OPTIMAL ESTIMATION OF CHANGE IN SAMPLE SURVEYS

Gösta Forsman and Tomas Garås, Statistics Sweden, Stockholm

1. Introduction

This paper deals with the problem of efficient estimation of the mean and of the total when measurements from times t_1 and t_2 are available for the variable under study. We also consider estimation of change from t_1 and t_2, change being expressed either as a difference of means and totals or as a ratio of totals. These problems are often thoroughly examined in standard texts when the population is assumed to be infinite and composed of the same units at t_1 and t_2. See e.g. Cochran (1977) and Raj (1968).

Here, by contrast, we assume that
(i) The population is changeable, that is, units may have joined or left the population between t_1 and t_2.
(ii) The population is finite.

Linear minimum-variance unbiased (LMVU) estimators and their variances are presented. Results for the ratio estimator are presented also. Optimum fractions are given for the matched part of the sample taken at time t_2, with the corresponding minimal variances.

2. Basic concepts

We assume a situation where the mean and total of a variable under study are to be estimated at times t_1 and t_2 in a repetitive sample survey, and where the change between t_1 and t_2 is of special interest. The variable under study is called x at time t_1 and y at time t_2. The changeable population is called U at t_1 and U' at t_2, with size N and N' respectively.

A sample of n units drawn at time t is denoted s_t (n).

At time t_1 the sample design is as follows (see figure 1): n units are drawn with Simple Random Sampling (SRS) from the N units of U. Measurements are made of the variable x. Naturally, at this time, the population U' is unknown.

At time t_2 it is assumed that:
- Between t_1 and t_2, N_1 units have left U and N_2 units have joined it. The corresponding subpopulations are called U_1 and U_2, respectively.
- The new population U' contains $N' = N_{12} + N_2$ units where N_{12} is the number of units in U_{12}, that is the intersection of U and U'.
- For the sample of n drawn at t_1, n_1 units belong to U_1 and n_{12} to U_{12}.

At time t_2 the sampling design is as follows:
- The sample consists of n' units and is partitioned into three parts, namely

  n_{12m} (m = matched) units are drawn with SRS from s_{t_1} (n_{12})

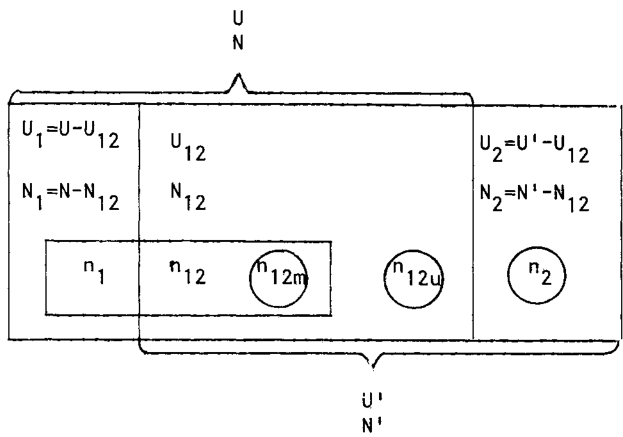
  n_{12u} (u = unmatched) units are drawn with SRS from $U_{12}-s_{t_1}$ (n_{12})

  n_2 units are drawn with SRS from the N_2 units of U_2.

The sample is drawn in a way that the following relations hold true:

$$n = n_1 + n_{12}$$
$$n' = n_{12m} + n_{12u} + n_2$$
$$n_{12} = n_{12m} + n_{12u}$$

Figure 1: The sampling design



The relative size of the unmatched part of the sample of U_{12} at t_2 is denoted

$$\Theta = \frac{n_{12u}}{n_{12}} \quad \text{which implies that } 1-\Theta = \frac{n_{12m}}{n_{12}}$$

The relative change of the population is denoted by Q_1 and Q_2, where

$$Q_1 = \frac{N_{12}}{N} \qquad 1 - Q_1 = \frac{N_1}{N}$$
$$Q_2 = \frac{N_{12}}{N'} \qquad 1 - Q_2 = \frac{N_2}{N'}$$

The population means, totals and variances for x and y in U_1, U_{12} and U_2 are defined by

$$\bar{X}_1 = \frac{\sum^{N_1} x_i}{N_1} \; ; \; X_1 = N_1\bar{X}_1 \quad ; \; S_1^2 = \frac{\sum^{N_1} (x_i-\bar{X}_1)^2}{N_1-1}$$

$$\bar{X}_{12} = \frac{\sum^{N_{12}} x_i}{N_{12}} ; \; X_{12} = N_{12}\bar{X}_{12} \; ; S_{12x}^2 = \frac{\sum^{N_{12}} (x_i-\bar{X}_{12})^2}{N_{12}-1}$$

$$\bar{Y}_{12} = \frac{\sum^{N_{12}} y_i}{N_{12}} ; \; Y_{12} = N_{12}\bar{Y}_{12} \; ; S_{12y}^2 = \frac{\sum^{N_{12}} (y_i-\bar{Y}_{12})^2}{N_{12}-1}$$

$$\bar{Y}_2 = \frac{\sum^{N_2} y_i}{N_2} \; ; \; Y_2 = N_2\bar{Y}_2 \quad ; \; S_2^2 = \frac{\sum^{N_2} (y_i-\bar{Y}_2)^2}{N_2-1}$$

Furthermore, it is assumed that

$$S_{12x}^2 = S_{12y}^2 = S^2$$

Then ρ, the coefficient of correlation between x and y in U_{12}, is

$$\rho = \frac{\text{Cov}(x,y)}{S^2} \Rightarrow \text{Cov}(x,y) = \rho S^2$$

where

$$\text{Cov}(x,y) = \frac{\sum^{N_{12}} (x_i-\bar{X}_{12})(y_i-\bar{Y}_{12})}{N_{12}-1}$$

3. Two useful theorems

The following two theorems will be useful in the derivation of the estimators and their variances.

Theorem 1: Consider a population U which is fixed over times t_1 and t_2 and contains M units. The following sampling design is assumed:

At t_1 a SRS of u units is drawn from the M units of U.

At t_2 a SRS of g units is drawn from $s_{t_1}(u)$ and a SRS of v-g units is drawn from $U-s_{t_1}(u)$

(Note that $s_{t_2}(v) = s_{t_2}(g) \cup s_{t_2}(v-g)$ is a SRS from U)

Assume also that x and y are variables under study at t_1 and t_2 respectively and that $\bar{x}_u$ and $\bar{y}_v$ are simple estimators of means and that $Cov(x,y) = \rho S^2$.

Then

$$Cov(a\bar{x}_u, b\bar{y}_v|g) = ab(\frac{g}{u\cdot v} - \frac{1}{M})\rho S^2$$

where a and b are arbitrary constants. This can be proved easily.

In our case u, v and M correspond to n_{12}, n_{12m}, n_{12u} and N_{12}. With theorem 1 covariances are easily calculated.

The next theorem is due to Rao, C.R. (1952). See also Raj (1968).

Theorem 2: A necessary and sufficient condition for T_0 to be the minimum-variance-unbiased estimator (MVU) of a parameter is that T_0 is unbiased and that $Cov(T_0,z) = 0$ for all z where z is a zero function, i.e., $E(z) = 0$.

Corollary 2.1: If T_0 is a MVU-estimator of a parameter and T_1 is an unbiased estimate of that parameter then $V(T_0) = Cov(T_0,T_1)$. This follows from Theorem 2 since

$$Cov(T_0,T_0-T_1) = 0 \Rightarrow Cov(T_0,T_0) = V(T_0) = Cov(T_0,T_1)$$

Corollary 2.2: Let α and β be arbitrary constants. If T_0'' and T_0' are MVU-estimators of the parameters P'' and P' respectively, then $D_0 = \alpha T_0'' - \beta T_0'$ is a MVU-estimator of $\alpha P'' - \beta P'$.

This follows from theorem 2 since

$$Cov(T_0'',z) = Cov(T_0',z) = 0 \Rightarrow$$

$$\Rightarrow Cov(D_0,z) = \alpha Cov(T_0'',z) - \beta\, Cov(T_0',z) = 0$$

Theorem 2.2 with corollaries will be used to derive linear MVU-estimators (LMVU) and their variances.

4. LMVU- and MV-estimators

Linear MVU-estimators (LMVU) and their variances have been derived for the following parameters:

a. $\bar{X} = \frac{\sum^{N} x_i}{N}$ and $X = N\bar{X}$ in U at t_1

b. $\bar{Y} = \frac{\sum^{N'} y_i}{N'}$ and $Y = N'\bar{Y}$ in U' at t_2

c. $\bar{Y} - \bar{X}$ and $Y - X$.

Minimum variance-estimator (MV) and its variance have been derived for the parameter

d. Y/X.

Optimum variances have been calculated by minimizing the variances of the LMVU- and MV-estimators with respect to $\Theta = \frac{n_{12u}}{n_{12}}$.

The method of calculating LMVU-estimators and their variances is, in principal, the same for the above parameters a-c. The techniques to derive the LMVU-, the MV-estimators and their variances are shown in Forsman and Garås (1982).

4.1. Results for the LMVU- and MV-estimators

LMVU- and MV-estimators, their variances and optimal variances are given in this section for the parameters X, $\bar{X}$, Y, $\bar{Y}$, Y-X, $\bar{Y}-\bar{X}$ and Y/X. Let us introduce the following concepts:

An LMVU-estimator, here considered, is denoted by e_{LMVU}. Then

$$V_{opt}(e_{LMVU}) = \min_{\Theta} V(e_{LMVU}) \text{ where } \Theta = \frac{n_{12u}}{n_{12}},\ 0 \le \Theta \le 1$$

The Θ-value minimizing $V(e_{LMVU})$ is denoted Θ_{min}.

If $\rho = \pm 1$ and $\Theta \to 1 \Rightarrow V(e_{LMVU}) \to V_{opt}(e_{LMVU})$.

$\rho = 0 \Rightarrow V_{opt}(e_{LMVU})$ is independent of Θ.

The formulas below are valid for $-1 < \rho < 1$.

The above mentioned about e_{LMVU} is also valid for the MV-estimator, although negative ρ is not considered here.

a. Estimator of total at t_1

$$\hat{X}_{LMVU} = \hat{X}_1 + \hat{X}_{12} - \frac{\rho(1-\Theta)\Theta}{1-\rho^2\Theta^2}\left[(\hat{Y}_{12m} - \hat{Y}_{12u}) + \rho(\hat{X}_{12} - \hat{X}_{12m})\right]$$

$$V(\hat{X}_{LMVU}) = N_1^2 S_1^2 (\frac{1}{n_1} - \frac{1}{N_1}) + N_{12}^2 \frac{S^2}{n_{12}}\left[\frac{1-\rho^2\Theta}{1-\rho^2\Theta^2} - \frac{n_{12}}{N_{12}}\right]$$

$$\Theta_{min} = \frac{1}{1+\sqrt{1-\rho^2}} \Rightarrow V_{opt}(\hat{X}_{LMVU}) = N_1^2 S_1^2 (\frac{1}{n_1} - \frac{1}{N_1}) + N_{12}^2 \frac{S^2}{n_{12}}\left[\frac{(1+\sqrt{1-\rho^2})}{2} - \frac{n_{12}}{N_{12}}\right]$$

b. Estimator of mean at t_1

$$\hat{\bar{X}}_{LMVU} = \hat{X}_{LMVU}/N$$

$$V(\hat{\bar{X}}_{LMVU}) = V(\hat{X}_{LMVU})/N^2 \ ;\ V_{opt}(\hat{\bar{X}}_{LMVU}) = V_{opt}(\hat{X}_{LMVU})/N^2$$

c. Estimator of total at t_2

$$\hat{Y}_{LMVU} = \hat{Y}_2 + \hat{Y}_{12u} + \frac{(1-\Theta)}{1-\rho^2\Theta^2}\left[(\hat{Y}_{12m} - \hat{Y}_{12u}) + \rho(\hat{X}_{12} - \hat{X}_{12m})\right]$$

$$V(\hat{Y}_{LMVU}) = N_2^2 S_2^2 (\frac{1}{n_2} - \frac{1}{N_2}) + N_{12}^2 \frac{S^2}{n_{12}}\left[\frac{1-\rho^2\Theta}{1-\rho^2\Theta^2} - \frac{n_{12}}{N_{12}}\right]$$

$$\Theta_{min} = \frac{1}{1+\sqrt{1-\rho^2}} \Rightarrow V_{opt}(\hat{Y}_{LMVU}) = N_2^2 S_2^2 (\frac{1}{n_2} - \frac{1}{N_2}) + N_{12}^2 \frac{S^2}{n_{12}}\left[\frac{(1+\sqrt{1-\rho^2})}{2} - \frac{n_{12}}{N_{12}}\right]$$

d. Estimator of mean at t_2

$$\hat{\bar{Y}}_{LMVU} = \hat{Y}_{LMVU}/N'$$

$$V(\hat{\bar{Y}}_{LMVU}) = V(\hat{Y}_{LMVU})/N'^2 ;\ V_{opt}(\hat{\bar{Y}}_{LMVU}) = V_{opt}(\hat{Y}_{LMVU})/N'^2$$

e. Estimator of difference between totals

$$(\hat{Y}-\hat{X})_{LMVU} = \hat{Y}_{LMVU} - \hat{X}_{LMVU} = \hat{Y}_2 - \hat{X}_1 + \hat{Y}_{12u} - \hat{X}_{12} + $$
$$+ \frac{(1-\Theta)}{1-\rho\Theta}\left[(\hat{Y}_{12m}-\hat{Y}_{12u}) + \rho(\hat{X}_{12} - \hat{X}_{12m})\right]$$

$$V(\hat{Y}-\hat{X})_{LMVU} = N_2^2S_2^2(\frac{1}{n_2} - \frac{1}{N_2}) + N_1^2S_1^2(\frac{1}{n_1} - \frac{1}{N_1}) +$$
$$+ \frac{2N_{12}^2S^2(1-\rho)}{n_{12}}\left[\frac{1}{1-\rho\Theta} - \frac{n_{12}}{N_{12}}\right]$$

$$\underset{0<\rho<1}{\Theta_{min}} = 0 \Rightarrow V_{opt}(\hat{Y}-\hat{X})_{LMVU} = N_2^2S_2^2(\frac{1}{n_2} - \frac{1}{N_2}) +$$
$$+ N_1^2S_1^2(\frac{1}{n_1} - \frac{1}{N_1}) + 2N_{12}^2S^2(1-\rho)\ (\frac{1}{n_{12}} - \frac{1}{N_{12}})$$

$$\underset{-1<\rho<0}{\Theta_{min}} = 1 \Rightarrow V_{opt}(\hat{Y}-\hat{X})_{LMVU} = N_2^2S_2^2(\frac{1}{n_2} - \frac{1}{N_2}) +$$
$$+ N_1^2S_1^2(\frac{1}{n_1} - \frac{1}{N_1}) + 2N_{12}^2\frac{S^2}{n_{12}}\left[1- \frac{n_{12}}{N_{12}}(1-\rho)\right]$$

f. Estimator of difference between means

$$(\hat{\bar{Y}}-\hat{\bar{X}})_{LMVU}=\hat{\bar{Y}}_{LMVU}-\hat{\bar{X}}_{LMVU}=(1-Q_2)\bar{y}_2-(1-Q_1)\bar{x}_1+Q_2\bar{y}_{12u}-$$
$$-Q_1\bar{x}_{12}+\frac{(1-\Theta)\ (Q_2+Q_1\rho\Theta)}{1-\rho^2\Theta^2}\left[(\bar{y}_{12m}-\bar{y}_{12u})+\rho(\bar{x}_{12}-\bar{x}_{12m})\right]$$

$$V(\hat{\bar{Y}}-\hat{\bar{X}})_{LMVU}=(1-Q_2)^2S_2^2(\frac{1}{n_2} - \frac{1}{N_2})+(1-Q_1)^2S_1^2(\frac{1}{n_1} - \frac{1}{N_1})+$$
$$+\frac{S^2}{n_{12}}\left[\frac{(Q_1^2+Q_2^2)(1-\rho^2\theta)-2Q_1Q_2\rho(1-\theta)}{1-\rho^2\theta^2} - \frac{n_{12}}{N_{12}}(Q_1^2+Q_2^2-2Q_1Q_2\rho)\right]$$

For the calculation of Θ_{min} we consider two cases.

Case 1: $0 < \rho < 1$

$$V(\Theta) = V(\hat{\bar{Y}}-\hat{\bar{X}})_{LMVU};\ \frac{dV(\Theta)}{d\Theta} = 0 \text{ gives}$$

$$\Theta = k\pm\sqrt{k^2-\ell};\ k=\frac{Q_1^2+Q_2^2-2Q_1Q_2\ \rho}{(Q_1^2+Q_2^2)\rho^2-2Q_1Q_2\rho}$$

$$\text{And } \ell = \frac{1}{\rho^2} \qquad (4.1.1)$$

It can be proved that $V(\Theta) < V(1)$ when $0 \leq \Theta < 1$ and $0 < \rho < 1$.
Then it follows that
a. If there exist a solution $\Theta\in[0,1[$ then $\Theta=\Theta_{min}$, which gives $V_{opt}(\hat{\bar{Y}}-\hat{\bar{X}})_{LMVU}$.
b. If there does not exist such a solution, then $\Theta_{min} = 0$.

Case 2: $-1 < \rho < 0$
According to (4.1.1) $\Theta = k\pm\sqrt{k^2-\ell}$
It can be proved that $V(\theta) > V(\Theta)$ when $0 < \Theta \leq 1$ and $-1 < \rho < 0$.
Then it follows that
a. If there exist a solution $\Theta \in]0,1]$ then $\Theta = \Theta_{min}$, which gives $V_{opt}(\hat{\bar{Y}}-\hat{\bar{X}})_{LMVU}$.
b. If there does not exist such a solution, then $\Theta_{min} = 1$.

g. Estimator of ratio between totals

$$\hat{R}_{MV} = \frac{\hat{Y}_{LMVU}}{\hat{X}_{LMVU}} =$$
$$= \frac{\hat{Y}_2+\hat{Y}_{12u} + \frac{(1-\Theta)}{(1-\rho^2\Theta^2)}\left[(\hat{Y}_{12m}-\hat{Y}_{12u})+\rho(\hat{X}_{12}-\hat{X}_{12m})\right]}{\hat{X}_1+\hat{X}_{12} - \frac{\rho\theta(1-\Theta)}{(1-\rho^2\Theta^2)}\left[(\hat{Y}_{12m}-\hat{Y}_{12u})+\rho(\hat{X}_{12}-\hat{X}_{12m})\right]}$$

$$V(\hat{R}_{MV}) \doteq \frac{1}{X^2}\Bigg[N_2^2S_2^2(\frac{1}{n_2}-\frac{1}{N_2})+R^2N_1^2S_1^2(\frac{1}{n_1} - \frac{1}{N_1}) -$$
$$- N_{12}^2\frac{S^2}{n_{12}}\left[\frac{n_{12}}{N_{12}}(1+R^2-2R\rho)\right] +$$
$$+ N_{12}^2\frac{S^2}{n_{12}}\left[\frac{(1+R^2)(1-\rho^2\Theta)-2R\rho(1-\Theta)}{(1-\rho^2\Theta^2)}\right]\Bigg]$$

For the calculation of Θ_{min} we here only consider $0 < \rho < 1$.

The minimum variance of $\hat{R}_{MV}$ is obtained below by minimizing $V(\hat{R}_{MV})$ with respect to Θ. This leads to a second degree equation of Θ.

Let $R > 0, 0 < \rho < 1$ and $V(\Theta) = V(\hat{R}_{MV})$.

Then $\frac{dV(\Theta)}{d\Theta} = 0$ leads to

$$\Theta = k\pm\sqrt{k^2-\ell};\ k = \frac{(1+R^2-2R\rho)}{\rho(\rho+\rho R^2-2R)} \text{ and } \ell = \frac{1}{\rho^2}$$

It can be proved that $V(\Theta) < V(1)$ when $0\leq\Theta<1$, $R>0$ and $0<\rho<1$. Then we can calculate optimum values of Θ.
a. If there exist a solution $\Theta\in[0,1[$ then $\Theta=\Theta_{min}$, which gives $V_{opt}(\hat{R}_{MV})$.

b. If there does not exist such a solution, then $\Theta_{min} = 0$.

4.2 Design comparisons

The variances of the LMVU-estimators of the design described above have been compared to the variances of the estimators of the same parameters under a simple design. This design is at t_1 a SRS of n units drawn from the N units of U and, at t_2, a SRS of n units from the N' units of U', independent of the sample at t_1. The following simplifying assumptions have been done: $Q_1=Q_2=Q$, $0\leq\rho<1$, $n_1=n_2$, $n_2/n=N_2/N$ and $n_{12}/n=N_{12}/N$. The population variances of U, U', U_1, U_{12} and U_2 are assumed equal and the sampling fractions are assumed ignorable.

The estimators for the simple design and their variances are shown in Forsman and Garås (1982).

In table 1 (levels) and table 2 (differences) the reductions of variances by using the matched design described in section 2 instead of the simple design, are expressed as the ratio

$$Z = \frac{V_{opt}(\text{LMVU-estimator, matched sample design})}{V(\text{estimator, simple design})}$$

Table 1: $Z=1-Q\rho$ (levels)

Q \ ρ	0.99	0.80	0.60	0.40	0.20
1	0.57	0.80	0.90	0.96	0.99
0.99	0.57	0.80	0.90	0.96	0.99
0.90	0.61	0.82	0.91	0.96	0.99
0.80	0.66	0.84	0.92	0.97	0.99
0.70	0.70	0.86	0.93	0.97	0.99
0.60	0.74	0.88	0.94	0.97	0.99
0.50	0.79	0.90	0.95	0.98	0.99

Table 2: $Z=(1-Q) + Q\frac{(1+\sqrt{1-\rho^2})}{2}$ (differences)

Q \ ρ	0.99	0.80	0.60	0.40	0.20
1	0.01	0.20	0.40	0.60	0.80
0.99	0.02	0.21	0.41	0.60	0.80
0.90	0.11	0.28	0.46	0.64	0.82
0.80	0.21	0.36	0.52	0.68	0.84
0.70	0.31	0.44	0.58	0.72	0.86
0.60	0.41	0.52	0.64	0.76	0.88
0.50	0.50	0.60	0.70	0.80	0.90

The results are following:

Large variance reductions can be achieved with the matched sample strategy for the estimators of change, even when the population is changeable. For the estimators of levels there are also variance reductions when using the matched sample strategy, but they are smaller.

5. Concluding remarks

The variances of the LMVU-estimators for levels and differences and of the MV-estimators for ratios are minimized with respect to the relative size of the overlapping part of the samples at t_1 and t_2. The optimum size of this part differs depending on whether the estimation concerns levels or changes. In practise the size chosen is a matter of judgment. As for the comparisons made between the sampling design and a simpler one with independent samples a numerical example shows that considerable variance reductions are possible. This is true even if the change between t_1 and t_2 is as large as 10 to 20 percent provided that the correlation coefficient in the overlapping part of the population is large. The variance reduction is especially large for estimation of change.

6. References

Cochran, W.G. (1977): Sampling Techniques.3rd, Wiley

Forsman, G. and Garås, T. (1982): Optimal estimation of change in Sample Surveys. Research report. Statistics Sweden.

Rao, C.R. (1952): Some Theorems on Minimum Variance Unbiased Estimation. Sankhyā, 12

Raj, D. (1968): Sampling Theory. McGraw-Hill.