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 is 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: $Y_h (h = 1, \ldots, H)$ and $X_l (l = 1, \ldots, L)$, high cost on $Y_h$ and low cost on $X_l$, 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 $Y_h$ and $C_L$ the unit cost on $X_l$. 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 $n_2 = (n - n_1)/C_H/C_L$ units to collect data on $X_l$. Which is the better?

For simple random cases, if we are interested in the population means, the estimators of choice $a$ are

\[ \bar{X} = \frac{\sum_{i=1}^{n_1} x_{ih} + \sum_{i=1}^{n_2} x_{il}}{n_1 + n_2}, \]

those of choice $b$ are

\[ \bar{X} = \frac{\sum_{i=1}^{n_1} x_{ih} + \sum_{i=1}^{n_2} x_{il} + \sum_{i=1}^{n_2} d_{h,i}(\hat{x}_{ih} - \sum_{i=1}^{n_1} x_{ih})}{n_1 + n_2}. \]

Obviously, the variance of $\bar{X}^b$ will not great than that of $\bar{X}^a$, because $n_1 + n_2 > n$. What about the variance of $\hat{y}_h^b$ compared with that of $\hat{y}_h^a$? 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 $P(s_1)$ and $P(s_2|s_1)$. 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 the data on $X_l$ which are low cost. If we view $P(s_2|s_1)$ as generated by $P(s_2|s_1)$, i.e., $P(s_2) = \sum_{s_1 \subset U} P(s_2|s_1)$, then $P(s_2)$ is a design in common sense. Therefore, in the sense of selecting a sample, a compounded design with size $(n_1, n_2)$ is defined as combination of designs $P(s_1)$ and $P(s_2)$ 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 $P(s_1 + s_2)$ is a design. We denote compounded design as $P(s_1 + s_2)$.

For variables $X_l$, 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 $\hat{x}_{ih}^s = \sum_{i \in s_1 + s_2} x_{ih}/(n_1 + \pi_{i}^s)$. Although we only have the data of variables $Y_h$ in $s_1$, the sample data of $X_l$ in $s_1 + s_2$ can be used as auxiliary information to construct relevant estimators for $Y_h$, for instance,

\[ \hat{y}_h^s = \sum_{i \in s_1} y_{ih}/\pi_{i}^s, \]

\[ \hat{x}_{ih}^s = \sum_{i \in s_1} x_{ih}/\pi_{i}^s. \]

**Theorem 1** Given compounded design $P(s_1 + s_2)$

1. The variance of the estimator $\hat{y}_h^s = \hat{y}_h^{s_1} + \sum_{i=1}^{n_1} d_{h,i}(\hat{x}_{ih}^s - \hat{x}_{ih}^{s_1})$ is

\[ \psi(\hat{y}_h^s) = \psi(\hat{y}_h^{s_1}) + \sum_{i=1}^{n_1} d_{h,i}^2 \psi(\hat{x}_{ih}^{s_1}) + 2 \sum_{i=1}^{n_1} d_{h,i} \psi(\hat{x}_{ih}^s - \hat{x}_{ih}^{s_1}) \]

where

\[ \psi(\hat{x}_{ih}^s) = x_{ih} + d_{h,i}^2 \psi(\hat{x}_{ih}^{s_1}) + 2 \sum_{i=1}^{n_1} d_{h,i} \psi(\hat{x}_{ih}^s - \hat{x}_{ih}^{s_1}), \]

\[ \pi_{i}^{s_1} \text{ and } \pi_{i}^{s} \text{ refer to the inclusion probabilities with respect to design } P(s_1) \text{ and } P(s_1 + s_2) \text{ respectively.} \]

1.2 The optimal coefficient vector $d_{h,i}$ is $d_{h,i} = (d_{h,i})_{L \times 1} = A^{-1}b_{h,i}$ and the corresponding variance is $\psi(\hat{y}_h^s) = \psi(\hat{y}_h^{s_1}) - b_{h,i}^T A^{-1} b_{h,i}$.

In simple random case, it can be showed that $\psi(\hat{y}_h^s) = N^2(\hat{y}_h^s - \frac{1}{N} \sum_{i=1}^{n_1} y_{ih}) + N^2(\hat{y}_h^{s_1} - \frac{1}{N} \sum_{i=1}^{n_1} y_{ih})^2 + R_x R_y + S_y$,

where, $R_{y,h} = (\rho_{y,h})_{L \times 1}$ and $R_x = (\rho_{x_1, x_2})_{L \times L}$. 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 $x_i$. 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 $P(s_1 + s_2)$ for variable $y_h$ are:

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

The variance of estimator $\hat{y}_h^* = \sum_{i \in S_1} \frac{y_h}{n_1} + \sum_{i \in S_2} d_{h,i} \left( \sum_{i \in S_1 + S_2} \frac{x_{i,j}}{1 + x_{i,j}} - \sum_{i \in S_1} \frac{x_{i,j}}{1 + x_{i,j}} \right)$ is

$$\nu_{s_1} \left( \hat{y}_h^* \right) = \frac{s^2_h}{C^*(1 - C^*)} \left( \frac{1}{\sum_{i \in S_1 + S_2} \frac{x_{i,j}}{1 + x_{i,j}}} \right)^2 + \frac{C^*}{C^* - \frac{d_h}{C^*}} - \frac{1}{1 - \frac{d_h}{C^*}}$$

where $C^* = \frac{C^*}{C^* + \sum_{i \in S_1 + S_2} \frac{x_{i,j}}{1 + x_{i,j}}} \cdot \frac{1}{\sum_{i \in S_1 + S_2} \frac{x_{i,j}}{1 + x_{i,j}}}$.

2.2 If and only if $C^* < r_{y_h}^* r_{y_h}^*$, then exist $n_1 < n$ such that

$$\nu_{s_1} \left( \hat{y}_h^* \right) = \frac{n_1}{n_1^2} \left( \frac{1}{1 - \frac{d_h}{C^*}} \right)^2 < \nu_{s_1} \left( \hat{y}_h^* \right)$$

2.3 If and only if $C^* < r_{y_h}^* r_{y_h}^*$ holds for each $y_h$, then exist $n_1 < n$ such that

$$\nu_{s_1} \left( \hat{y}_h^* \right) = \frac{n_1}{n_1^2} \left( \frac{1}{1 - \frac{d_h}{C^*}} \right)^2$$

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 $x_i$ without cost to construct regression estimator for $y_h$. 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 relevant estimator is $\nu_{s_1} = N^2 \left( \frac{1}{n} - \frac{1}{n^2} \right) s^2_h$. But we have another choice with same budget: interview $n_1(< n)$ units collecting the data on all variables and $n_2 = \left( \frac{n - n_1}{C^*} \right)$ units only collecting on $x_i$, the variance of corresponding estimator is

$$\nu_{s_1} = N^2 \left( \frac{1}{n_1} - \frac{1}{n_1^2} \right) s^2_h - N^2 \left( \frac{n_1}{n_1^2} \right) \sum_{i \in S_1} \frac{x_{i,j}}{1 + x_{i,j}} r_{y_h}^* r_{y_h}^* s^2_h < N^2 \left( \frac{1}{n} - \frac{1}{n^2} \right) s^2_h$$

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

$$\frac{\nu_{s_1} - \nu_{s_1}}{\nu_{s_1}} = \sqrt{1 - \left( \frac{C^*}{C^* - \frac{d_h}{C^*}} \right)^2}$$

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

$$\begin{align*}
C^* & = 0.1, & r_{y_h}^* & = 0.1, & 0.2, & 0.3, & 0.4, & 0.5, & 0.6, & 0.7, & 0.8, & 0.9
\end{align*}$$

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

3.1 The variance of the estimator $\hat{x}_i^* = \sum_{i \in S_1 + S_2} \frac{x_{i,j}}{\pi_i + \pi_{i,j}}$ under $P(s_1 + s_2)$ is not greater than that of $\hat{x}_i^* = \sum_{i \in S_1} \frac{x_{i,j}}{\pi_i}$ under $P(s^*)$: $\nu(x^*) = X^* \nu(x^*)$ for any $x^*$, where $n = \left( \frac{\sum x_{i,j}}{\pi_i} \right)$.

3.2 The optimal size of $P(s_1 + s_2)$ for variable $y_h$ is:

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

The variance of estimator $\hat{y}_h^* = \sum_{i \in S_1} \frac{y_h}{n_1} + \sum_{i \in S_2} d_{h,i} \left( \sum_{i \in S_1 + S_2} \frac{x_{i,j}}{1 + x_{i,j}} - \sum_{i \in S_1} \frac{x_{i,j}}{1 + x_{i,j}} \right)$ is

$$\nu_{s_1} \left( \hat{y}_h^* \right) = \frac{y_h}{n_1} + \sum_{i \in S_1} \frac{\pi_i}{\pi_i + \pi_{i,j}} Y' \sum_{i \in S_1 + S_2} \frac{x_{i,j}}{1 + x_{i,j}} \cdot \pi_i$$

where $\pi_h = (x' \sum_{i \in S_1} x_{i,j}) (x' \sum_{i \in S_1 + S_2} x_{i,j})^{-1} x' \sum_{i \in S_1 + S_2} y_h \pi_i \sum_{i \in S_1 + S_2} x_{i,j} \cdot \pi_i \sum_{i \in S_1 + S_2} x_{i,j}$.

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

$$\nu_{s_1} \left( \hat{y}_h^* \right) < \frac{n_1}{n_1^2} \left( \frac{1}{1 - \frac{d_h}{C^*}} \right)^2$$

holds for each $h$.}