{l,i}}{n_1+n_2} \\ \hat{\mathbf{Y}}_h^b = \sum_{i=1}^{n_1} \frac{y_{h,i}}{n_1} + \sum_{l=1}^{L} d_{h,l}(\hat{X}_l^b - \sum_{i=1}^{n_1} \frac{x_{l,i}}{n_1}) \end{cases}$

Obviously, the variance of $\hat{\mathbf{x}}_l^b$ will not great than that of $\hat{\mathbf{x}}_l^a$, because $n_1 + n_2 > n$. What about the variance of $\hat{\mathbf{y}}_h^b$ compared with that of $\hat{\mathbf{y}}_h^a$? In most practical cases, we can get benefit in deduction of variance by choice **b**.

From above intuitive ideas, a *compounded sample design* is proposed. It is composed of two generalized designs $\mathbf{P}(s_1)$ and $\mathbf{P}(s_2|s_1)$. The process of a *compounded design* is first selecting a sample s_1 from the population U by $P(s_1)$, collecting the data on all variables; then selecting a sample s_2 from $U - s_1$ by $P(s_2|s_1)$, only collecting the data on $\mathbf{x}_l$ which are low cost. If we view $\mathbf{P}(s_2)$ as generated by $\mathbf{P}(s_2|s_1)$, i.e., $P(s_2) = \sum_{s_1 \subset U} P(s_2|s_1)$, then $\mathbf{P}(s_2)$ is a design in common sense. Therefore, in the sense of probability selecting a sample, a *compounded design* with size (n_1, n_2) is defined as combination of designs $\mathbf{P}(s_1)$ and $\mathbf{P}(s_2)$ with size n_1, n_2 respectively through conditional probabilities $P(s_2|s_1)$ or joint probabilities $P(s_1, s_2)$, if exist $P(s_2|s_1)$ or $P(s_1, s_2)$ such that $\mathbf{P}(s_1 + s_2)$ is a design. We denote *compounded design* as $\mathbf{P}(s_1 + s_2)$.

For variables $\mathbf{x}_l$, we have all the data in s_1 and s_2, so we can use sample $s_1 + s_2$ to construct the relevant estimators, such as $\hat{\mathbf{x}}_l^s = \sum_{i \in s_1+s_2} x_{l,i}/(\pi_i^{s_1} + \pi_i^{s_2})$. Although we only have the data of variables $\mathbf{y}_h$ in s_1, the sample data of $\mathbf{x}_l$ in $s_1 + s_2$ can be used as auxiliary information to construct relevant estimators for $\mathbf{y}_h$, for instance,

$\hat{\mathbf{y}}_h^s = \hat{\mathbf{y}}_h^{s_1} + \sum_{l=1}^{L} d_{h,l}(\hat{\mathbf{x}}_l^s - \hat{\mathbf{x}}_l^{s_1})$, where, $\hat{\mathbf{y}}_h^{s_1} = \sum_{i \in s_1} y_{h,i}/\pi_i^{s_1}$, $\hat{\mathbf{x}}_l^{s_1} = \sum_{i \in s_1} x_{l,i}/\pi_i^{s_1}$.

Theorem 1 Given *compounded design* $\mathbf{P}(s_1 + s_2)$

1.1 The variance of the estimator $\hat{\mathbf{y}}_h^s = \hat{\mathbf{y}}_h^{s_1} + \sum_{l=1}^{L} d_{h,l}(\hat{\mathbf{x}}_l^s - \hat{\mathbf{x}}_l^{s_1})$ is

$V(\hat{\mathbf{y}}_h^s) = Y_h' \Pi^{s_1} Y_h + D_h' A D_h - 2D_h' b_h$ if $D_h = (d_{h,1}, \ldots, d_{h,L})'$ is fixed beforehand.

where $\begin{cases} X = (X_1, \ldots, X_L) \\ \Pi^{s_1} = (\pi_{ij}^{s_1}/\pi_i^{s_1}\pi_j^{s_1} - 1), \Pi^{s} = (\pi_{ij}^{s}/\pi_i^{s}\pi_j^{s} - 1), \Pi^{s_1 s} = (\pi_{ij}^{s_{12}}/\pi_i^{s_1}\pi_j^{s} - 1) \\ \mathbf{A} = X'(\Pi^{s} + \Pi^{s_1} - \Pi^{s s_1} - \Pi^{s_1 s})X \\ b_h = X'(\Pi^{s_1} - \Pi^{s s_1})Y_h \end{cases}$

π^{s_1} and π^{s} refer to the inclusion probabilities with respect to design $\mathbf{P}(s_1)$ and $\mathbf{P}(s_1+s_2)$ respectively. $\pi_{ij}^{s_{12}}$ is the joint probability of unit i included in s_1 and unit j included in s_2.

1.2 The optimal coefficient vector $\mathbf{D}_h$ is $\mathbf{D}_h = (d_{h,l})_{L \times 1} = \mathbf{A}^{-1}\mathbf{b}_h$ and the corresponding variance is $V(\hat{\mathbf{Y}}_h^s) = Y_h' \Pi^{s_1} Y_h - \mathbf{b}_h' \mathbf{A}^{-1} \mathbf{b}_h$.

In simple random case, it can be showed that $V(\hat{\mathbf{Y}}_h^s) = N^2(\frac{1}{n_1} - \frac{1}{N})S_{y_h}^2 - N^2 \frac{n_2}{n_1(n_1+n_2)} \mathbf{r}_{y_h x}' \mathbf{R}_x^{-1} \mathbf{r}_{y_h x} S_{y_h}^2$, where, $\mathbf{r}_{y_h x} = (\rho_{x_l, y_h})_{L \times 1}$ and $\mathbf{R}_x = (\rho_{x_{l_1}, x_{l_2}})_{L \times L}$. This coincides with the regression estimator using auxiliary information under simple random design.

Evidently, variance deduction of a *compounded design* is related to the correlation degree between the study variables and the cost on $\mathbf{x}_l$. Now let's see under what kind of conditions that we can get benefit by *compounded design* and what is the optimal size (n_1, n_2).

Theorem 2 For any given simple random design with size n and total budget C, there exists a simple random *compounded design* with size (n_1, n_2) under same budget, such that:

2.1 The optimal size of $\mathbf{P}(s_1+s_2)$ for variable $\mathbf{y}_h$ are:

$n_1(h) = \frac{n}{1-C^*}(1+\sqrt{\frac{C^*}{1-C^*}\frac{\alpha_h}{1-\alpha_h}})^{-1}$, $n_2(h) = min\{\frac{n-n_1(h)}{C^*}, N-n_1(h)\}$

The variance of estimator $\hat{\mathbf{y}}_h^s = \sum_{i\in s_1}\frac{y_{h,i}}{n_1} + \sum_{l=1}^{L} d_{h,l}\{\sum_{i\in s_1+s_2}\frac{x_{l,i}}{n_1+n_2} - \sum_{i\in s_1}\frac{x_{l,i}}{n_1}\}$ is

$V_{opt}(\hat{Y}_h^s) = S_{y_h}^2\{\frac{1}{n}[\sqrt{(1-C^*)(1-\alpha_h)} + \sqrt{C^*\alpha_h}]^2 - \frac{1}{N}\}$, where, $C^* = \frac{C_L}{C_L+C_H}$, $\alpha_h = \mathbf{r}'_{y_h x}\mathbf{R}_x^{-1}\mathbf{r}_{y_h x}$

2.2 If and only if $C^* < \mathbf{r}'_{y_h x}\mathbf{R}_x^{-1}\mathbf{r}_{y_h x}$, then exist $n_1 < n$ such that

$V(\hat{\mathbf{Y}}_h^s) = (\frac{1}{n_1} - \frac{1}{N})S_{y_h}^2 - \frac{n_2}{n_1(n_1+n_2)}\mathbf{r}'_{y_h x}\mathbf{R}_x^{-1}\mathbf{r}_{y_h x}S_{y_h}^2 < (\frac{1}{n} - \frac{1}{N})S_{y_h}^2$

2.3 If and only if $C^* < \mathbf{r}'_{y_h x}\mathbf{R}_x^{-1}\mathbf{r}_{y_h x}$ holds for each $\mathbf{y}_h$, then exist $n_1 < n$ such that

$V(\hat{\mathbf{Y}}_h^s) < (\frac{1}{n} - \frac{1}{N})S_{y_h}^2$ holds for each $\mathbf{y}_h$.

When $C_L \to 0$, size solution(n_1, n_2) to compounded design is: $n_1 = n$ and $n_2 = N - n$. This situation means using auxiliary information $\mathbf{x}_l$ without cost to construct regression estimator for $\mathbf{y}_h$. In other words, the traditional regression estimator using auxiliary information is a special case of Theorem 2.

The Theorem 2 tells us following facts in simple random case.

With limited total budget C, we can interview n units collecting the data on all variables, the variance of relavant estimator is $V_1 = N^2(\frac{1}{n} - \frac{1}{N})S_y^2$. But we have another choice with same budget: interview $n_1(< n)$ units collecting the data on all variables and $n_2(= \frac{n-n_1}{C^*})$ units only collecting the data on $\mathbf{x}_l$, the variance of corresponding estimator is

$$V_2 = N^2(\frac{1}{n_1} - \frac{1}{N})S_{y_h}^2 - N^2\frac{n_2}{n_1(n_1+n_2)}\mathbf{r}'_{y_h x}\mathbf{R}_x^{-1}\mathbf{r}_{y_h x}S_{y_h}^2 < N^2(\frac{1}{n} - \frac{1}{N})S_{y_h}^2$$

The percentage variance deduction of the latter choice compared with the former is:

$$\frac{V_1 - V_2}{V_1} = \frac{N}{N-n}\{1 - [\sqrt{(1-C^*)(1-\mathbf{r}'\mathbf{R}^{-1}\mathbf{r})} + \sqrt{C^*\mathbf{r}'\mathbf{R}^{-1}\mathbf{r}}]^2\}$$

Following example gives us an objective view to see the benefit of compounded design:

$\left\{ \begin{array}{l} C^* = 0.1 \\ \frac{n}{N} = 0.1 \end{array} \right.$

$\mathbf{r}'\mathbf{R}^{-1}\mathbf{r}$	=	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\frac{V_1-V_2}{V_1}$	=	0%	2.2%	7.3%	14.0%	22.2%	31.8%	42.8%	55.6%	71.1%

Theorem 3 Given any fixed design $\mathbf{P}(s^*)(n_{s^*} = n)$ with certain restrictions on π_i and π_{ij}, there exists a *compounded design* $\mathbf{P}(s_1+s_2)$ generated by $\mathbf{P}(s^*)$, such that:

3.1 The variance of the estimator $\hat{\mathbf{x}}_l^s = \sum_{i\in s_1+s_2}\frac{x_{l,i}}{\pi^{s_1}+\pi^{s_2}}$ under $\mathbf{P}(s_1+s_2)$ is not greater than that of $\hat{\mathbf{x}}_l^{s^*} = \sum_{i\in s^*}\frac{x_{l,i}}{\pi_i}$ under $\mathbf{P}(s^*)$: $V(\hat{\mathbf{x}}_l^s) = X'\Pi^s X \le X'\Pi X$, where $\Pi = (\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j})$.

3.2 The optimal size of $\mathbf{P}(s_1+s_2)$ for variable $\mathbf{y}_h$ 1s:

$n_1(h) = \frac{n}{1-C^*}(1+\sqrt{\frac{C^*}{1-C^*}\frac{\beta_h}{1-\beta_h}})^{-1}$, $n_2(h) = min\{\frac{n-n_1(h)}{C^*}, N-n_1(h)\}$

The variance of estimator $\hat{\mathbf{y}}_h^s = \sum_{i\in s_1}\frac{y_{h,i}}{\pi_i^{s_1}} + \sum_{l=1}^{L} d_{h,l}\{\sum_{i\in s_1+s_2}\frac{x_{l,i}}{\pi_i^{s_1}+\pi_i^{s_2}} - \sum_{i\in s_1}\frac{x_{l,i}}{\pi_i^{s_1}}\}$ is:

$V(\hat{\mathbf{Y}}_h^s) = Y'\Pi Y + \frac{n}{n-1}Y'\sum_{(n-1)}Y[\frac{1}{n_1} - \frac{n_2}{n_1(n_1+n_2)}\beta_h - \frac{1}{n}]$

where, $\beta_h = (X'\sum_{(n-1)}Y_h)'(X'\sum_{(n-1)}X)^{-1}(X'\sum_{(n-1)}Y_h)/(Y_h'\sum_{(n-1)}Y_h)$. $\sum_{(n-1)} = (a_{ij})$, $a_{ii} = \frac{n-1}{\pi_i}$, $a_{ij} = -\frac{\pi_{ij}}{\pi_i\pi_j}$

3.3 If and only if $C^* < \beta(h)$ holds for each h, there exist $n_1 < n$ such that

$V(\hat{\mathbf{Y}}_h^s) \le Y'\Pi Y_h$ holds for each h.