Analytic Studies in Double Sampling*

by

Ing-tzer Wey

I. Introduction

The estimation of parameters and the design of sampling surveys depend on the possession of advance information about an auxiliary variate x which is highly correlated with the y variate under investigation. Ratio and regression estimations require a knowledge of the population mean $\bar{X}$ or total X. If it is desired to stratify the population according to the values of the x, their frequency distribution must be known. When such information is lacking, the usual procedure is to apply the technique known as two-phase or double sampling. The technique consists in taking a larger preliminary sample of size n' to estimate the population mean $\bar{X}$ or total X or the frequency distribution of x while a smaller sample of size n is often drawn from this preliminary sample to observe the y variate.

We shall use the data obtained from double sampling to estimate means and totals of certain subdivisions of the y variate. Such subdivisions(or subpopulations) have been termed as "domains of study" by the United Nations Subcommission on Sampling. The estimates thus obtained are then used for what is known as an "analytic study" in making

* This reseach was supported by National Science Council of the Republic of China

comparisons of means and totals of these domains. Various estimators of the domain mean and total will be derived in different sampling schemes, and their precisions will be compared.

II. Double Sampling with Equal Probability and without Replacement

(A) Simple Estimation of the Domain Total

A larger preliminary sample of size n' is taken from a population of size N by simple random sampling without replacement. Let the *J-th* domain contain N_J units, and let n_J' be the number of units in the preliminary sample that happen to fall into this domain. Since the value of N_J is usually not known, an unbiased estimator of N_J is provided by

$$\hat{N}_J = N n_J'/n' = N p_J$$

where p_J is an unbiased estimator of $P_J = N_J/N$.

A subsample of size n is drawn by simple random sampling without replacement from the preliminary sample in which the y-variates under study are observed. Let n_J be the number of units in the subsample that happen to fall into the *J-th* domain. Denote y_{Jk} by the observed value which belongs to the *J-th* domain. Then, the population total Y_J of y-variates in the *J-th* domain is estimated as follows:

$$_d\hat{Y}_J = \frac{\hat{N}_J}{n_J} \sum_{k=1}^{n_J} y_{Jk} = \hat{N}_J \bar{y}_J \qquad (1)$$

Theorem 1. The estimator $_d\hat{Y}_J$ is unbiased and its variance is given by

$$Var(_d\hat{Y}_J) = N^2\left(P_J + \frac{g'}{n'}Q_J\right)(1-f)\left(1+\frac{Q_J}{nP_J}\right)\frac{S_J^2}{n} + \frac{g'}{n'}N^2\bar{Y}_J^2 P_J Q_J \qquad (2)$$

where $g' = (N-n')/(N-1)$, $f = n/N$, $P_J = N_J/N$, $Q_J = 1-P_J$, $\bar{Y}_J$ is the

population mean of y-variates in the J-th domain and S_J^2 is the variance.

Proof: We first prove the unbiasedness as follows:

$$E({}_d\hat{Y}_J)=E[({}_d\hat{Y}_J \mid n_J')]=E[\hat{N}_J\bar{Y}_J]=N_J\bar{Y}_J=Y_J$$

since $\bar{y}_J$ is the mean of a simple random sample and since p_J is the proportion of the J-th domain in a simple random sample. The variance of ${}_d\hat{Y}_J$ can be obtained from the fact that the variance of an estimator is equal to the expectation of the conditional variance given an event plus the variance of the expectation given the event [see Hansen, et al (1953)], i.e.

$$Var({}_d\hat{Y}_J)=E[Var({}_d\hat{Y}_J \mid n_J')]+Var[E({}_d\hat{Y}_J \mid n_J')]$$

Now, from the familiar result of the variance of a proportion in a simple random sample, we have

$$Var[E({}_d\hat{Y}_J \mid n_J')]=Var[Np_J\bar{Y}_J]=N^2\bar{Y}_J^2Var(p_J)$$

$$=N^2\bar{Y}_J^2\frac{N-n'}{N-1}\frac{P_JQ_J}{n'}$$

$$=\frac{g'}{n'}N^2\bar{Y}_J^2P_JQ_J$$

Further, the conditional variance of $\bar{y}_J$, given n_J' is, to terms of order n^{-2} (see Wey, 1968),

$$Var(\bar{y}_J \mid n_J')=\frac{S_J^2}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)$$

and

$$E(\hat{N}_J^2)=N_J^2+Var(\hat{N}_J)$$

$$=N_J^2+N^2\frac{g'}{n'}P_JQ_J$$

Hence,

$$E[Var({}_d\hat{Y}_J \mid n_J')]=E[\hat{N}_J^2Var(\bar{y}_J \mid n_J')]$$

$$=E\left[\hat{N}_J^2(1-f)\left(1+\frac{Q_J}{nP_J}\right)\frac{S_J^2}{nP_J}\right]$$

$$=N^2\left(P_J^2+\frac{g'}{n'}P_JQ_J\right)(1-f)\left(1+\frac{Q_J}{nP_J}\right)\frac{S_J^2}{nP_J}$$

This completes the proof of the theorem.

Q. E. D.

(B) Comparison of Single and Double Sampling

Ignoring the terms in $(n'n)^{-1}$ and n^{-2} and assuming n/N and n'/N negligible in the variance of ${}_d\hat{Y}_J$ we have

$$Var({}_d\hat{Y}_J)=N^2P_J\left[\frac{S_J^2}{n}+\frac{\bar{Y}_J^2Q_J}{n'}\right] \tag{3}$$

This approximate expression for the variance is now minimized by choice of n and n' for a given cost of

$$C=n'c_1+nc_2 \tag{4}$$

where c_2 is usually large in relation to c_1. It is easily found that

$$\frac{n}{n'}=\frac{V_J}{\sqrt{Q_J}}\sqrt{\frac{c_1}{c_2}} \tag{5}$$

where $V_J=S_J/\bar{Y}_J$ is the coefficient of variation in the J-*th* domain. From equations (4) and (5), we have

$$n'=C\Big/\left[c_1+\frac{V_J}{\sqrt{Q_J}}\sqrt{c_1c_2}\right] \tag{6}$$

$$n=\frac{CV_J}{\sqrt{Q_J}}\sqrt{\frac{c_1}{c_2}}\Big/\left[c_1+\frac{V_J}{\sqrt{Q_J}}\sqrt{c_1c_2}\right] \tag{7}$$

Substituting these solutions into (3), we obtain the minimum variance as follows:

$$Var({}_d\hat{Y}_J)_{min}=N^2P_J\sqrt{Q_J\bar{Y}_J}\left(c_1+\frac{V_J}{\sqrt{Q_J}}\sqrt{c_1c_2}\right)$$

$$\cdot\left(S_J\sqrt{\frac{c_2}{c_1}}+\sqrt{Q_J\bar{Y}_J}\right)\Big/C \tag{8}$$

If all resources are devoted to a single simple random sample, this sample has size $n_s=C/c_2$, and an unbiased estimator of the *J-th* domain total Y_J is given by

$$\hat{Y}_J=\frac{N}{n_s}\sum_{k=1}^{m_J} y_{Jk} \tag{9}$$

where m_J is the number of units in the single sample that happen to fall into the *J-th* domain. Assuming n_s/N negligible, we have the variance of $\hat{Y}_J$ as follows [see Cochran (1963), p.37]:

$$Var(\hat{Y}_J)=\frac{N^2}{n_s}P_J(S_J^2+Q_J\bar{Y}_J^2) \tag{10}$$

$$=\frac{1}{C}N^2P_Jc_2(S_J^2+Q_J\bar{Y}_J^2) \tag{10^1}$$

Hence, double sampling gives a smaller variance if

$$S_J^2\left(\frac{1}{n_s}-\frac{1}{n}\right)+Q_J\bar{Y}_J^2\left(\frac{1}{n_s}-\frac{1}{n'}\right)>0$$

Substituting equations (6), (7) and $n_s=C/c_2$ into the above inequality and simplifying, we have

$$\frac{V_J}{\sqrt{Q_J}}>\frac{c_2-c_1}{2\sqrt{c_1c_2}} \tag{11}$$

The above inequality shows that for given c_1 and c_2 the ratio of the coefficient of variation in the *J-th* domain to the square root of the proportion of units that do not belong to the *J-th* domain must exceed a critical value before double sampling brings an increase in precision. We summarize the foregoing into the following theorem.

Theorem 2. In a double sample, the minimum variance of ${}_d\hat{Y}_J$ for a given cost of $C=n'c_1+nc_2$ is smaller than the variance of $\hat{Y}_J$ in a single simple random sample if

$$\frac{V_J}{\sqrt{Q_J}}>\frac{c_2-c_1}{2\sqrt{c_1c_2}}$$

where V_J is the coefficient of variation in the *J-th* domain and Q_J is the proportion of units that do not belong to the *J-th* domain. The comparison is made to terms of order n^{-1} and assume that n/N and n'/N are negligible.

(C) *Estimated Variance of the Simple Estimator for the Domain Total in Double Sampling.*

We now come to estimate the variance of ${}_d\hat{Y}_J$ in double sampling. Since in most applications $f=n/N$ will be negligible, we construct an unbiased estimator of the variance under this condition.

Theorem 3. If $f=n/N$ is negligible, then an unbiased estimator for $Var({}_d\hat{Y}_J)$ in double sampling is given by

$$var({}_d\hat{Y}_J)=N^2p_J^2\frac{s_J^2}{n_J}+\frac{N(N-n')}{n'-1}\left(\bar{y}_J^2-\frac{s_J^2}{n_J}\right)p_Jq_J \quad (12)$$

where

$$p_J=n_J'/n', \quad q_J=1-p_J, \quad \bar{y}_J=\sum_{k=1}^{n_J} y_{Jk}/n_J$$

and

$$s_J^2=\frac{1}{n_J-1}\sum_{k=1}^{n_J}(y_{Jk}-\bar{y}_J)^2$$

Proof: It is obvious that the conditional expected value of s_J^2 given n_J is S_J^2. Expanding n_J^{-1} around $E(n_J)=nP_J$ by Taylor's series and retaining to terms of order n^{-2}, we have the conditional expected value of n_J^{-1} with fixed p_J as follows (see Wey, 1968, p. 98)

$$\frac{1}{nP_J}\left(1+\frac{Q_J}{nP_J}\right)$$

Thus, the expected value of the first term in the right-hand side of expression (12) is

$$E\left[E\left(N^2 p_J^2 \frac{s_J^2}{n_J} \mid p_J\right)\right]$$

$$=E\left[N^2 p_J^2\left(1+\frac{Q_J}{nP_J}\right)\frac{S_J^2}{nP_J}\right]$$

$$=N^2\left(P_J^2+\frac{g'}{n'}P_J Q_J\right)\left(1+\frac{Q_J}{nP_J}\right)\frac{S_J^2}{nP_J}$$

Next, since the conditional expected value of $\bar{y}_J^2-s_J^2/n_J$ given n_J is $\bar{Y}_J^2$ the expected value of the second term in the right-hand side of expression (12) is

$$E\left[E\left\{N^2\left(\bar{y}_J^2-\frac{s_J^2}{n_J}\right)\frac{N-n'}{N}\frac{p_J q_J}{n'-1} \mid p_J\right\}\right]$$

$$=E\left[N^2\bar{Y}_J^2\frac{N-n'}{N}\frac{p_J q_J}{n'-1}\right]$$

$$=\frac{g'}{n'}N^2\bar{Y}_J^2 P_J Q_J$$

Thus,

$$E[var(_d\hat{Y}_J)]=N^2\left(P_J+\frac{g'}{n'}Q_J\right)\left(1+\frac{Q_J}{nP_J}\right)\frac{S_J^2}{n}+\frac{g'}{n'}N^2\bar{Y}_J^2 P_J Q_J$$

By comparison of the above expression with equation (2), we find that both two equations are the same under the assumption of n/N negligible.

Q. E. D.

(D) Double Sampling for Stratification

Suppose that the population of size N is to be stratified into L strata with respect to the values of an auxiliary variate x. The first sample of size n' is taken from the population by simple random sampling without replacement. On the basis of this the sample units are allocated to the L strata. Let

$W_h=N_h/N$: proportion of population falling into stratum h,

$w_h=n_h'/n'$: proportion of first sample falling into stratum h,

Then, w_h is an unbiased estimator for W_h. From n_h' a simple random subsample of size n_h is taken to collect information on y-variates. Let n_{Jh} be the number of units in the subsample that happen to fall into the J-th domain in stratum h. Denote y_{Jhi} by the observed valve which belongs to th J-th domain in stratum h.

(1) Estimation of the Domain Total.

Theorem 4. In double sample, an unbiased estimator for the J-th domain total Y_J is given by

$$ {}_d\hat{Y}_{Jst}=N\sum_{h=1}^{L}\frac{w_h}{n_h}\sum_{i=1}^{n_{Jh}}y_{Jhi} \tag{13}$$

If the values of the n_h do not depend on the w_h and assuming $1/N_h$ negligible, then the variance of ${}_d\hat{Y}_{Jst}$ is

$$\begin{aligned} Var({}_d\hat{Y}_{Jst}) = N^2\sum_{h=1}^{L}\left[W_h^2+\frac{g'}{n'}W_h(1-W_h)\right]\frac{(1-f_h)}{n_h} \\ \cdot P_{Jh}(S_{Jh}^2+Q_{Jh}\bar{Y}_{Jh}) \\ +\frac{g'}{n'}N^2\sum_{h=1}^{L}W_h(P_{Jh}\bar{Y}_{Jh}-P_J\bar{Y}_J)^2 \end{aligned} \tag{14}$$

where

$$g'=(N-n')/(N-1),\quad f_h=n_h/N_h,\quad P_{Jh}=N_{Jh}/N_h,$$
$$Q_{Jh}=1-P_{Jh},\quad P_J=N_J/N.$$

Proof: In order to prove the theorem, a new variate y_{hi}^* is defined on each unit in the h-th stratum as follows:

$$y_{hi}^*=\begin{cases} y_{Jhi} & \text{if the } i\text{-}th \text{ unit of stratum } h \text{ belongs} \\ & \text{to the } J\text{-}th \text{ domain,} \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Then, the estimator ${}_d\hat{Y}_{Jst}$ can be written as

$$ {}_d\hat{Y}_{Jst} = N \sum_{h=1}^{L} w_h \bar{y}_h^* $$

where $$ \bar{y}_h^* = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}^* = \frac{n_{Jh}}{n_h} \bar{y}_{Jh} $$

Now, $$ E({}_d\hat{Y}_{Jst}) = E\left[E(N \sum_{h=1}^{L} w_h \bar{y}_h^* \mid w_h)\right] $$

$$ = E\left[N \sum_{h=1}^{L} w_h \bar{Y}_h^*\right] $$

$$ = N \sum_{h=1}^{L} W_h Y_h^* $$

$$ = Y^* = Y_J $$

Thus, ${}_d\hat{Y}_{Jst}$ is an unbiased estimator for Y_J.

To find the variance, we have

$$ Var({}_d\hat{Y}_{Jst} \mid w_h) = N^2 \sum_{h=1}^{L} w_h^2 Var(\bar{y}_h^* \mid w_h) $$

$$ = N^2 \sum_{h=1}^{L} w_h^2 (1-f) \frac{S_h^{*2}}{n_h} \qquad (16) $$

where

$$ S_h^{*2} = \frac{1}{N_h - 1} \sum_{h=1}^{N_h} (y_{hi}^* - \bar{Y}_h^*)^2 $$

$$ = \frac{1}{N_h - 1}\left[\sum_{h=1}^{N_{Jh}} y_{Jhi}^2 - \frac{1}{N_h}\left(\sum_{h=1}^{N_{Jh}} y_{Jhi}\right)^2\right] $$

$$ = P_{Jh} S_{Jh}^2 + P_{Jh} Q_{Jh} \bar{Y}_{Jh}^2 . $$

Then the expected value of above conditional variance with respect to w_h is

$$E[Var({}_d\hat{Y}_{Jst}|w_h)]$$

$$= N^2 \sum_{h=1}^{L} \left[W_h^2 + \frac{g'}{n'} W_h (1 - W_h) \right] (1 - f_h) \frac{S_h^{*2}}{n_h} \qquad (17)$$

Next, the variance of the conditional expected value of ${}_d\hat{Y}_{Jst}$ given w_h is

$$Var[E({}_d\hat{Y}_{Jst}|w_h)]$$

$$= N^2 [\sum_{h=1}^{L} \bar{Y}_h^{*2} Var(w_h) + 2 \sum_{h<k} \bar{Y}_h^* \bar{Y}_k^* Cov(w_h, w_k)]$$

$$= \frac{g'}{n'} N^2 [\sum_{h=1}^{L} \bar{Y}_h^{*2} W_h (1 - W_h) - 2 \sum_{h<k} W_h W_k \bar{Y}_h^* \bar{Y}_k^*]$$

$$= \frac{g'}{n'} N^2 \sum_{h=1}^{L} W_h (\bar{Y}_h^* - \bar{Y}^*)^2$$

$$= \frac{g'}{n'} N^2 \sum_{h=1}^{L} W_h (P_{Jh} \bar{Y}_{Jh} - P_J \bar{Y}_J)^2 \qquad (18)$$

Adding equations (17) and (18), we obtain the variance of ${}_d\hat{Y}_{Jst}$ as is to be proved.

Q. E. D.

In a single stratified random sampling in which the W_h are known exactly, an unbiased estimator of the J-th domain total is given by

$$\hat{Y}_{Jst} = N \sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{h=1}^{n_{Jh}} y_{Jhi} = N \sum_{h=1}^{L_{Jh}} W_h \bar{y}_h^* \qquad (19)$$

and its variance is

$$Var(\hat{Y}_{Jst}) = N^2 \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^{*2}}{n_h}$$

$$= N^2 \sum_{h=1}^{L} W_h^2 (1 - f_h) P_{Jh} (S_{Jh}^2 + Q_{Jh} \bar{Y}_{Jh}^2) / n_h \qquad (20)$$

By comparing equation (14) with (20), we find that the effects of errors in double sampling for stratification are to increase slightly the within-stratum contribution and to introduce a between-stratum component.

(2) Estimated Variance of the Domain Total Estimate

Let $var({}_d\hat{Y}_{Jst})$ be an unbiased estimator of $Var({}_d\hat{Y}_{Jst})$ in (14). Using the $y_{hi}{}^*$ defined in expression (15) and putting

$$s_h{}^{*2} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi}{}^* - \bar{y}_h{}^*)^2$$

we see that $s_h{}^{*2}$ is an unbiased estimator of $S_h{}^{*2}$ in expression (17). It follows that an unbiased estimator of $E[Var({}_d\hat{Y}_{Jst} | w_h)]$ is

$$N^2 \sum_{h=1}^{L} w_h{}^2 (1 - f_h) \frac{s_h{}^{*2}}{n_h} \tag{21}$$

To obtain an unbiased estimator of $Var({}_d\hat{Y}_{Jst} | w_h)$ in (18), we start with the following expression:

$$\sum_{h=1}^{L} w_h (\bar{y}_h{}^* - \bar{y}_{st}{}^*)^2$$

$$= \sum_{h=1}^{L} w_h [(\bar{Y}_h{}^* - \bar{Y}) + (\bar{y}_h{}^* - \bar{Y}_h{}^*) - (\bar{y}_{st}{}^* - \bar{Y}^*)]^2$$

where

$$\bar{y}_{st}{}^* = \sum_{h=1}^{L} w_h \bar{y}_h{}^* = {}_d\hat{Y}_{Jst}/N.$$

$$E\Big[\sum_{h=1}^{L} w_h [\bar{y}_h{}^* - \bar{y}_{st}{}^*)^2]\Big]$$

$$= \sum_{h=1}^{L} W_h (\bar{Y}_h{}^* - \bar{Y}^*)^2 + \sum_{h=1}^{L} W_h (1 - f_h) \frac{S_h{}^{*2}}{n_h} - Var({}_d\hat{Y}_{Jst})/N^2$$

Thus, an unbiased estimator of $Var[E({}_d\hat{Y}_{Jst} | w_h)]$ is

$$\frac{g'}{n'}N^2\left[\sum_{h=1}^{L} w_h(\bar{y}_h^*-\bar{y}_{st}^*)^2-\sum_{h=1}^{L} w_h(1-f_h)\frac{s_h^{*2}}{n_h}\right]$$

$$+\frac{g'}{n'}var({}_d\hat{Y}_{Jst}) \qquad (22)$$

Therefore, an unbiased estimator of $Var({}_d\hat{Y}_{Jst})$ is equal to the sum of expressions (21) and (23), i. e.

$$var({}_d\hat{Y}_{Jst})$$

$$=N^2\sum_{h=1}^{L}\left[\left(w_h^2-\frac{g'}{n'}w_h\right)(1-f_h)\frac{s_h^{*2}}{n_h}+\frac{g'}{n'}w_h(\bar{y}_h^*-\bar{y}_{st}^*)^2\right]+\frac{g'}{n'}var({}_d\hat{Y}_{Jst})$$

Herce.

$$var({}_d\hat{Y}_{Jst})$$

$$=\frac{n'N(N-1)}{n'-1}\sum_{h=1}^{L}\left[\left(w_h^2-\frac{g}{n'}w_h\right)(1-f_h)\frac{s_h^{*2}}{n_h}+\frac{g'}{n'}w_h(\bar{y}_h^*-\bar{y}_{st}^*)^2\right]$$

We have thus proved the following theorem.

Theorem 5. An unbiased estimator of $Var({}_d\hat{Y}_{Jst})$ in (14) is

$$var({}_d\hat{Y}_{Jst})=\frac{n'N(N-1)}{n'-1}\sum_{h=1}^{L}\left[\left(w_h^2-\frac{g'}{n'}w_h\right)(1-f_h)\frac{s_h^{*2}}{n_h}\right.$$

$$\left.+\frac{g'}{n'}w_h(\bar{y}_h^*-\bar{y}_{st}^*)^2\right] \qquad (23)$$

where

$$s_h^{*2}=\frac{1}{n_h-1}\left[\sum_{i=1}^{n_{Jh}} y_{Jhi}^2-\frac{1}{n_h}(n_{Jh}\bar{y}_{Jh})^2\right]$$

$$\bar{y}_{st}^*=\sum_{h=1}^{L} w_h\bar{y}_h^*={}_d\hat{Y}_{Jst}/N,\quad \bar{y}_h^*=n_{Jh}\bar{y}_{Jh}.$$

(3) Estimation of the Domain Mean.

If N_J, the total number of units in the *J-th* domain is known, then an unbiased estimator of the domain mean $\bar{Y}_J$ is ${}_d\hat{Y}_{Jst}/N_J$. However, it

is usually not known the value of N_J. An unbiased estimator of N_J is given by

$$_d\hat{N}_{Jst}=N\sum_{h=1}^{L}w_hp_{Jh} \tag{24}$$

where

$$p_{Jh}=n_{Jh}/n_h$$

Hence, an estimator of the domain mean is

$$_d\hat{\bar{Y}}_{Jst}={}_d\hat{Y}_{Jst}/{}_d\hat{N}_{Jst}=\sum_{h=1}^{L}\frac{w_h}{n_h}\sum_{i=1}^{n_{Jh}}y_{Jhi}/\sum_{h=1}^{L}w_hp_{Jh} \tag{25}$$

In order to discuss the properties of the estimator, the following two variates are defined on each unit in stratum h.

$$y_{hi}^*=\begin{cases}y_{Jhi} & \text{if the } i\text{-th unit of stratum } h \text{ belongs to the } J\text{-}th \text{ domain,}\\ 0 & \text{otherwise.}\end{cases}$$

$$x_{hi}^*=\begin{cases}1 & \text{if the } i\text{-}th \text{ unit of stratum } h \text{ belongs to the } J\text{-}th \text{ domain,}\\ 0 & \text{otherwise.}\end{cases}$$

Then, the subsample means of these two variates are

$$\bar{y}_h^*=\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi}^*=\frac{1}{n_h}\sum_{i=1}^{n_{Jh}}y_{Jhi}=\frac{n_{Jh}}{n_h}\bar{y}_{Jh}$$

$$\bar{x}_h^*=\frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi}^*=\frac{n_{Jh}}{n_h}=p_{Jh}$$

so that the estimated domain mean may be written as

$$_d\hat{\bar{Y}}_{Jst}=\bar{y}_{st}^*/\bar{x}_{st}^*$$

where $\bar{y}_{st}^*$ and $\bar{x}_{st}^*$ are unbiased estimators of the population means $\bar{Y}^*$ and $\bar{X}^*$, respectively, made from a double sample, i. e.

$$\bar{y}_{st}^*=\sum_{h=1}^{L}w_h\bar{y}_h^*,\qquad \bar{x}_{st}^*=\sum_{h=1}^{L}w_h\bar{x}_h^*$$

Note that $\bar{X}^*$ is the proportion of units in the population that fall into the *J-th* domain, i. e.

$$\bar{X}^*=\frac{1}{N}\sum_{h=1}^{L} N_h\bar{X}_h^*=\frac{1}{N}\sum_{h=1}^{L} N_{Jh}=\frac{N_J}{N}=P_J$$

and the population mean $\bar{Y}^*$ is

$$\bar{Y}^*=\frac{1}{N}\sum_{h=1}^{L} N_h\bar{Y}_h^*=\frac{1}{N}\sum_{h=1}^{L} Y_{Jh}=P_J\bar{Y}_J$$

The estimator ${}_d\hat{\bar{Y}}_{Jst}$ is biased, since it is a ratio estimator. An upper bound for the ratio of the bias of the estimator to its standard deviation will be obtained in the following theorem.

Theorem 6. In a double sample, an upper bound for the ratio of the absolute value of the bias of ${}_d\hat{\bar{Y}}_{Jst}$ to its standard deviation $\sigma({}_d\hat{\bar{Y}}_{Jst})$ is given by

$$\frac{|Bias({}_d\hat{\bar{Y}}_{Jst})|}{\sigma({}_d\hat{\bar{Y}}_{Jst})}\leq\frac{1}{P_J}\left[\sum_{h=1}^{L}\left\{W_h^2+\frac{g'}{n'}W_h(1-W_h)\right\}\frac{N_h-n_h}{N_h-1}\frac{P_{Jh}Q_{Jh}}{n_h}\right.$$
$$\left.+\frac{g'}{n'}\sum_{h=1}^{L}W_h(P_{Jh}-P_J)^2\right]^{\frac{1}{2}} \qquad (26)$$

where

$$g'=(N-n)/(N-1),\ P_{Jh}=N_{Jh}/N_h,\ Q_{Jh}=1-P_{Jh}.$$

Proof: We begin with the covariance between ${}_d\hat{\bar{Y}}_{Jst}$ and $\bar{x}_{st}^*$.

$$Cov({}_d\hat{\bar{Y}}_{Jst},\bar{x}_{st}^*)=E({}_d\hat{\bar{Y}}_{Jst}\bar{x}_{st}^*)-E({}_d\hat{\bar{Y}}_{Jst})E(\bar{x}_{st}^*)$$
$$=\bar{Y}^*-E({}_d\hat{\bar{Y}}_{Jst})\bar{X}^*$$

Hence,

$$E({}_d\hat{\bar{Y}}_{Jst})=[\bar{Y}^*-Cov({}_d\hat{\bar{Y}}_{Jst},\bar{x}_{st}^*)]/\bar{X}^*$$
$$=\bar{Y}_J-\frac{1}{\bar{X}^*}Cov({}_d\hat{\bar{Y}}_{Jst},\bar{x}_{st}^*)$$

and

$$\frac{|Bias({}_d\hat{\bar{Y}}_{Jst})|}{\sigma({}_d\hat{\bar{Y}}_{Jst})} = \frac{|\rho\sigma({}_d\bar{x}_{st}^*)|}{\bar{X}^*} \leqq \frac{\sigma({}_d\bar{x}_{st}^*)}{\bar{X}^*}$$

The variance of $\bar{x}_{st}^*$ is given by (see Cochran, 1963, p, 229)

$$Var(\bar{x}_{st}^*)$$

$$= \sum_{h=1}^{L}\left[\left\{W_h^2+\frac{g'}{n'}W_h(1-f_h)\right\}\frac{N_h-n_h}{N_h-1}\frac{P_{Jh}Q_{Jh}}{n_h}+\frac{g'}{n'}W_h(P_{Jh}-P_J)^2\right]$$

and this completes the proof of the theorem.

Theorem 7. In a double sample, an approximate variance of ${}_d\hat{\bar{Y}}_{Jst}$ is, to terms of order n'^{-1},

$$Var({}_d\hat{\bar{Y}}_{Jst}) = \frac{1}{P_J^2}\sum_{h=1}^{L}\left[W_h^2+\frac{g'}{n'}W_h(1-W_h)\right]$$

$$\cdot\frac{(1-f_h)}{n_h}P_{Jh}[S_{Jh}^2+Q_{Jh}(\bar{Y}_{Jh}-\bar{Y}_J)^2]$$

$$+\frac{g'}{n'P_J^2}\sum_{h=1}^{L}W_hP_{Jh}^2(\bar{Y}_{Jh}-\bar{Y}_J)^2 \qquad (27)$$

Proof: The deviation of ${}_d\hat{\bar{Y}}_{Jst}$ from the true mean $\bar{Y}_J$ is

$${}_d\hat{\bar{Y}}_{Jst}-\bar{Y}_J = \frac{1}{\bar{x}_{st}^*}\sum_{h=1}^{L}w_h(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*)$$

Expanding $(\bar{x}_{st}^*)^{-1}$ around $\bar{X}^*$ by Taylor's series and retaining the first term in the expansion, we have

$${}_d\hat{\bar{Y}}_{Jst}-\bar{Y}_J = \frac{1}{\bar{X}^*}\sum_{h=1}^{L}w_h(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*)$$

To this order of approximation, the estimator is unbiased, since

$$E({}_d\hat{\bar{Y}}_{Jst}-\bar{Y}_J) = \frac{1}{P_J}\sum_{h=1}^{L}W_h(\bar{Y}^*{}_h-\bar{Y}_J\bar{X}_h^*)$$

$$= \frac{1}{N_J}\sum_{h=1}^{L}N_{Jh}(\bar{Y}_{Jh}-\bar{Y}_J) = 0$$

Now, it is easily verified that the variance of the conditional expected value of ${}_d\hat{Y}_{Jst}$, given w_h is

$$Var[E({}_d\hat{Y}_{Jst}|w_h)]=\frac{g'}{n'P_J{}^2}\sum_{h=1}^{L}W_h(\bar{Y}_h{}^*-\bar{Y}_J\bar{X}_h{}^*)^2$$

$$=\frac{g'}{n'P_J{}^2}\sum_{h=1}^{L}W_hP_{Jh}{}^2(\bar{Y}_{Jh}-\bar{Y}_J)^2 \qquad (28)$$

The conditional variance of the estimator, given w_h is

$$Var({}_d\hat{Y}_{Jst}|w_h)=\frac{1}{P_J{}^2}\sum_{h=1}^{L}w_h{}^2Var(\bar{y}_h{}^*-\bar{Y}_J\bar{x}_h{}^*) \qquad (29)$$

where

$$Var(\bar{y}_h{}^*-\bar{Y}_J\bar{x}_h{}^*)$$

$$=\frac{1-f_h}{n_h(N_h-1)}\sum_{i=1}^{N_h}[(y_{hi}{}^*-\bar{Y}_h{}^*)-\bar{Y}_J(x_{hi}{}^*-\bar{X}_h{}^*)]^2$$

$$=\frac{1-f_h}{n_h}P_{Jh}[S_{Jh}{}^2+Q_{Jh}(\bar{Y}_{Jh}-\bar{Y}_J)^2]$$

Taking the expectation of the conditional variance with respect to w_h, we have

$$E[Var({}_d\hat{Y}_{Jst}|w_h)]$$

$$=\frac{1}{P_J{}^2}\sum_{h=1}^{L}\left[W_h{}^2+\frac{g'}{n'}W_h(1-W_h)\right]Var(\bar{y}_h{}^*-\bar{Y}_J\bar{x}_h{}^*)$$

Thus, an approximate variance of ${}_d\hat{Y}_{Jst}$ is given by substituting the above results into the following expression:

$$Var({}_d\hat{Y}_{Jst})=E[Var({}_d\hat{Y}_{Jst}|w_h)]+Var[E({}_d\hat{Y}_{Jst}|w_h)] \qquad (30)$$

This completes the proof of the theorem.

Q. E. D.

(4) Estimated Variance of th Domain Mean Estimate.

The variance of $\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*$ in the equation (29) is approximately estimated by the following expression

$$var(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*)$$

$$=\frac{1-f_h}{n_h(n_h-1)}\sum_{i=1}^{n_h}[(y_{hi}^*-\bar{y}_h^*)-{}_d\hat{Y}_{Jst}(x_{hi}^*-\bar{x}_h^*)]^2$$

$$=\frac{1-f_h}{n_h-1}p_{Jh}[s_{Jh}^2+q_{Jh}(\bar{y}_{Jh}-{}_d\hat{Y}_{Jst})^2] \tag{31}$$

where

$$p_{Jh}=n_{Jh}/n_h, \quad q_{Jh}=1-p_{Jh}.$$

Thus, the first term in the right-hand side of the equation (30) can be approximately estimated by

$$\frac{1}{p_J^2}\sum_{h=1}^{L}w_h^2 var(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*) \tag{32}$$

where

$$p_J={}_d\hat{N}_{Jst}/N.$$

Next, the expression $(\bar{Y}_h^*-\bar{Y}_J\bar{X}_h^*)^2$ in the equation (28) can be estimated by

$$(\bar{y}_h^*-{}_d\hat{Y}_{Jst}\bar{x}_h^*)^2-var(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*)$$

$$=p_{Jh}(\bar{y}_{Jh}-{}_d\hat{Y}_{Jst})^2-var(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*)$$

Thus, an estimator of the variance of the conditional expected value of ${}_d\hat{Y}_{Jst}$, given w_h is

$$\frac{g'}{n'p_J^2}\sum_{h=1}^{L}w_h[p_{Jh}(\bar{y}_{Jh}-{}_d\hat{Y}_{Jst})^2-var(\bar{y}_h^*-\bar{Y}_J\bar{x}_h^*)] \tag{33}$$

Hence, adding equations (32) and (33) up, we have an estimated variance of the domain mean estimate as follows:

$$var({}_d\hat{Y}_{Jst})=\frac{1}{p_J{}^2}\sum_{h=1}^{L}\left(w_h{}^2-\frac{g'}{n'}w_h\right)var(\bar{y}_h{}^*-\bar{Y}_J\bar{x}_h{}^*)$$

$$+\frac{g'}{n'p_J{}^2}\sum_{h=1}^{L}w_hp_{Jh}(\bar{y}_{Jh}-{}_d\hat{Y}_{Jst})^2 \qquad (34)$$

where

$$g'=(N-n)/(N-1),\quad p_J=\sum_{h=1}^{L}w_hp_{Jh}.$$

(E) Ratio Estimation of the Domain Total

In a larger preliminary sample of size n' drawn from a population of size N by simple random sampling without replacement, we measure only the auxiliary variate x_i which is highly correlated with the variate y_i under study. In a smaller subsample of size n drawn from the preliminary sample by simple random sampling without replacement, we measure both x_i and y_i. Let n_J be the number of units in the subsample that happen to fall into the *J-th* domain and denote (x_{Jk}, y_{Jk}) by the observed pair which belongs to the *J-th* domain.

We define a new pair $(x_i{}^*, y_i{}^*)$ as (x_{Jk}, y_{Jk}) if the *i-th* unit of the population or sample belong to the *k-th* member of the domain *J*, and (0, 0) otherwise. Assume that the total number N_J of units belonging to the *J-th* domain in the population is not known, then an unbiased estimator of the domain total X_J of the auxiliary variate x_{Jk} obtained from the first sample is

$$\hat{X}_J=\frac{N}{n'}\sum_{i=1}^{n'}x_i{}^*=N\bar{x}_L{}^*$$

Now, the domain total Y_J of y_{Jk} variate is estimated by

$${}_d\hat{Y}_{JR}=\frac{\bar{y}_J}{\bar{x}_J}\hat{X}_J \qquad (35)$$

where $\bar{y}_J$ and $\bar{x}_J$ are subsample means of the *J-th* domain. In order to find an approximate bias and variance of the ratio estimator, the above expresssion can be written as follows:

$$ {}_d\hat{Y}_{JR}=\frac{\bar{y}^*}{\bar{x}^*}\hat{X}_J \tag{36}$$

where

$$\bar{y}^*=\frac{1}{n}\sum_{i=1}^{n} y_i^*=p_J\bar{y}_J,\quad p_J=n_J/n,$$

$$\bar{x}^*=\frac{1}{n}\sum_{i=1}^{n} x_i^*=p_J\bar{x}_J.$$

(1) An Approximate Bias of the Ratio Estimator

Let $\bar{e}_y^*=\bar{y}^*-\bar{Y}^*$, $\bar{e}_x^*=\bar{x}^*-\bar{X}^*$, and

$u=\hat{X}_J-X_J=N(\bar{x}_L^*-\bar{X}^*)$

then expression (36) can be written as

$$ {}_d\hat{Y}_{JR}=\frac{\bar{Y}^*+\bar{e}_y^*}{\bar{X}^*+\bar{e}_x^*}(X_J+u)$$

The bias of the estimator is given by

$$Bias({}_d\hat{Y}_{JR})=E({}_d\hat{Y}_{JR})-Y_J$$

$$=E\left[\frac{\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*+u\bar{e}_y^*}{\bar{X}^*+\bar{e}_x^*}\right]$$

Let

$$f(\theta)=E\left[\frac{\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*+u\bar{e}_y^*}{\bar{X}^*+\theta\,\bar{e}_x^*}\right]$$

Then, $Bias({}_d\hat{Y}_{JR})$ is equal to $f(\theta)$ evaluated at $\theta=1$. Now we shall find Taylor's expansion of $f(\theta)$ around $\theta=0$. This expansion is

$$f(\theta)=f(0)+\theta f'(0)+\frac{1}{2}\theta^2 f''(0)+\cdots\cdots \tag{37}$$

where

$$f(0)=\frac{1}{\bar{X}^*}E(\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*+u\bar{e}_y^*)$$

$$f'(0)=-\frac{1}{\bar{X}^{*2}}E[(\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*+u\bar{e}_y^*)\bar{e}_x^*]$$

The leading terms in the expansion are the first two terms which are of order n^{-1}. If only the first two terms in the expansion are retained, then an approximate bias is obtained as

$$Bias\,({}_d\hat{Y}_{JR})=\frac{1}{\bar{X}^*}E(u\bar{e}_y^*)-\frac{1}{\bar{X}^{*2}}E[(\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*)\bar{e}_x^*]$$

The term $E(u\bar{e}_x^*\bar{e}_y^*)$ in $f'(0)$ is ignored, since it is of order n^{-3}. It may be verified that

$$E(u\bar{e}_y^*)=N\ Cov(\bar{x}_L^*,\ \bar{y}^*)=N\left(1-\frac{n'}{N}\right)\frac{S_{xy}^*}{n'}$$

$$E(u\bar{e}_x^*)=N\ Cov(\bar{x}_L^*,\ \bar{x}^*)=N\left(1-\frac{n'}{N}\right)\frac{S_x^{*2}}{n'}$$

$$E(\bar{e}_x^*\bar{e}_y^*)=Cov\,(\bar{x}^*,\ \bar{y}^*)=\left(1-\frac{n}{N}\right)\frac{S_{xy}^*}{n}$$

$$E(\bar{e}_x^{*2})=Var(\bar{x}^*)=\left(1-\frac{n}{N}\right)\frac{S_x^{*2}}{n}$$

and

$$S_x^{*2}=P_J(S_{Jx}^2+Q\bar{X}_J^2)$$

$$S_y^{*2}=P_J(S_{Jy}^2+Q\bar{Y}_J^2)$$

$$S_{xy}^*=P_J(S_{Jxy}+Q_J\bar{X}_J\bar{Y}_J)$$

Hence, the approximate bias is

$$\begin{aligned}Bias({}_d\hat{Y}_{JR})&=\frac{N^2}{X_J}\left(\frac{1}{n}-\frac{1}{n'}\right)(R_JS_x^{*2}-S_{xy}^*)\\&=\frac{N}{\bar{X}_J}\left(\frac{1}{n}-\frac{1}{n'}\right)(R_JS_{Jx}^2-S_{Jxy})\\&=N\left(\frac{1}{n}-\frac{1}{n'}\right)\bar{Y}_J(C_{Jxx}-C_{Jxy})\end{aligned}\qquad(38)$$

where

$$C_{Jxx}=S_{Jx}^2/\bar{X}_J^2,\ C_{Jxy}/\bar{X}_J\bar{Y}_J.$$

We summarize the foregoing in the following theorem.

Theorem 8. In a double sample, an approximate bias of the ratio estimator ${}_d\hat{Y}_{JR}$ in (35) is given by

$$Bias\,({}_d\hat{Y}_{JR})=N\left(\frac{1}{n}-\frac{1}{n'}\right)\bar{Y}_J(C_{Jxx}-C_{Jxy}$$

We note that the bias of the ratio estimator becomes negligible as the sizes of the first and second sample are suffciently large.

(2) An Approximate Variance of the Ratio Estimator

We now come to the question of the precision of the ratio estimator. Since the estimator ${}_d\hat{Y}_{JR}$ is generally biased, its mean-squared error around Y_J would be more appropriate than the variance as a measure of its precision. However, since the bias is negligible as n' and n become large, the variance will be very close to the mean-squared error. We shall therefore consider the mean-squared error as the variance when the sample sizes n' and n are sufficiently large.

Theorem 9. In a double sample, an approximate variance of ${}_d\hat{Y}_{JR}$ is obtained under the assumption of both the first and second sample sizes being large enough to make the bias negligible as follows:

$$Var({}_d\hat{Y}_{JR})=\frac{1}{n}N^2(1-f)P_J(S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy})$$
$$+\frac{1}{n'}N^2(1-f')P_J(2R_JS_{Jxy}-R_J^2S_{Jx}^2+Q_J\bar{Y}_J^2) \quad (39)$$

where

$$f=n/N,\quad f'=n'/N,\quad P_J=N_J/N,$$
$$Q_J=1-P_J,\quad R_J=Y_J/X_J.$$

Proof: The mean-squared error of ${}_d\hat{Y}_{JR}$ around Y_J is

$$E({}_d\hat{Y}_{JR}-Y_J)^2=E\left[\frac{\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*+u\bar{e}_y^*}{\bar{X}^*+\bar{e}_x^*}\right]^2$$

Then we note that mean-squared error of ${}_d\hat{Y}_{JR}$ is the value at $\theta=1$ of the function

$$g(\theta)=E\left[\frac{\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*+u\bar{e}_y^*}{\bar{X}^*+\theta\bar{e}_x^*}\right]^2$$

Expanding the function $g(\theta)$ around $\theta=0$ by Taylor's series and retaining the leading term being the first term in the expansion, we have an approximate variance as

$$Var({}_d\hat{Y}_{JR})=\frac{1}{\bar{X}^{*2}}E(\bar{Y}^*u+X_J\bar{e}_y^*-Y_J\bar{e}_x^*)^2$$

where the term $u\bar{e}_y^*$ is ignored. Using the results obtained in the preceding section, we have after simplification,

$$\begin{aligned}Var({}_d\hat{Y}_{JR})&=\frac{1}{n}N^2(1-f)(S_y^{*2}+R_J^2S_x^{*2}-2R_JS_{xy}^*)\\&\quad+\frac{1}{n'}N^2(1-f')(2R_JS_{xy}^*-R_J^2S_x^{*2})\\&=\frac{1}{n}N^2(1-f)P_J(S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy})\\&\quad+\frac{1}{n'}N^2(1-f')P_J(2R_JS_{Jxy}-R_J^2S_{Jx}^2+Q_J\bar{Y}_J^2)\end{aligned}$$

This completes the proof of the theorem.

Q. E. D.

Now let us consider the condition under which the ratio estimator ${}_d\hat{Y}_{JR}$ will be superior to the simple estimator ${}_d\hat{Y}_J$. We assume that the sample sizes n' and n are sufficiently large so as to make the bias of ${}_d\hat{Y}_{JR}$ negligible. Then, the approximate variance of ${}_d\hat{Y}_{JR}$ given in (39) is valid. Hence, the estimator ${}_d\hat{Y}_{JR}$ will give a more precise result whenever

$$Var({}_d\hat{Y}_J) > Var({}_d\hat{Y}_{JR})$$

Ignoring the terms of order n^{-2} in expression (2), we have

$$N^2P_J(1-f)\frac{S_{Jy}^2}{n}+\frac{1}{n'}(1-f')N^2P_JQ_J\bar{Y}_J^2$$

$$> \frac{1}{n}N^2(1-f)P_J(S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy})$$

$$+\frac{1}{n'}N^2(1-f')P_J(2R_JS_{Jxy}-R_J^2S_{Jx}^2+Q_J\bar{Y}_J^2)$$

or

$$\rho_J > \frac{R_JS_{Jx}}{2S_{Jy}} = \frac{CV(x_{Jk})}{2CV(y_{Jk})}$$

where ρ_J is the correlation coefficient between y_{Jk} and x_{Jk} in the domain J. Note that this result is just the same as that obtained in a single simple random sampling (see Wey, 1970).

(3) Comparison of Double Sampling for Ratio Estimation and Single Sampling

Assuming the finite population correction terms negligible, we write the variance of the ratio estimator of the domain total as

$$Var({}_d\hat{Y}_{JR}) = N^2P_J\left(\frac{V_2}{n}+\frac{V_1}{n'}\right) \tag{40}$$

where

$$V_2 = S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy}$$

$$V_1 = 2R_JS_{Jxy}-R_J^2S_{Jx}^2+Q_J\bar{Y}_J^2$$

Using the simple cost function $C=n'c_1+nc_2$, it is easily found that

$$\frac{n}{n'} = \left(\frac{V_2}{V_1}\frac{c_1}{c_2}\right)^{\frac{1}{2}} \tag{41}$$

For this subsampling rate, the minimum variance is

$$Var({}_d\hat{Y}_{JR})_{min} = \frac{1}{C}(\sqrt{c_1V_1}+\sqrt{c_2V_2})^2 \tag{42}$$

If a single simple random sample of size $n_s=C/c_2$ is taken at the same cost, the variance of the simple estimator of the domain total is given by equation (10),

$$Var(\hat{Y}_J) = \frac{1}{C}N^2P_Jc_2(S_{Jy}^2+Q_J\bar{Y}_J^2)$$

$$= \frac{1}{C}N^2P_Jc_2(V_1+V_2)$$

Thus the condition under which the double sampling for ratio estimation is better than the single sampling for simple estimation at the same cost is

$$(\sqrt{c_1V_1}+\sqrt{c_2V_2})^2 < c_2(V_1+V_2)$$

or

$$\frac{V_1}{V_2} > \frac{4c_1c_2}{(c_2-c_1)^2} \tag{43}$$

We have therefore proved the following theorem.

Theorem 10. In a double sample, the minimum variance of the ratio estimator of the domain total for a given cost of $C=n'c_1+nc_2$ is smaller than the variance of the simple estimator in a single simple random sample of size C/c_2 if

$$\frac{V_1}{V_2} > \frac{4c_1c_2}{(c_2-c_1)^2}$$

where

$$V_1=2R_JS_{Jxy}-R_J^2S_{Jx}^2+Q_J\bar{Y}_J^2$$

is the quantity contributed by the first sample and

$$V_2=S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy}$$

by the second.

(F) Ratio Estimation of the Domain Mean

Let n_J' be the number of units in the first sample of size n' that happen to fall into the *J-th* domain and let $\bar{x}_J'$ denote the first sample mean of this domain. Then, a ratio esitmator of the domain mean $\bar{Y}_J$ is

$$_d\hat{\bar{Y}}_{JR}=\frac{\bar{y}_J}{\bar{x}_J}\bar{x}_J'=\hat{R}_J\bar{x}_J' \tag{44}$$

where $\bar{y}_J$ and $\bar{x}_J$ are subsample means of the *J-th* domain.

Let $e_x=\bar{x}_J-\bar{X}_J$, $e_y=\bar{y}_J-\bar{Y}_J$, and $u_x=\bar{x}_J'-\bar{X}_J$, then expression (44) can be written as

$$_d\hat{\bar{Y}}_{JR}=\frac{\bar{Y}_J+e_y}{\bar{X}_J+e_x}(\bar{X}_J+u_x)$$

The deviation of $_d\hat{\bar{Y}}_{JR}$ from $\bar{Y}_J$ is

$$_d\hat{\bar{Y}}_{JR}-\bar{Y}_J=\frac{\bar{Y}_Ju_x+\bar{X}_Je_y-\bar{Y}_Je_x+e_yu_x}{\bar{X}_J+e_x}$$

Now, the mean-squared error of $_d\hat{\bar{Y}}_{JR}$ around $\bar{Y}_J$ is the value at $\theta=1$ of the function

$$g(\theta)=E\left[\frac{\bar{Y}_Ju_x+\bar{X}_Je_y-\bar{Y}_Je_x+e_yu_x}{\bar{X}_J+\theta e_x}\right]$$

Expanding the function $g(\theta)$ around $\theta=0$ by Taylor's series and retaining the leading term being the first term in the expansion, we have an approximate variance as

$$Var(_d\hat{\bar{Y}}_{JR})=\frac{1}{\bar{X}_J^2}E(\bar{Y}_Ju_x+\bar{X}_Je_y-\bar{Y}_Je_x)^2 \tag{45}$$

where the term e_yu_x is ignored, since it is of order n^{-2} as compared to the other terms of order n^{-1}. Taking the expectation term by term in (45), we have (see Wey, 1968),

$$E(u_x^2)=Var(\bar{x}_J')=\frac{S_{Jx}^2}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)$$

where

$$f'=n'/N,\ P_J=N_J/N,\ Q_J=1-P_J$$

$$E(e_y^2)=Var(\bar{y}_J)=\frac{S_{Jy}^2}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)$$

where

$$f=n/N.$$

$$E(e_x^2)=Var(\bar{x}_J)=\frac{S_{Jx}^2}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)$$

$$E(u_xe_y)=Cov(\bar{x}_J',\ \bar{y}_J)=\frac{S_{Jxy}}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)$$

$$E(u_xe_y)=Cov(\bar{x}_J',\ \bar{x}_J)=\frac{S_{Jx}^2}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)$$

$$E(e_xe_y)=Cov(\bar{x}_J,\ \bar{y}_J)=\frac{S_{Jxy}}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)$$

Substituting the above results into (45), we have

$$Var({}_d\hat{\bar{Y}}_{JR})=\frac{1}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)(S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy})$$
$$+\frac{1}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)(2R_JS_{Jxy}-R_J^2S_{Jx}^2) \quad (46)$$

We have thus proved the following theorem.

Theorem 11. In a doudle sample, under the assumpion of both the first and second sample sizes being large enough to make the bias negligible, an approximate variance of the ratio estimator ${}_d\hat{\bar{Y}}_{JR}$ of the domain mean $\bar{Y}_J$ is

$$Var({}_d\hat{\bar{Y}}_{JR})=\frac{1}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)(S_{Jy}^2+R_J^2S_{Jx}^2-2R_JS_{Jxy})$$
$$+\frac{1}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)(2R_JS_{Jxy}-R_J^2S_{Jx}^2)$$

where $f=n/N,\ f'=n'/N,\ P_J=N_J/N,\ Q_J=1-P_J,$

and $R_J=\bar{Y}_J/\bar{X}_J.$

An approximate estimated variance of ${}_d\hat{\bar{Y}}_J$ is given by

$$var({}_d\hat{\bar{Y}}_{JR}) = \frac{1-f}{n_J}(s_{Jy}^2 + \hat{R}_J^2 s_{Jx}^2 - 2\hat{R}_J s_{Jxy}) + \frac{1-f'}{n_J'}(2\hat{R}_J s_{Jxy} - \hat{R}_J^2 s_{Jx}^2) \qquad (47)$$

where s_{Jy}^2, s_{Jx}^2 and s_{Jxy} are unbiased estimators of their respect parameters.

(G) Multivariate Ratio Estimation of the Domain Mean

The discussion so far has been restricted to the situation in which only one auxiliary x-variate is used for improving the precision of estimates. We shall now extend the ratio estimation to the situation in which p auxiliary x-variates $(x_{1Jk}, x_{2Jk}, \cdots\cdots, x_{pJk})$ are available. From each x-variate, we first construct a ratio estimator for the *J-th* domain mean $\bar{Y}_J$ and then combine them by using appropriate weights. This approach was first considered by Olkin (1958) to derive an approximate minimum variance of the multivariate ratio estimator in a single simple random sampling. In a double sample, let ${}_d\hat{\bar{Y}}_{JMR}$ denote a multivariate ratio estimator of the *J-th* domain mean $\bar{Y}_J$, then

$${}_d\hat{\bar{Y}}_{JMR} = \sum_{i=1}^{p} w_i \hat{\bar{Y}}_{JRi} \qquad (47)$$

$$\hat{\bar{Y}}_{JRi} = \frac{\bar{y}}{\bar{x}_{Ji}}\bar{x}_{Ji}' = \hat{R}_{Ji}\bar{x}_{Ji}'$$

where $\bar{x}_{Ji}'$ is the first sample mean of the *i-th* auxiliary x-variate in the *J-th* domain, $\bar{y}_J$ and $\bar{x}_{Ji}$ are subsample means of the domain J. The w_i are weights to be determined to maximize the precision of the estimator, subject to $\sum w_i = 1$.

(1) The Optimum Weights of the Multivariate Ratio Estimator

It is clear that the multivariate ratio estimator is, in general, biased. Each estimator $\hat{Y}_{JRi}$ has bias of order n^{-1}. The bias of the estimator ${}_d\hat{Y}_{JMR}$ is also of order n^{-1} since it is the weighted mean of the biases of estimators $\hat{Y}_{JRi}$. Thus, we assume that the sample size be large enough to make the bias negligible. Under this assumption, the variance of ${}_d\hat{Y}_{JMR}$ is

$$Var({}_d\hat{Y}_{JMR})=\sum_{h=1}^{p}\sum_{i=1}^{p}w_hw_iCov(\hat{Y}_{JRh},\ \hat{Y}_{JRi})$$

where the covariance between $\hat{Y}_{JRh}$ and $\hat{Y}_{JRi}$ can be obtained from expressien (46) as follows:

$$\begin{aligned}Cov(\bar{Y}_{JRh},\ \hat{Y}_{JRi})&=\frac{1}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)(S_{Jy}{}^2+R_{Jh}R_{Ji}S_{Jhi}-R_{Jh}S_{Jyh}-R_{Ji}S_{Jyi})\\&+\frac{1}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)(R_{Jh}S_{Jyh}+R_{Ji}S_{Jyi}-R_{Jh}R_{Ji}S_{Jhi})\end{aligned}$$

where S_{Jhi} is the covariance between the h-th and i-th auxiliary x-variates, and S_{Jyh} is the covariance between y-variate and the h-th auxiliary x-variate in the domain J. Thus, the variance of ${}_d\hat{Y}_{JMR}$ is

$$\begin{aligned}Var({}_d\hat{Y}_{JMR})=&\frac{\bar{Y}_J{}^2}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)\sum_{h=1}^{p}\sum_{i=1}^{p}w_hw_ia_{hi}\\&+\frac{\bar{Y}_J{}^2}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)\sum_{h=1}^{p}\sum_{i=1}^{p}w_hw_ib_{hi}\end{aligned}\quad(48)$$

where

$$a_{hi}=C_{Jyy}-C_{Jyh}-C_{Jyi}+C_{Jhi}$$

$$b_{hi}=C_{Jyh}+C_{Jyi}-C_{Jhi}$$

$$C_{Jyy}=S_{Jy}^2/\bar{Y}_J^2,\quad C_{Jyh}=S_{Jyh}/\bar{Y}_J\bar{X}_{Jh}$$

$$C_{Jhi}=S_{Jhi}/\bar{X}_{Jh}\bar{X}_{Ji}$$

In order to simplify the expression for the derivation of optimum weights, we shall write expression (48) in the matrix notation as follows:

$$Var(_d\hat{\bar{Y}}_{JMR})=\bar{Y}^2{}_J(K_2\boldsymbol{w}'\boldsymbol{A}\boldsymbol{w}+K_1\boldsymbol{w}'\boldsymbol{B}\boldsymbol{w})$$

where $\boldsymbol{w}'=(w_1,w_2,\cdots\cdots,w_p)$ is the weight vector,

$$\boldsymbol{A}=\boldsymbol{T}\boldsymbol{C}\boldsymbol{T}' \tag{49}$$

$$\boldsymbol{C}=\begin{pmatrix} C_{Jyy} & C_{Jy1}\cdots\cdots C_{Jyp} \\ C_{Jy1} & C_{J11}\cdots\cdots C_{J1p} \\ \vdots & \vdots \\ C_{Jyp} & C_{Jp1}\cdots\cdots C_{Jpp} \end{pmatrix}$$ is $(p+1)\times(p+1)$ positive definite matrix.

$$\boldsymbol{T}=\begin{pmatrix} 1 & -1 & 0\cdots\cdots 0 \\ 1 & 0 & -1\cdots\cdots 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0\cdots -1 \end{pmatrix}$$ is $p\times(p+1)$ matrix of rank p.

$$\boldsymbol{B}=C_{Jyy}\boldsymbol{H}-\boldsymbol{A}$$

$\boldsymbol{H}$ is $p\times p$ matrix whose elements are all equal to unity.

$$K_1=\frac{1}{n'P_J}(1-f')\left(1+\frac{Q_J}{n'P_J}\right)$$

$$K_2=\frac{1}{nP_J}(1-f)\left(1+\frac{Q_J}{nP_J}\right)$$

Since the matrix $\boldsymbol{T}$ is of rank p and the matrix $\boldsymbol{C}$ is positive definite, it follows that the matrix $\boldsymbol{A}$ is positive definite. Now, let $\boldsymbol{u}$ be $p\times 1$ vector whose elements are all equal to unity. Then, under the restriction $\boldsymbol{u}'\boldsymbol{w}=1$, the variance can be written as

$$Var(_d\hat{\bar{Y}}_{JMR})=\bar{Y}_J^2[(K_2-K_1)\boldsymbol{w}'\boldsymbol{A}\boldsymbol{w}+K_1C_{Jyy}] \tag{50}$$

since $\boldsymbol{w}'\boldsymbol{H}\boldsymbol{w}=1$. We shall use the theory of Lagrange's multiplier and minimize the following function to obtain the optimum weights:

$$L = \boldsymbol{w}'\boldsymbol{A}\boldsymbol{w} - 2\lambda(\boldsymbol{w}'\boldsymbol{u} - 1)$$

Equating to zero the derivative of L with respect to $\boldsymbol{w}$, we have

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\boldsymbol{A}\boldsymbol{w} - 2\lambda\boldsymbol{u} = 0$$

$$\boldsymbol{w} = \lambda\boldsymbol{A}^{-1}\boldsymbol{u}$$

By the restriction $\boldsymbol{w}'\boldsymbol{u} = 1$, it follows that

$$\lambda = 1/\boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u}$$

and
$$\boldsymbol{w} = \boldsymbol{A}^{-1}\boldsymbol{u}/\boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u} \tag{51}$$

Insertion of $\boldsymbol{w}$ in (50) yields

$$Var(\hat{\bar{Y}}_{JMR}) = \bar{Y}_J^2[(K_2 - K_1)/\boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u} + K_1 C_{Jyy}] \tag{52}$$

This completes the proof of the following theorem.

Theorem 12. In a double sample, the multivariate ratio estimator of the *J-th* domain mean $\bar{Y}_J$,

$$\hat{\bar{Y}}_{JMR} = \sum_{i=1}^{p} w_i \hat{\bar{Y}}_{JRi}$$

has the minimum variance when the weight vector $\boldsymbol{w}$ satisfies

$$\boldsymbol{w} = \frac{\boldsymbol{A}^{-1}\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u}}$$

where the matrix $\boldsymbol{A}$ is defined in (49). And the minimum variance of the estimator is

$$Var(\hat{\bar{Y}}_{JRM}) = \bar{Y}_J^2[(K_2 - K_1)/\boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u} + K_1 C_{Jyy}]$$

where

$$K_1 = \frac{1}{n'P_J}(1 - f')\left(1 + \frac{Q_J}{n'P_J}\right)$$

$$K_2 = \frac{1}{nP_J}(1 - f)\left(1 + \frac{Q_J}{nP_J}\right)$$

It can be verified that an approximate bias of the univariate ratio estimator $\hat{\bar{Y}}_{JRi}$ is

$$Bias(\hat{\bar{Y}}_{JRi})=\bar{Y}_J(K_2-K_1)(C_{Jii}-C_{Jyi}) \tag{53}$$

Hence, an approximate bias of the multivariate ratio estimator is

$$\begin{aligned} Bias(_d\hat{\bar{Y}}_{JMR}) &= \bar{Y}_J(K_2-K_1)\sum_{i=1}^{p} w_i(C_{Jii}-C_{Jyi}) \\ &= \bar{Y}_J(K_2-K_1)\boldsymbol{w}'\boldsymbol{d} \\ &= \bar{Y}_J(K_2-K_1)\frac{\boldsymbol{w}'\boldsymbol{A}^{-1}\boldsymbol{d}}{\boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u}} \end{aligned} \tag{54}$$

where $\boldsymbol{d}$ is $p\times 1$ vector whose elements are $d_i=C_{Jii}-C_{Jyi}$. We note that the bias of the multivariate ratio estimator becomes negligible as the sample sizes n and n' are sufficiently large. We also note that the multivariate ratio estimator based on p auxiliary x-variates has smaller variance than those based on q auxiliary x-variates if $p>q$ [see Wey, 1970].

(2) Estimation of the Variance

For estimation of the variance of $_d\hat{\bar{Y}}_{JMR}$ we first note that an element of the matrix $\boldsymbol{A}$ can be expressed as

$$\begin{aligned} a_{hi} &= C_{Jyy}-C_{Jyh}-C_{Jyi}+C_{Jhi} \\ &= \frac{\bar{Y}_J^{-2}}{N_J-1}\sum_{k=1}^{N_J}(y_{Jk}-R_{Jh}x_{Jhk})(y_{Jk}-R_{Ji}x_{Jik}) \end{aligned}$$

Then, an estimator of $\bar{Y}_J^2 a_{hi}$ is

$$\hat{a}_{hi}=\frac{1}{n_J-1}\sum_{k=1}^{n_J}(y_{Jk}-\hat{R}_{Jh}x_{Jhk})(y_{Jk}-\hat{R}_{Ji}x_{Jik}) \tag{55}$$

Now, from expression (50), the conditional variance of $_d\hat{\bar{Y}}_{JMR}$, given n_J and n_J' is

$$\begin{aligned} &Var(_d\hat{\bar{Y}}_{JMR}\,|\,n_J,n_J') \\ &=\left[\frac{1}{n_J}\left(1-\frac{n_J}{N_J}\right)-\frac{1}{n_J'}\left(1-\frac{n_J'}{N_J}\right)\right]\bar{Y}_J^2\sum_{h=1}^{p}\sum_{i=1}^{p}a_{hi}w_hw_i \\ &\quad+\frac{1}{n_J'}\left(1-\frac{n_J'}{N_J}\right)S_{Jy}^2 \end{aligned} \tag{56}$$

Since the value of N_J is usually not known, the proportions n_J/N_J and n_J'/N_J are respectively replaced by their expected values $f=n/N$ and $f'=n'/N$ in the above expression to obtain an estimated variance as follows:

$$var(\hat{\bar{Y}}_{JMR}) = \left(\frac{1-f}{n_J} - \frac{1-f'}{n_J'}\right) \sum_{h=1}^{p} \sum_{i=1}^{p} \hat{a}_{hi}\hat{w}_h\hat{w}_i + \frac{1-f'}{n_J'} s_{Jy}^2$$

$$= \left(\frac{1}{n_J} - \frac{1}{n_J'}\right) \Big/ \boldsymbol{u}'\hat{A}^{-1}\boldsymbol{u} + \frac{1-f'}{n_J'} s_{Jy}^2 \qquad (57)$$

where the matrix $\hat{A}$ has the $\hat{a}_{hi}$ given by (55) as its elements and s_{Jy}^2 is an unbiased estimator of S_{Jy}^2.

III. Double Sampling with Unequal Probabilities and with Replacement

In order to estimate the sub-population mean or total for y in the *J-th* domain, it may be considered desirable to select the sample with probabilities proportional to x although information on x of the whole population is not available. This information is then collected from the first sample of size n' drawn by simple random sampling without replacement from which the second sample of size n is selected with probabilities proportional to x and with replacement. We may also consider the case in which information on a character z, an estimate or measure of size is available. In this situation we may select the first sample of size n' with probabilities π_i proportional to z_i and collect information on x. The second sample is a subsample of the first, selected with equal probability and without replacement in which information on y is collected. We shall consider the estmation of the *J-th* domain total in these two situations.

(A) *Selection of the Second Sample with Unequal Probabilities and with Replacement.*

In a larger preliminary sample of size n' drawn from a population of size N by simple random sampling without replacement, we measure only the auxiliary variate x_i which is highly correlated with the variate y_i, the character under study. The subsample of size n is then selected with probabilities proprtional to x_i from which the variate y_i are collected. Let n_J' and n_J be, respectively, the number of units in the first and second sample that happen to fall into the *J–th* domain. Denote (x_{Jk}, y_{Jk}) by the paired variate which belongs to the *J–th* domain.

We define a new pair (x_i^*, y_i^*) as (x_{Jk}, y_{Jk}) if the *i–th* unit of the population or sample belongs to the *k–th* member of the domain *J*, and (0, 0) otherwise. Assume that the total number N_J of units belonging to the *J-th* domain in the population is not known, then the domain total Y_J is estimated by

$$ {}_d\hat{Y}_J = \frac{N}{n'}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^*}{w_i}\right)=\frac{N}{n'}\bar{y}_{Jr} \qquad (1) $$

where $w_i = x_i / \sum_{k=1}^{n'} x_k$

Denote $\boldsymbol{G}$ by the set of $(x_1, \cdots\cdots, x_{n'})$ in the first sample. Then, the conditional expected value of ${}_d\hat{Y}_J$, given $\boldsymbol{G}$ is

$$ E\,({}_d\hat{Y}_J \mid \boldsymbol{G}) = \frac{N}{n'}\left[\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{y_i^*}{w_i}\mid \boldsymbol{G}\right)\right] $$

$$ = \frac{N}{n'}\sum_{i=1}^{n'} y_i^* $$

$$ = N\bar{y}^* $$

Hence, the unconditional expected value of ${}_d\hat{Y}_J$ is

$$E({}_d\hat{Y}_J)=E[E({}_d\hat{Y}_J|\boldsymbol{G})]=Y^*=Y_J$$

Therefore, ${}_d\hat{Y}_J$ is an unbiased estimator of Y_J. Before proceding to find the variance of ${}_d\hat{Y}_J$, we note that

$$E\left(\frac{y_i^*}{w_i}|\boldsymbol{G}\right)=\sum_{k=1}^{n'} y_k^* \qquad for \quad i=1,\cdots\cdots,n$$

$$Var\left(\frac{y_i^*}{w_i}|\boldsymbol{G}\right)=\sum_{k=1}^{n'} w_k\left(\frac{y^*_k}{w_k}-y^*\right)^2 for \quad i=1,\cdots\cdots,n \qquad (2)$$

and

$$Cov\left(\frac{y_i^*}{w_i},\ \frac{y_j^*}{w_j}|\boldsymbol{G}\right)=0 \qquad for\ i\neq j$$

where

$$y^*=n'\bar{y}^*=\sum_{i=1}^{n'} y_i^*$$

Now, the variance of ${}_d\hat{Y}_J$ is obtained as follows:

$$Var({}_d\hat{Y}_J)=E[Var({}_d\hat{Y}_J|\boldsymbol{G})]+Var[E({}_d\hat{Y}_J|\boldsymbol{G})] \qquad (3)$$

Since the first sample is drawn by simple random sampling without replacement, it follows that

$$Var[E({}_d\hat{Y}_J|\boldsymbol{G})]=Var(N\bar{y}^*)$$

$$=N^2(1-f')\frac{S_*^2}{n'} \qquad (4)$$

where

$$S_*^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i^*-\bar{Y}^*)^2$$

$$\doteqdot P_J S_J^2+P_J Q_J \bar{Y}_J^2$$

$$P_J=N_J/N,\ Q_J=1-P_J,\ f'=n'/N.$$

From the results given in (2), we have the conditional variance of ${}_d\hat{Y}_J$, given $\boldsymbol{G}$ as follows:

$$Var({}_d\hat{Y}_J \mid \boldsymbol{G}) = \left(\frac{N}{n'}\right)^2 \frac{1}{n} \sum_{i=1}^{n'} w_i \left(\frac{y_i^*}{w_i} - y^*\right)^2$$

$$= \left(\frac{N}{n'}\right)^2 \frac{1}{n} \sum_{i=1}^{n'} \sum_{i<j} x_i x_j \left(\frac{y_i^*}{x_i} - \frac{y_j^*}{x_j}\right)^2 \qquad (5)$$

Since the probability of a specified pair of units being selected in the sample of size n' is $n'(n'-1)/N(N-1)$, we have

$$E[Var({}_d\hat{Y}_J/\boldsymbol{G})]$$

$$= \left(\frac{N}{n'}\right)^2 \frac{1}{n} \frac{n'(n'-1)}{N(N-1)} \sum_{i=1}^{N} \sum_{i<j} x_i x_j \left(\frac{y_i^*}{x_i} - \frac{y_j^*}{x_j}\right)^2$$

$$= \frac{N}{N-1} \frac{n'-1}{nn'} \sum_{i=1}^{N} W_i \left(\frac{y_i^*}{W_i} - Y^*\right)^2$$

$$= \frac{N}{N-1} \frac{n'-1}{nn'} \left[\sum_{k=1}^{N_J} W_k \left(\frac{y_{Jk}}{W_k} - Y_J\right)^2 + \left(1 - \sum_{k=1}^{N_J} W_k\right) Y_J^2\right] \qquad (6)$$

where

$$W_k = x_k/X.$$

Substituting (4) and (6) inte (3), we have

$$Var({}_d\hat{Y}_J) = \frac{N}{N-1} \frac{n'-1}{nn'} \left[\sum_{k=1}^{N_J} W_k \left(\frac{y_{Jk}}{W_k} - Y_J\right)^2 + \left(1 - \sum_{k=1}^{N_J} W_k\right) Y_J^2\right]$$

$$+ N^2 (1-f') \frac{S_*^2}{n'} \qquad (7)$$

We have thus proved the following theorem.

Theorem 1. If the first sample of size n' is drawn from a population of size N by simple random sampling without replacement and the subsample of size n is selected with probabilities proportional to x_i, then an unbiased estimator of the domain total Y_J is

$${}_d\hat{Y}_J = \frac{N}{n'} \left(\frac{1}{n} \sum_{k=1}^{n_J} \frac{y_{Jk}}{w_k}\right), \quad w_k = x_k / \sum_{i=1}^{n'} x_i$$

with variance, ignoring terms in N^{-1},

$$Var({}_d\hat{Y}_J)=\frac{n'-1}{nn'}\left[\sum_{k=1}^{N_J} W_k\left(\frac{y_{Jk}}{W_k}-Y_J\right)^2+\left(1-\sum_{k=1}^{N_J} W_k\right)Y_J{}^2\right]$$

$$+\frac{N^2}{n'}(1-f')P_J(S_J{}^2+Q_J\bar{Y}_J{}^2) \qquad (8)$$

where $W_k=x_k/X$, $f'=n'/N$, $P_J=N_J/N$, and $Q_J=1-P_J$.

(B) Estimation of the Variance.

We first obtain an estimated variance of the conditional expected value of ${}_d\hat{Y}_J$, given $\boldsymbol{G}$. It is obvious that an unbiased estimator of $S_*{}^2$ in expression (4) is

$$s_*{}^2=\frac{1}{n'-1}\sum_{i=1}^{n'}(y_i{}^*-\bar{y}^*)^2$$

$$=\frac{1}{n'-1}\left[\sum_{i=1}^{n'} y_i{}^{*2}-\frac{1}{n'}\left(\sum_{i=1}^{n'} y_i{}^*\right)^2\right]$$

We note that given the first sample $\boldsymbol{G}$,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i{}^{*2}}{w_i}\,|\boldsymbol{G}\right]=\sum_{i=1}^{n'} y_i{}^{*2}$$

and

$$E\left[\frac{2}{n(n-1)}\sum_{i<j}^{n}\frac{y_i{}^*}{w_i}\frac{y_j{}^*}{w_j}\,|\boldsymbol{G}\right]=\left(\sum_{i=1}^{n'} y_i{}^*\right)^2$$

Thus, an unbiased estimator of $s_*{}^2$, given $\boldsymbol{G}$ is provided by

$$\hat{s}_*{}^2=\frac{1}{n'-1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i{}^{*2}}{w_i}-\frac{2}{n'n(n-1)}\sum_{i<j}^{n}\frac{y_i{}^*}{w_i}\frac{y_j{}^*}{w_j}\right]$$

$$=\frac{1}{n(n'-1)}\left[\sum_{i=1}^{n}\frac{y_i{}^{*2}}{w_i}-\frac{1}{n'(n-1)}\left\{\left(\sum_{i=1}^{n}\frac{y_i{}^*}{w_i}\right)^2-\sum_{i=1}^{n}\left(\frac{y_i{}^*}{w_i}\right)^2\right\}\right]$$

Hence, an unbiased estimator of $Var[E({}_d\hat{Y}_J|\boldsymbol{G})]$ is

$$var\,[E\,({}_d\hat{Y}_J|\boldsymbol{G})]=$$

$$\frac{N^2(1-f')}{nn'(n'-1)}\left[\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k}-\frac{1}{n'(n-1)}\left\{\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{w_k}\right)^2-\sum_{k=1}^{n_J}\left(\frac{y_{Jk}}{w_k}\right)^2\right\}\right]$$

Next, for estimating the expected value of the conditional variance of ${}_d\hat{Y}_J$, given $\boldsymbol{G}$, we can obtain an unbiased estimator of expression (5) by using the result derived in the estimation of the population total with PPES sampling (see Cochran, 1963, *p.*254):

$$var\,({}_d\hat{Y}_J|\boldsymbol{G})=\left(\frac{N}{n'}\right)^2\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i^*}{w_i}-\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^*}{w_i}\right)^2$$

$$=\left(\frac{N}{n'}\right)^2\frac{1}{n(n-1)}\left[\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k}-\frac{1}{n}\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{w_k}\right)^2\right]$$

It follows that an unbiased estimator of the variance of ${}_d\hat{Y}_J$ is

$$var({}_d\hat{Y}_J)=\left(\frac{N}{n'}\right)^2\frac{1}{n(n-1)}\left[\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k}-\frac{1}{n}\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{w_k}\right)^2\right]$$

$$+\frac{N^2(1-f')}{nn'(n'-1)}\left[\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k}-\frac{1}{n'(n-1)}\left\{\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{w_k}\right)^2-\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k^2}\right\}\right]$$

We have thus proved the following theorem.

Theorem 2. Under the conditions given in *Theorem* 1, an unbiased estimator of the variance of ${}_d\hat{Y}_J$ is

$$var({}_d\hat{Y}_J)=\left(\frac{N}{n'}\right)^2\frac{1}{n(n-1)}\left[\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k^2}-\frac{1}{n}\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{w_k}\right)^2\right]$$

$$+\frac{N^2(1-f')}{nn'(n'-1)}\left[\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k}-\frac{1}{n'(n-1)}\left\{\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{w_k}\right)^2-\sum_{k=1}^{n_J}\frac{y_{Jk}^2}{w_k^2}\right\}\right]$$

where

$$f'=n'N,\quad w_k=x_k/\sum_{i=1}^{n'}x_i.$$

(C) *Selection of the First Sample with Unequal Probabilities and with Replacement*

We now consider the case in which the first sample of size n' is selected with probabilities π_i proportional to z_i, an estimate or measure of size and collect information on x_i. The second sample is a subsample of the first, selected with equal probability and without replacement in which information on y_i is collected. Then, the proposed estimator for the *J-th* domain total Y_J is

$$ {}_d\hat{Y}_{JR}=\frac{\hat{Y}^*}{\hat{X}^*}\hat{X}_L^*=\hat{R}^*\hat{X}_L^* \tag{10}$$

where

$$\hat{Y}^* =\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^*}{\pi_i}, \quad \hat{X}^*=\frac{1}{n}\sum_{i=1}^{n}\frac{x_i^*}{\pi_i}$$

$$\hat{X}_L^*=\frac{1}{n'}\sum_{i=1}^{n'}\frac{x_i^*}{\pi_i}, \quad \hat{R}^*=\hat{Y}^*/\hat{X}^*$$

the x_i^* and y_i^* are defined in the preceding section.

Since the estimator ${}_d\hat{Y}_{JR}$ is a ratio estimator, it is, in general, biased. Assume that the sample size is sufficiently large to make the bias negligible. Under this assumption, we shall find an approximate variance of the estimator. Let $\boldsymbol{G}$ denote as before the set of $(x_1,\cdots\cdots,x_{n'})$ in the first sample.

Given the set $\boldsymbol{G}$, we expand $\hat{X}^{*-1}$ around $E(\hat{X}^*|\boldsymbol{G})=\hat{X}_L^*$ by Taylor's series and retain the first term in the expansion. Then, we have

$$E({}_d\hat{Y}_{JR}|\boldsymbol{G}) = E(\hat{Y}^*|\boldsymbol{G}) = \frac{1}{n'}\sum_{i=1}^{n'}\frac{y_i^*}{\pi_i}$$

so $$E({}_d\hat{Y}_{JR}) = \sum_{i=1}^{N} y_i^* = Y_J$$

Hence, to this order of approximation the estimator is unbiased. An approximate variance is obtained as follows:

$$Var[E({}_d\hat{Y}_{JR}|\boldsymbol{G})]$$

$$= \frac{1}{n'}\sum_{k=1}^{N}\pi_i\left(\frac{y_i^*}{\pi_i} - Y^*\right)^2$$

$$= \frac{1}{n'}\left[\sum_{k=1}^{N_J}\pi_k\left(\frac{y_{Jk}}{\pi_k} - Y_J\right)^2 + \left(1 - \sum_{k=1}^{N_J}\pi_k\right)Y_J^2\right] \qquad (11)$$

Here we assume for simplifying summation that the first N_J units in the population of size N belong to the *J-th* domain.

$$Var\,({}_d\hat{Y}_{JR}|\boldsymbol{G}) = \hat{X}_L^{*2} Var(\hat{R}^*|\boldsymbol{G})$$

$$= \hat{X}_L^{*2} Var\left[\frac{1}{\hat{X}^*}(\hat{Y}^* - R^*\hat{X}^*)\,|\,\boldsymbol{G}\right]$$

$$\doteqdot Var\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^* - R^* x_i^*)/\pi_i|\,\boldsymbol{G}\right]$$

$$= \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{1}{n'-1}\sum_{i=1}^{n'}\left[\frac{y_i^* - R^* x_i^*}{\pi_i} - (\hat{Y}_L^* - R^*\hat{X}_L^*)\right]^2 \qquad (12)$$

where $R^* = Y^*/X^* = R_J$.

Taking expectation of the above expression with respect to $\boldsymbol{G}$, we have

$$E[Var({}_d\hat{Y}_{JR}|\boldsymbol{G})]$$

$$= \left(\frac{1}{n} - \frac{1}{n'}\right)\sum_{i=1}^{N}\pi_i\left[\frac{y_i^* - R^* x_i^*}{\pi_i} - (Y^* - R^* X^*)\right]^2$$

$$= \left(\frac{1}{n} - \frac{1}{n'}\right)\left[V\left(\frac{y}{\pi}\right) + R_J^2 V\left(\frac{x}{\pi}\right) - 2R_J C\left(\frac{y}{\pi}, \frac{x}{\pi}\right)\right]$$

where

$$V\left(\frac{y}{\pi}\right)=\sum_{k=1}^{N_J}\pi_k\left(\frac{y_{Jk}}{\pi_k}-Y_J\right)^2+\left(1-\sum_{k=1}^{N_J}\pi_k\right)Y_J^2 \tag{13}$$

$$V\left(\frac{x}{\pi}\right)=\sum_{k=1}^{N_J}\pi_k\left(\frac{x_{Jk}}{\pi_k}-X_J\right)^2+\left(1-\sum_{k=1}^{N_J}\pi_k\right)X_J^2 \tag{14}$$

$$C\left(\frac{y}{\pi},\frac{x}{\pi}\right)=\sum_{k=1}^{N_J}\pi_k\left(\frac{y_{Jk}}{\pi_k}-Y_J\right)\left(\frac{x_{Jk}}{\pi_k}-X_J\right)+\left(1-\sum_{k=1}^{N_J}\pi_k\right)X_JY_J \tag{15}$$

Therefore, an approximate variance of ${}_d\hat{Y}_{JR}$ is

$$Var({}_d\hat{Y}_{JR})=Var[E({}_d\hat{Y}_{JR}|\boldsymbol{G})]+E[Var({}_d\hat{Y}_{JR}|\boldsymbol{G})]$$
$$=\frac{1}{n}V\left(\frac{y}{\pi}\right)+\left(\frac{1}{n}-\frac{1}{n'}\right)R_J\left[R_JV\left(\frac{x}{\pi}\right)-2C\left(\frac{y}{\pi},\frac{x}{\pi}\right)\right]$$

We summarize the foregoing in the following theorem.

Theorem 3. If the first sample of size n' is selected with probabilities π_i proportional to z_i and the second sample is a subsample of the first, drawn with equal probability and without replacement, then the *J-th* domain total Y_J is estimated by

$${}_d\hat{Y}_{JR}=\hat{R}^*\hat{X}^*_L$$

with an approximate variance (to terms of order n^{-1})

$$Var({}_d\hat{Y}_{JR})=\frac{1}{n}V\left(\frac{y}{\pi}\right)+\left(\frac{1}{n}-\frac{1}{n'}\right)R_J\left[R_JV\left(\frac{x}{\pi}\right)-2C\left(\frac{x}{\pi},\frac{y}{\pi}\right)\right] \tag{16}$$

(D) Estimation of the Approximate Variance.

In a single-phase PPES sampling, an unbiased estimator of expression (11) is known as

$$\frac{1}{n'(n'-1)}\sum_{i=1}^{n'}\left[\frac{y_i^*}{\pi_i}-\frac{1}{n'}\sum_{i=1}^{n'}\frac{y_i^*}{\pi_i}\right]^2$$

Now, since the second sample is a subsample of the first, drawn by equal probability and without replacement, it is obvious that an unbiased estimator of the above expression, given $\boldsymbol{G}$ is

$$\frac{1}{n'(n-1)}\sum_{i=1}^{n}\left[\frac{y_i^*}{\pi_i}-\frac{1}{n}\sum_{k=1}^{n}\frac{y_k^*}{\pi_k}\right]^2$$

$$=\frac{1}{n'(n-1)}\left[\sum_{i=1}^{n_J}\left(\frac{y_{Ji}}{\pi_i}-\frac{1}{n}\sum_{k=1}^{n_J}\frac{y_{Jk}}{\pi_k}\right)^2+\frac{(n-n_J)}{n^2}\left(\sum_{k=1}^{n_J}\frac{y_{Jk}}{\pi_k}\right)^2\right]$$

Next, we assume for a moment that R^* is known. Then, an unbiased estimator of expression (12) is given by

$$\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{y_i^*-R^*x_i^*}{\pi_i}-(\hat{Y}^*-R^*\hat{X}^*)\right]^2$$

Substituting $\hat{R}^*=\hat{Y}^*/\hat{X}^*$ for R^* in the above expression, we have an approximate estimator of the expected value of the conditional variance as follows:

$$\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i^*-\hat{R}^*x_i^*}{\pi_i}\right)^2$$

$$=\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{1}{n-1}\sum_{k=1}^{n_J}\left(\frac{y_{Jk}-\hat{R}_J x_{Jk}}{\pi_k}\right)^2$$

where $\hat{R}_J=\hat{R}^*$.

It follows that the variance of ${}_d\hat{Y}_{JR}$ is approximately estimated by

$$var({}_d\hat{Y}_{JR})=\frac{1}{n'(n-1)}\left[\sum_{k=1}^{n_J}\left(\frac{y_{Jk}}{\pi_k}-\hat{Y}_J\right)^2+(n-n_J)\hat{Y}_J^{\,2}\right]$$

$$+\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{1}{n-1}\sum_{k=1}^{n_J}\left(\frac{y_{Jk}-\hat{R}_J x_{Jk}}{\pi_k}\right)^2 \qquad (17)$$

where $\hat{Y}_J=\frac{1}{n}\sum_{k=1}^{n_J} y_{Jk}/\pi_k=\hat{Y}^*$.

Theorem 4. Under the conditions given in *Theorem* 3, the approximate variance of ${}_d\hat{Y}_{JR}$ is estimated by

$$var({}_d\hat{Y}_{JR})=\frac{1}{n'(n-1)}\left[\sum_{k=1}^{n_J}\left(\frac{y_{Jk}}{\pi_k}-\hat{Y}_J\right)^2+(n-n_J)\hat{Y}_J{}^2\right]$$

$$+\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{1}{n-1}\sum_{k=1}^{n_J}\left(\frac{y_{Jk}-\hat{R}_J x_{Jk}}{\pi_k}\right)^2$$

where

$$\hat{Y}_J=\frac{1}{n}\sum_{k=1}^{n_J} y_{Jk}/\pi_k,\quad \hat{R}_J=\hat{Y}_J/\hat{X}_J.$$

IV. Summary

This work is concerned with using the data obtained from double sampling to estimate means and totals of certain subdivisions (domains of study) of the y-variate. The estimates thus obtained are then used for what is known as an "analytic study" in making comparisons of means and totals of these domains.

In double sampling with equal probability and without replacement, three methods of estimation are considered, namely simple estimation, stratified estimation and ratio estimation. The conditions for these estimations to be more precise than the estimation with single simple random sampling are found under a certain cost function. The extension of ratio estimation to the situation in which $p(\geq 2)$ auxiliary x-variates are available is also considered.

Furthermore, two additional methods of estimation are considered under different double sampling schemes, i. e. one is to select the first sample by simple random sampling without replacement and the second

sample being a subsample of the first drawn with probabilities proportional to auxiliary x-variates and with replacement; and the other is to select the first sample with probabilities proportional to an estimate of size from which auxiliary information on x is collected and the second sample being a subsample of the first drawn by simple random sampling without replacement.

V. References

1. Cochran, W. G. (1963): Sampling Techniques. Second edition. John Wiley & Sons, Inc. New York.
2. Olkin, I. (1958): Multivariate ratio estimation for finite populations. Biometrika 45: 154-165.
3. Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953) Sample Survey Methods and Theory. Vol. II, John Wiley and Sons, Inc. New York.
4. Wey, I. T. (1968): Theory of Sample Surveys with Applications. The Directorate - General of Budgets, Accounts and Statistics, The Executive Yuan, Republic of China.
5. Wey, I. T. (1970): Estimation of parameters in domains of study (1). Report to the National Science Council, Republic of China.