Evaluating the MSE of the Sample Mean or Sample Inflated Total for Unequal Size Clustered Population Predictors of a Realized Random Cluster Means
Ed Stanek

Introduction

We evaluate the variance of the sample mean as a predictor of a primary sampling unit (PSU) mean. We also evaluate the variance of the inflated sample total as a predictor of a PSU total. The evaluation is in the context of two stage unequal size cluster sample.

PPS Sampling with Totals

We assume probability proportional to size (pps) second stage sampling to develop an unbiased predictor based on minimizing the expected value of the mean squared error, representing the common second stage sampling fraction as \( f \). Note that \( f = \bar{f}, \delta_f = f \bar{\delta} \), \( \tau_f = f \tau \), and \( \tau_{1-f} = (1-f)\tau \). As a result, assuming pps sampling, the collapsed mixed model simplifies to

\[
Y^* = X^*\bar{\tau} + Z^*T + E^*
\]

where \( X^* = \left( \begin{array}{c} f \\ 1-f \end{array} \right) \otimes I_N \), \( Z^* = \left( \begin{array}{c} f \\ 1-f \end{array} \right) \otimes I_N \), \( T = (T_1, T_2, \ldots, T_N)' \) and \( T_s = \sum_{i=1}^{N_s} U_{is} \tau_s - \bar{\tau} \) and

where \( \text{var}_{\bar{\xi}} \left( Y^* \right) = \sigma^2 \left( \begin{array}{c} f \\ 1-f \end{array} \right) \otimes P_N + \nu^2 \left( \begin{array}{c} 1 \\ -1 \\ 1 \end{array} \right) \otimes I_N \) with \( \sigma^2 = \frac{1}{N-1} \sum_{s=1}^{N} (\tau_s - \bar{\tau})^2 \),

\( \nu^2 = f(1-f)\sigma^2_{\nu} \) where \( \sigma^2_{\nu} = \frac{1}{N} \sum_{s=1}^{N_s} M_s \sigma^2_s \). [see c05ed18.doc]

A Simple Predictor of the PSU Total

We define a target random variable as a linear combination of the totals, \( P^*_i = g_i^c Y^* \). In particular, we define \( g_i^c = \mathbf{1}_2 \otimes e_i \) so that \( P^*_i \) is the total for PSU \( i \). Notice that \( g_i \) can not be stochastic. We develop the bias and MSE of a predictor that corresponds to the sample PSU total, divided by the common sampling fraction.

The predictor of the total \( P^*_i = g_i^c Y^* \) where \( g_i^x = \mathbf{1}_2 \otimes e_i \) that we consider is the simple inflated total given by \( \hat{P}^*_i = \tilde{g}_i^e Y^* \) where \( \tilde{g}_i^e = \left( \begin{array}{c} 1/\bar{f} \\ 0 \end{array} \right) \otimes e_i \). Notice that \( \tilde{g}_i^e - g_i^e = \left( \begin{array}{c} 1-\bar{f} \\ -f \end{array} \right) \otimes e_i \) and

\( \tilde{g}_i^e - g_i^e = \left( \begin{array}{c} 1-\bar{f} \\ -f \end{array} \right) \otimes e'_i \). As a result,
\[
E_{\tilde{y}_{\tilde{y}2}}\left(\tilde{P}^r - P^r\right) = (\tilde{g}^{i'} - g^{i''}) E_{\tilde{y}_{\tilde{y}2}}\left(Y^r\right) = \left(1 - \frac{f}{f'}\right) \otimes \epsilon^r_i \mid E_{\tilde{y}_{\tilde{y}2}}\begin{pmatrix} Y^r_1 \\ Y^r_2 \\ \vdots \\ Y^r_N \\ Y^r_{N+1} \\ \vdots \\ Y^r_{N+M} \end{pmatrix}
\]

where

\[
E_{\tilde{y}_{\tilde{y}2}}\left(Y^r_i\right) = \sum_{i=1}^{N} U_{is} \sum_{j=1}^{M} Y_{sj} = \sum_{i=1}^{N} U_{is} \sum_{j=m_f}^{m_f} \mu_s = \sum_{i=1}^{N} \left(1 - \frac{f}{f'}\right) M_s \mu_s
\]

Then

\[
E_{\tilde{y}_{\tilde{y}2}}\left(\tilde{P}^r - P^r\right) = \left(1 - \frac{f}{f'}\right) \otimes \epsilon^r_i \left[\begin{pmatrix} \tilde{\tau}_f \\ \tilde{\tau}_{1-f} \end{pmatrix} \otimes 1_N\right] = \frac{(1-f)\tilde{\tau}_f - f\tilde{\tau}_{1-f}}{f}
\]

When sampling is pps, \(\tilde{\tau}_f = f \frac{1}{N} \sum_{s=1}^{N} M_s \mu_s\) and \(\tilde{\tau}_{1-f} = (1-f) \frac{1}{N} \sum_{s=1}^{N} M_s \mu_s\). As a result, with pps sampling, \(E_{\tilde{y}_{\tilde{y}2}}\left(\tilde{P}^r - P^r\right) = (1-f) \frac{f - (1-f)}{f} \left(1 \sum_{s=1}^{N} M_s \mu_s\right) = 0\), and the predictor is unbiased.

The MSE of this predictor is given by
\[ \text{var}_{\tilde{Y}_{ii}} (\tilde{P}_i - P_i) = (\tilde{g}_i^* - g_i^*) \text{var}_{\tilde{Y}_{ii}} (\tilde{Y}^*) (\tilde{g}_i^* - g_i^*) \]
\[ = \sigma_i^2 \left( \frac{1-f}{f} - 1 \right) \left( \frac{f}{1-f} \right) \left( \frac{f}{1-f} \right) \odot \epsilon_i^P e_i \]
\[ + \nu^2 \left( \frac{1-f}{f} - 1 \right) \left( \frac{1-c_i}{1-c_i} - 1 \right) \left( \frac{1-f}{f} \right) \odot \epsilon_i^I e_i \]
\[ = \frac{\nu^2}{f^2} \]

Since \( \nu^2 = f(1-f)\sigma_{\epsilon_i}^2 \), \[ \text{var}_{\tilde{Y}_{ii}} (\tilde{P}_i - P_i) = \frac{(1-f)\sigma_{\epsilon_i}^2}{f} \]

**Simple Predictor of the PSU mean**

We define a target random variable as a linear combination of PSU means, \( P_i^* = g_i^*Y^* \). In particular, we define \( g_i^* = \left( \frac{c_i}{1-c_i} \right) \odot e_i \) so that \( P_i^* \) is the mean for PSU \( i \). Notice that \( g_i^* \) can not be stochastic. This implies that \( c_i \) must be constant for PSU \( i \), regardless as to which cluster occupies that position in a realization. We develop the bias and MSE of a predictor of \( P_i^* \) that corresponds to the sample PSU mean. The predictor of the mean is given by \( \tilde{P}_i^* = \tilde{g}_i^*Y^* \) where \( \tilde{g}_i^* = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \odot e_i \). Notice that \( \tilde{g}_i^* - g_i^* = \begin{pmatrix} 1-c_i \\ -(1-c_i) \end{pmatrix} \odot e_i \) and \( \tilde{g}_i^* - g_i^* = (1-c_i)(1 -1) \odot e_i' \). As a result,

\[ E_{\tilde{Y}_{ii}} (\tilde{P}_i^* - P_i^*) = (\tilde{g}_i^* - g_i^*) E_{\tilde{Y}_{ii}} (Y^*) \]
\[ = (1-c_i)(1 -1) \odot e_i' E_{\tilde{Y}_{ii}} \]

where
\[
E_{\frac{\tilde{Y}^*}{\tilde{\eta}, \tilde{\xi}}} (Y^\ast_{\tilde{\eta}, \tilde{\xi}}) = E_{\frac{Y^\ast_{\tilde{\eta}, \tilde{\xi}}}{\tilde{\eta}, \tilde{\xi}}} \left( \sum_{s=1}^{N} U_{is} \sum_{j=1}^{M_s} Y_{ij} \right)
= E_{\frac{\tilde{Y}^*}{\tilde{\eta}}} \left( \sum_{s=1}^{N} U_{is} \sum_{j=1}^{M_s} \frac{Y_{ij}}{M_s - m_s} \right)
= E_{\frac{\tilde{Y}^*}{\tilde{\eta}}} \left( \sum_{s=1}^{N} U_{is} \frac{1}{m_s} \sum_{j=1}^{M_s} \mu_s \right)
= E_{\frac{\tilde{Y}^*}{\tilde{\eta}}} \left( \sum_{s=1}^{N} U_{is} \mu_s \right)
= \frac{1}{N} \sum_{s=1}^{N} \mu_s
= \mu
\]

Then
\[
E_{\frac{\tilde{Y}^*}{\tilde{\eta}, \tilde{\xi}}} (P_{\tilde{\eta}}^* - P_{\tilde{\xi}}^*) = \left[ (1 - c_i) \right] (1 - 1) \otimes e_i^T \left[ \begin{array}{c} \mu \\ \mu \end{array} \right] \otimes I_N,
\]

illustrating that the predictor is unbiased for all choices of \( c_i \).

We evaluate the MSE of the predictor using the expression for the variance given by
\[
\text{var}_{\frac{\tilde{Y}^*}{\tilde{\eta}, \tilde{\xi}}} (Y^\ast_{\tilde{\eta}, \tilde{\xi}}) = \begin{pmatrix} \frac{\sigma_{\tilde{y}i}}{\tilde{m}} & 0 \\ \frac{\sigma_{\tilde{y}i}}{\tilde{m}} & \frac{\sigma_{\tilde{y}i}}{\tilde{M} - \tilde{m}} \end{pmatrix} \otimes I_N + \sigma_{\tilde{y}i}^2 J_2 \otimes I_N
\]

define \( \frac{\sigma_{\tilde{y}i}^2}{\tilde{M}} = \frac{1}{N} \sum_{s=1}^{N} \frac{\sigma_{i}^2}{M_s} \), \( \frac{\sigma_{\tilde{y}i}^2}{\tilde{m}} = \frac{1}{N} \sum_{s=1}^{N} \frac{\sigma_{i}^2}{m_s} \), \( \frac{\sigma_{\tilde{y}i}^2}{\tilde{M} - \tilde{m}} = \frac{1}{N} \sum_{s=1}^{N} \frac{\sigma_{i}^2}{M_s - m_s} \) and \( \sigma_{\tilde{y}i}^2 = \sigma_i^2 - \frac{\sigma_{\tilde{y}i}^2}{\tilde{M}} \). Then
\[
\text{var}_{\frac{\tilde{Y}^*}{\tilde{\eta}, \tilde{\xi}}} (P_{\tilde{\eta}}^* - P_{\tilde{\xi}}^*) = \left( \tilde{g}_{\tilde{\eta}}^* - g_{\tilde{\xi}}^* \right) \text{var}_{\frac{\tilde{Y}^*}{\tilde{\eta}}} (Y^\ast_{\tilde{\eta}}) \left( \tilde{g}_{\tilde{\eta}}^* - g_{\tilde{\xi}}^* \right)
\]
\[
= (1 - c_i)^2 \left[ \begin{array}{c} (1 - 1) \otimes e_i^T \\ 0 \end{array} \right] \otimes I_N \left[ \begin{array}{c} \frac{\sigma_{\tilde{y}i}}{\tilde{m}} \\ \frac{\sigma_{\tilde{y}i}}{\tilde{M} - \tilde{m}} \end{array} \right] \left[ \begin{array}{c} 1 \\ -1 \end{array} \right] \otimes e_i
\]
\[
\quad + \sigma_{\tilde{y}i}^2 (1 - c_i)^2 \left[ \begin{array}{c} (1 - 1) \otimes e_i^T \\ 0 \end{array} \right] \left[ \begin{array}{c} \frac{\sigma_{\tilde{y}i}}{\tilde{m}} \\ \frac{\sigma_{\tilde{y}i}}{\tilde{M} - \tilde{m}} \end{array} \right] \left[ \begin{array}{c} 1 \\ -1 \end{array} \right] \otimes e_i
\]
\[
\quad - \frac{\sigma_{\tilde{y}i}^2}{N} (1 - c_i)^2 \left[ \begin{array}{c} (1 - 1) \otimes e_i^T \\ 0 \end{array} \right] \left[ \begin{array}{c} \frac{\sigma_{\tilde{y}i}}{\tilde{m}} \\ \frac{\sigma_{\tilde{y}i}}{\tilde{M} - \tilde{m}} \end{array} \right] \left[ \begin{array}{c} 1 \\ -1 \end{array} \right] \otimes e_i
\]
\[
= (1 - c_i)^2 \left( \frac{\sigma_{\tilde{y}i}^2}{\tilde{m}} + \frac{\sigma_{\tilde{y}i}^2}{\tilde{M} - \tilde{m}} \right)
\]

or
\[
\text{var}_{\frac{\tilde{Y}^*}{\tilde{\eta}, \tilde{\xi}}} (P_{\tilde{\eta}}^* - P_{\tilde{\xi}}^*) = (1 - c_i)^2 \left( \frac{\sigma_{\tilde{y}i}^2}{\tilde{m}} + \frac{\sigma_{\tilde{y}i}^2}{\tilde{M} - \tilde{m}} \right).
\]