# A VARIANCE WHEN DATA ARISE FROM MULTISTAGE SAMPLING AND RATIO-ESTIMATION

Jai W. Choi, National Center for Health Statistics

## 1 INTRODUCTION

A sample may be selected by multistage sampling procedures and sample elements are then weighted to estimate population totals. Such inflation is accomplished by post-stratified ratio-estimation that is intended to reflect not only its size but also composition such as age-sex-race and residential areas. These weighted data are then used to estimate the average, ratio, and other parameters of interest. The main purpose of this paper is to show how the variance estimators can reflect the actual steps of sampling procedures and ratio estimation.

Such variances can be approximated by combining the two known procedures, the linearization of ratio estimates (Woodruff, 1971) and generalized variance estimator for multistage sampling (Kendall and Stuart, 1968). By this combination, one can reflect these two features on variance estimators.

The first procedures linearize the ratio estimation by a Taylor series expansion and retain the variable portion of the linear expansion for the ratio estimations. The variance of the ratio is then approximated by the variance of the variable portion of linearized ratios.

The second procedures include the generalized form of the variance for aggregate data, whichever design might be used for the selection of the sample.

There would be two sets of summation signs after these two procedures, the first arising from the ratio estimations and the second from the sampling designs. The generalized variance can be obtained by exchanging the summation signs, moving those for sampling in front, and summing up those for the ratio estimation. Only summation signs arising from sampling procedures remain. Then, we can apply the variance formula previously developed to this final result.

These approaches can be applied to the variance estimations for the data collected by the National Center for Health Statistics and other government agencies, where they usually utilize complex sampling designs and estimations, and yet may not reflect both of these features on the variance estimations.

Hidiroglou and Rao (1983) and Shah (1981) used these types of approaches for the analyses of Canadian Health Survey Data and standard errors program for survey data, respectively; the former used variance formulas for equal probability sampling with replacement, while the latter did the equal probability without replacement for two-stage sampling. In parctice it is rare for multistage samplings to use such sampling throughout. For instance, one may select samples with replacement with probability proportional to the size of population (pps) for the first stage and equal probability without replacement for the second stage.

Section 2 introduces some notations used in the following sections. Section 3 presents the generalized variance estimators for aggregates from multistage sampling. In Section 4, the post-stratified ratio estimates are linearized and only the variable portions are retained. We then approximate the variance of the ratios by that of variable portion of linearized ratios. Finally an example and some comments are included in Section 5.

## 2 NOTATIONS

Suppose that the population was stratified into L independent strata, indexed by s = 1,.. ., L, and that the members of the s-th stratum was grouped into $N_{1s}$ PSU's, indexed by i = 1,.. .,$N_{1s}$ and the i-th PSU included $N_{2si}$ members, indexed by j = 1, ..., $N_{2si}$. The corresponding symbols for sample are denoted by the lower case n with the same subscripts as shown in Table 1. Since the variances for aggregates from these strata are additive, we show the variance arising from one stratum, dropping the subscript s for the strata in the following developments.

**Table 1**

Symbols for two-stage clustered sample data when three-stage sampling was performed within a stratum.

| | Population | Sample |
|---|---|---|
| 1st-stage units | $N_1$ | $n_1$ |
| 2nd-stage units | $N_{2i}$ | $n_{2i}$ |
| 3rd-stage units | $N_{3ij}$ | $n_{3ij}$ |
| 1st-stage index | $i = 1...N_1$ | $i = 1...n_1$ |
| 2nd-stage index | $j = 1...N_{2i}$ | $j = 1...n_{2i}$ |
| 3rd-stage index | $k = 1...N_{3ij}$ | $k = 1...n_{3ij}$ |
| Index for cells | $h = 1...q$ | $h = 1...q$ |
| Totals: | $N = \sum_i^{N_1} \sum_j^{N_{2i}} N_{3ij}$ | $n = \sum_i^{n_1} \sum_i^{n_{2i}} n_{2i}$ |
| Cell counts: | $Y_{h3} = \sum_i^{N_1}\sum_j^{N_{2i}}\sum_j^{N_{3ij}} y_{hijk}$ | $y_{h3} = \sum_i^{n_1}\sum_j^{n_{2i}}\sum_j^{n_{3ij}} y_{hijk}$ |
| | $X_{h3} = \sum_i^{N_1}\sum_j^{N_{2i}}\sum_j^{N_{3ij}} x_{hijk}$ | $x_{h3} = \sum_i^{n_1}\sum_j^{n_{2i}}\sum_j^{n_{3ij}} x_{hijk}$ |
| cell prop.: | $Y_{h3}/N$ | $y_{h3}/n$ |
| ratio: | $R_{h3} = X_{h3}/Y_{h3}$ | $r_{h3} = x_{h3}/y_{h3}$ |

$x_{hijk}$ and $y_{hijk}$ are variables for x and y characteristics, respectively.

Table 2 shows the variances of aggregates by the types of aggregate and sampling design. The some of the formulas for these variances are discussed in Section 3. The variances of the ratios in Table 1 can be linearized and thus fall in the same categories as aggregates.

## 3 VARIANCE

Sampling could be done with equal or unequal probability, or probability proportional to the size (pps), with or without replacement, and with symmetrical or asymmetrical designs. Within each stage, we may consider any combination of these options.

We present a generalized variance formula

**Table 2**

Variances by types of design and aggregate

| Design Types | | Aggregate from | | |
|---|---|---|---|---|
| | | 1 stage[1] sampling | 2 stage[2] sampling | 3 stage[6] sampling |
| Unequal prob. | WR [3] | $\underset{1}{\mathrm{var}}(y_{h1})$ | $\mathrm{var}(y_{h2})$ | $\mathrm{var}(y_{h3})$ |
| | WO [4] | $\underset{1}{\mathrm{var}}(y_{h1})$ | $\mathrm{var}(y_{h2})$ | $\mathrm{var}(y_{h3})$ |
| Equal prob. | WR [3] | $\underset{1}{\mathrm{var}}(y_{h1})$ | $\mathrm{var}(y_{h2})$ | $\mathrm{var}(y_{h3})$ |
| | WO [4] | $\underset{1}{\mathrm{var}}(y_{h1})$ | $\mathrm{var}(y_{h2})$ | $\mathrm{var}(y_{h3})$ |
| Combinations[5] | | | $\mathrm{var}(y_{h2})$ | $\mathrm{var}(y_{h3})$ |

[1] $y_{h1}=\sum_{i}^{n_1} y_{hi}$; [2] $y_{h2}=\sum_{i}^{n_1}\sum_{j}^{n_{2i}} y_{hij}$;

[6] $y_{h3}=\sum_{i}^{n_1}\sum_{j}^{n_{2i}}\sum_{j}^{n_{3ij}} y_{hij}$;

[3] with replacement;

[4] without replacement;

[5] combination of equal and unequal probability samplings.

for any estimate $\hat{\theta}_h$ in the h-th cell based on completely arbitrary probabilities of selection. The total variance is then the sum of the variances for all strata.

The symbol E is used for the operator of expectation, var for the variance, and $\hat{\mathrm{var}}$ for the unbiased estimate of var. We may write

$$\mathrm{var}(\theta_h) = \underset{1}{\mathrm{var}}(\underset{>1}{E}(\hat{\theta}_h)) + \underset{1}{E}(\underset{>1}{\mathrm{var}}(\hat{\theta}_h)) \qquad (3.1)$$

where ">1" is the symbol to represent all stages of sampling after the first.

if $\hat{\theta}_h = y_{h1}$ defined in Table 1,

its unbiased estimate can be written as

$$\hat{\mathrm{var}}(\hat{\theta}_h) = \underset{1}{\hat{\mathrm{var}}}(\hat{\theta}_h) + \sum_{i}^{n_1} \pi_i^{(1)} \underset{>1}{\hat{\mathrm{var}}}(y_{hi}) \qquad (3.2)$$

where $\pi_i^{(1)}$ is the probability of the i-th unit included among the $n_1$ PSU's.

The expression (3.1) may be written into three components as

$$\mathrm{var}(\hat{\theta}_h)=\underset{1}{\mathrm{var}}\underset{2}{E}\ \underset{>3}{E}(\hat{\theta}_h)+\underset{1}{E}\ \underset{2}{\mathrm{var}}\underset{>3}{E}(\hat{\theta}_h)+\underset{1}{E}\ \underset{2}{E}\ \underset{>3}{\mathrm{var}}(\hat{\theta}_h).(3.3)$$

If $y_{hi}=\sum_{j=1}^{n_{2i}} y_{hij}$, substituting in (3.2), the unbiased estimate of (3.3) can be written as

$$\hat{\mathrm{var}}(\hat{\theta}_h)= \underset{1}{\hat{\mathrm{var}}}(\hat{\theta}_h) + \sum_{i}^{n_1} \pi_i^{(1)}\underset{2}{\hat{\mathrm{var}}}(y_{hi})$$

$$+ \sum_{i}^{n_1} \pi_i^{(1)} \sum_{j}^{n_{2i}} \pi_{ij}^{(2)} \underset{>2}{\hat{\mathrm{var}}}(y_{hij}), \qquad (3.4)$$

where $\pi_{ij}^{(2)}$ is the probability of selecting the j-th second stage unit in the selected i-th first stage unit. This extension is now obvious for further stages of sampling.

We may summarize the above formulas in words: an unbiased estimator of sampling variance in multistage sampling, when the first stage sampling is without replacement, is obtained as the sum of two components. The first component estimates the variance as if only the first-stage sampling had taken place. The second component is the weighted sum of the estimates, within the selected first stage units, of the variance due to later stages of sampling (the first stage units being regarded as fixed); the weights are the probabilities of selection of these first stage units (Durbin, 1953).

If the sampling is done with replacement at the first stage, only the first term remains in (3.4), regarded as the limit of $\pi_i^{(1)} \to 0$. In this case, it is simple to estimate variances in multistage sampling with any number of stages when the first stage, with replacement, uses the same unequal probabilities at each drawing, while other stages are arbitrary, but carried out independently in different selected first-stage units.

Consider variances for various sampling situations.

**I)** $\underset{1}{\mathrm{var}}(y_{h1})$ for unequal probability without replacement in a single stage:

Let ${}_rP_i$ be the probability that the i-th individual is selected at the r-th drawing, and $\sum_{i}^{N_1} {}_rP_i = 1$, $\pi_i = \sum_{r}^{n_1} {}_rP_i$, $\pi_{ii'} = \sum_{r\neq s}^{n_1}\sum {}_rP_i\ {}_sP_j$.

Kendall and Stuart (1968 vol 3, p172) shows

$$\underset{1}{\mathrm{var}}(y_{h1})= \sum_{i=1}^{N_1} \pi_i\ (1 - \pi_i)\ y_{hi}^2$$

$$+ \sum_{i\neq i'}^{N_1}\sum (\pi_{ii'} - \pi_i\ \pi_{i'})\ y_{hi}\ y_{hi'}\ . \qquad (3.5)$$

From $E(\sum_{i}^{n} g(y_i)) = \sum_{i}^{N} \pi_i\ g(y_i)$ and

$$E(\sum_{i\neq i'}^{n} \sum g(y_i y_{i'}))=\sum_{i\neq i'}^{N} \sum \pi_{ii'} g(y_i y_{i'})$$

for any function g of observations, the unbiased estimate of (3.5) is given by

$$\underset{1}{\hat{\mathrm{var}}}(y_{h1}) = \frac{1}{2} \sum_{i\neq i'}^{n_1}\sum \frac{(\pi_i\pi_{i'} - \pi_{ii'})}{\pi_{ii'}} (y_{hi} - y_{hi'})^2, \qquad (3.6)$$

For one-stage sampling, $\hat{\mathrm{var}}(y_{h1})= \underset{1}{\hat{\mathrm{var}}}(y_{h1})$ in the general formula (3.2).

**II)** $\mathrm{var}(y_{h2})$ for unequal probability without replacement in two-stage sampling:

$$\hat{\mathrm{var}}(y_{h2})= \underset{1}{\hat{\mathrm{var}}}(y_{h2})+ \sum_{i}^{n_1} \pi_i^{(1)}\underset{2}{\hat{\mathrm{var}}}(y_{hi+}),$$

where the first term is given by (3.6) and the second term is the weighted sum of the variances for the second-stages in the selected 1st-stage units.

**III)** $\underset{1}{\mathrm{var}}(y_{h1})$ for equal probability without replacement in a single stage sampling: we have

$\pi_i = \frac{n_1}{N_1} = F_1$ and $\pi_{ii'} = \frac{n_1}{N_1}\frac{(n_1 - 1)}{(N_1 - 1)}$. Using these, we can write (3.6) as

$$\underset{1}{\hat{\text{var}}}(y_{h1}) = (1 - F_1)\frac{n_1}{n_1 - 1}\sum_i^{n_1}(y_{hi} - \bar{y}_h)^2 \qquad (3.7)$$

where $\bar{y}$ is the mean of $y_i'$s.

**IV)** $\underset{1}{\text{var}}(y_{h1})$ for unequal probability with replacement: We now have to allow $\pi_{ii'}$ for $i = i'$, but the term $\pi_i$ and $\pi_{i'}$ in the double summation must still have different suffixes. (3.6) still holds for this sampling.

**V)** $\underset{1}{\text{var}}(y_{h1})$ for equal probability with replacement: In this case, theory simplifies, i.e., $\pi_i = np_i$ and $\pi_{ii'} = n(n - 1)\, P_iP_{i'}$ where $p_i$ is for the probability that the i-th element included in the replacement sampling at any draw, and (3.6) under these definitions can be written as

$$\text{var}(y_{h1}) = \frac{n_1}{2}\sum_i^{n_1}\sum_j P_i\, P_j(y_i - y_j)^2, \qquad (3.8)$$

Since the same conditions as (3.6) hold, now allowing i = i', the unbiased estimate of (3.8) can be expressed as

$$\underset{1}{\hat{\text{var}}}(y_{h1}) = \frac{n_1}{(n_1 - 1)}\sum_i^{n_1}(y_{hi} - \bar{y}_h)^2, \qquad (3.9)$$

which differs from (3.7) only by $(1 - F_1)$, the factor arising from without replacement.

The aim of sample design ( i.e. a choice of the $\pi_{ii'}$ and hence the $\pi_i$) is partially to reduce variance of an estimator as much as possible. We can find some compromise set of $\pi_{ii'}$ which will be effective in producing small variances for all the estimates we may use. Brewer(1963) gives $\pi_i$, and $\pi_{ii'}$ values, which has desirable properties of small variance in (VIII) and $(\pi_i\pi_{i'} - \pi_{ii'}) > 0$ shown in (3.6) when we take two sample units (n=2).

**VI)** $\text{var}(y_{h3})$ for equal probability without replacement where

$$y_{h3} = \sum_i^{n_1}\sum_j^{n_{2i}}\sum_k^{n_{3ij}} y_{hijk}. \qquad (3.10)$$

Suppose that sampling is done with equal probability without replacement at each of three stages sampling in stratum. Substituting such probabilities as in III) in the general formula (3.4), it can be shown that

$$\text{var}(y_{h3}) = n_1\sigma^2_{y_h}(1 - F_1) + F_1\sum_i^{N_1} n_{2i}\sigma^2_{y_{hi}}(1 - F_{2i})$$
$$+ F_1\sum_i^{N_1} F_{2i}\sum_j^{N_{2i}} n_{3ij}\,\sigma^2_{hij}(1 - F_{3ij}), \qquad (3.11)$$

where $F_1 = n_1/N_1$, $F_{2i} = n_{2i}/N_{2i}$, and $F_{3i} = n_{3ij}/N_{3ij}$ are the sampling fractions

$$\sigma^2_{y_{hi}} = \frac{1}{N_1 - 1}\sum_i^{N_1}(y_{hi} - \frac{1}{N_1}\sum_i^{N_1} y_{hi})^2,$$

$$\sigma^2_{y_{hi}} = \frac{1}{N_{2i} - 1}\sum_j^{N_{2i}}(y^*_{hij} - \frac{1}{N_{2i}}\sum_j^{N_{2i}} y^*_{hij})^2,$$

$$\sigma^2_{hij} = \frac{1}{N_{3ij} - 1}\sum_k^{N_{3ij}}(y_{hijk} - \frac{1}{N_{3ij}}\sum_k^{N_{3ij}} y_{hijk})^2$$

$$y_{hi} = \frac{n_{2i}}{N_{2i}}\sum_j^{N_{2i}} y^*_{hij}, \text{ and } y^*_{hij} = \frac{n_{3ij}}{N_{3ij}}\sum_k^{N_{3ij}} y_{hijk}.$$

For symmetrical data, i.e. $n_1n_{2i}n_{3ij} = n_1n_2n_3 = n$, the unbiased estimate of (3.11) is given by

$$\hat{\text{var}}(y_{h3}) = n^2\Big(\frac{s_1^2}{n_1}(1 - F_1) + \frac{n_1}{N_1}\frac{s_2^2}{n_1n_2}(1 - F_2)$$
$$+ \frac{n_1}{N_1}\frac{n_2}{N_2}\frac{s_3^2}{n_1n_2n_3}(1 - F_3)\Big). \qquad (3.13)$$

Every terms in (3.13) after the first is multiplied by the earlier stage sampling fractions $(n_1/N_1)(n_2/N_2)....$, $\sigma^2$ is replaced by sample $s^2$. Note that, if $(n_1/N_1)$ is negligible, all other terms after the first are also negligible.

**VII)** $\text{var}(y_{h2})$ for equal probability without replacement for two-stage sampling:The two-stage result is obtained by putting $N_1 = n_1 = 1$ in (3.11) after appropriate changes of subscripts.

**VIII)** $\text{var}(y_{h3})$ for probability proportional to the population size (pps) with replacement for the first two stages and equal probability without replacement for the third stage:

It is rare in multistage sampling to use equal probabilities sampling throughout for the variance becomes big. When the units vary considerably in size, the effect of equal probability sampling is to make variances very large. This point does not arise in the symmetrical case when all the units at every stage are of equal size. Thus we are obliged to seek some other sampling scheme to reduce the sampling variance.

We may achieve this by varying probabilities at each stage. If overall probability of selection of single element in a multistage sampling is n/N, it is said to be self weighting as the members of sample are equally weighted. Then, the sample variance can be reduced for some estimates. A simple way of achieving the self-weighting pps sampling design is to select $n_1$ PSU's with probabilities $p_i^{(1)}$ at each drawing, $n_{2i}$ second stage units from each of the $n_1$ selected PSU's with probabilities $p_{ij}^{(2)}$, and $n_{3ij}$ third stage units with probabilities $p_{ijk}^{(3)}$ at each drawing, where $p_i^{(1)} = \frac{N_{3i+}}{N}$, $p_{ij}^{(2)} = \frac{N_{3ij}}{N_{3i+}}$ and $p_{ijk}^{(3)} = \frac{1}{N_{3ij}}$

The subscript "+" means the summation over the respective subscript.

Pps sampling design provides the necessary condition to reduce the sample variance as seen by, if $n_1 \, n_{2i} \, n_{3ij} = n_1 \, n_2 \, n_3 = n$,

$$(n_1 \, p_i^{(1)}) \, (n_{2i} \, p_{ij}^{(2)}) \, (n_{3ij} \, p_{ijk}^{(3)}) = \frac{n}{N}$$

which is the overall selection probability for each elementary unit.

For the total $y_{h3}$ of (3.10) from pps sampling with replacement for the first two stages and the equal probability sampling without replacement for the third stage, it can be shown that

$$\begin{aligned} var(y_{h3}) = & \frac{n_1}{N} \sum_{i}^{N_1} N_{3i+}(T_i - \bar{T})^2 \\ & + \frac{n_1}{N} \sum_{i}^{N_1} n_{2i} \sum_{j}^{N_{2i}} N_{3ij}(T_{ij} - \bar{T}_i)^2 \\ & + \frac{n_1}{N} \sum_{i=1}^{N_1} n_{2i} \sum_{j=1}^{N_{2i}} n_{3ij} \, N_{3ij} \, \sigma_{hij}^2 (1 - F_{3ij}) \end{aligned} \quad (3.14)$$

where $T_{ij} = n_{3ij}\mu_{ij}$, $\mu_{ij} = E(m_{ij})$; $T_i = \sum_{j}^{N_{2i}} T_{ij}$;

$m_{ij} = \frac{1}{n_{3ij}} \sum_{k}^{n_{3ij}} y_{hijk}$; $\bar{T} = n_1 \sum_{i}^{N_1} ( \frac{N_{3i+}}{N}) \, \bar{T}_i$ where

$\bar{T}_i = \sum_{j}^{N_{2i}} (\frac{N_{3ij}}{N_{3i+}}) \, T_{ij}$, $N_{3i+} = \sum_{j=1}^{N_{2i}} N_{3ij}$;

and $\sigma_{hij}^2 = \frac{1}{N_{3ij} - 1} \sum_{k=1}^{N_{3ij}} (y_{hijk} - \mu_{ij})^2$.

One may find an unbiased estimate of (3.14) and that the result is consistent with the previous discussions on with-replacement sampling at the first-stage.

**IX)** var($y_{h2}$) for pps with replacement for the first-stage and equal probability without replacement for the second-stage: The two-stage result is obtained from (3.14) by setting $n_1 = N_1 = 1$ and making appropriate changes in notation.

Similarly we can find the variance of ratio and covariance of ratios.

## 4 LINEARIZATION

Ratio estimates can be linearized by a Taylor series expansion under summation. Then the variance of the variable portion of this expansion is the same as the variance of the original ratio. Denote any ratio of variables $u_1, \ldots, u_k$, by a function $f(u_1, u_2, \ldots, u_k)$. It has been known that

$$var( f(u_1 \ldots u_k)) \approx var(\sum_{i}^{k} u_i \frac{\partial f(U_1 \ldots U_k)}{\partial U_i}) \quad (4.1)$$

where $E(u_i) = U_i$ $i = 1, \ldots, k$, and the symbol "≈" means that both sides of the symbol approximately equal.

EXAMPLE 1 Let x and y be the random variables with expected values X and Y, respectively. Consider the ratio x/y. The variance of the ratio is approximated by that of the variable portion of the linear expansion of the ratio:

$$var( \frac{x}{y} ) \approx var( \frac{x - R \, y}{Y}), \quad (4.2)$$

where R =X/Y.

EXAMPLE 2 Ratio $X' = (x/y)\hat{Y}$ is often used for estimation purposes where x and y are variables while $\hat{Y}$ is the known number.

$$var( \frac{x}{y} \hat{Y}) \approx var( \frac{\hat{Y}}{Y} (x - R \, y)). \quad (4.3)$$

Using these two procedures given in Section 3 and Section 4, we can find variance for a complex ratio estimate. We will demonstrate it using an actual example in Section 5.

## 5 AN EXAMPLE AND SUMMARY

Current Population Survey(CPS) estimated X of population characteristic x as follows: (Technical report 40, Bureau of the Census, p 155).

$$X' = \sum_{a}^{A} \frac{x_{aSR} + \sum_{c=1}^{C} \frac{x_{acNS}}{z_c} Z_c}{y_{aSR} + \sum_{c=1}^{C} \frac{y_{acNS}}{z_c} Z_c} \hat{Y}_a, \quad (5.1)$$

where the subscript c = 1, ..., C (C = 48 cells of color-residence) for the collapsed nonself-representing strata (NS) arising from the first ratio adjustment, and a = i, ...,A (A = 60 cells of age-sex-race categories) from the second ratio adjustment. We express (5.1) as

$$X' = \sum_{a}^{A} \frac{x_a'}{y_a'} \hat{Y}_a, \quad (5.2)$$

where $x_a'$ and $y_a'$ are so defined in (5.1)

The terms $x_{aSR}$, $y_{aSR}$, $x_{acNS}$, $y_{acNS}$, $z_c$, $Z_c$, and $\hat{Y}_a$ in (5.1) are given below.

$x_{aSR}$ = the weighted sample total, from ultimate sampling units (USU) in self-representing (SR) primary sampling units (PSUs), of population with the desired characteristic x in the a-th age-sex-race category. The weights are the inverse of the probability of selecting the USUs. In practice, the weights, represented by $w_{sij}$ , also included special weighting and the noninterview adjustment factors.

$y_{aSR}$ = same as for $x_{aSR}$, but for the total population.

$x_{acNS}$= same as $x_{aSR}$, but for the c-th race-residence category for the nonself-representing (NS) population.

$y_{acNS}$= same as for $x_{ac}$, but for the total population.

$z_c$ = estimated total 1980 Census NSR population in the c-th collapsed race-residence category, based on 1980 census population of the NSR sample PSUs, weighted and summed over all NS strata.

$z_{csij}$= 1 if the (sij)-th person in NSR-sample PSU belongs to the c-th color-residence category, 0 otherwise.

$Z_c$ = the 1980 census population in NS strata in the c-th, collapsed race-residence category;

$Y_a$ = independent population total for the U.S. in the a-th age-sex-race category for the current CPS month.

For the variance of X', first we linearize X' under the summation sign summing over a and then consider only the variable portion of linear expansion as shown. Secondly we exchange the summation signs moving those for sampling in front and sum up those for ratio estimations. Thirdly variance of the resulting aggregate can be found by the formula developed previously in Section 3.

These steps are illustrated as follows:

**Step 1** Linearize (5.2) and take the variable portion. The variance of the ratio estimate is approximated by that of linearized portion.

$$\mathrm{var}(X') \approx \mathrm{var}\Big(\sum_a^A \gamma_a \Big( x'_a - \frac{\tilde{X}_a}{\tilde{Y}_a}\, y'_a \Big)\Big), \quad (5.3)$$

Form (5.1), (5.3) can be rewritten as

$$\mathrm{var}(X') \approx \mathrm{var}\Big( \sum_a^A \gamma_a\Big(\Big(x_{aSR} + \sum_c^C \frac{x_{acNS}}{z_c} Z_c\Big) - \frac{\tilde{X}_a}{\tilde{Y}_a}\Big(y_{aSR} + \sum_c^C \frac{y_{acNS}}{z_c} Z_c\Big)\Big)\Big).$$

where $\gamma_a = \frac{\hat{Y}_a}{\tilde{Y}_a}$, $\tilde{X}_a = E(x'_a)$, and $\tilde{Y}_a = E(y'_a)$.

Collecting the terms of SR-strata and NS strata separately, and linearizing the ratio estimate under the summation over c for the second time, we then obtain

$$\mathrm{var}(X') \approx \mathrm{var}\Big( \sum_a^A \gamma_a\Big(x_{aSR} - \frac{\tilde{X}_a}{\tilde{Y}_a} y_{aSR}\Big) + \sum_a^A \gamma_a \sum_c^C \frac{Z_c}{\tilde{Z}_c} \Big(x_{acNS} - \frac{\tilde{X}_a}{\tilde{Y}_a} y_{acNS} + \frac{\tilde{X}_a}{\tilde{Y}_a}\frac{\tilde{Y}_{ac}}{\tilde{Z}_c} z_c - \frac{\tilde{X}_{ac}}{\tilde{Z}_c} z_c\Big)\Big) \quad (5.4)$$

where $\tilde{Z}_c = E(z_c)$, $\tilde{Y}_{ac} = E(y_{acNS})$, $\tilde{X}_{ac} = E(x_{acNS})$,

$x_{aSR} = \sum_s^{L_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} x_{asij}$, $y_{aSR} = \sum_s^{L_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij} y_{asij}$,

$x_{acNS} = \sum_s^{n'_1}\sum_i^{n_{2s}}\sum_j^{n_{3si}} w_{sij}\, x_{acsij}$,

$y_{acNS} = \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}} w_{sij} y_{acsij}$, $z_c = \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{h3i}} w_{sij} z_{csij}$.

$L_1$ is the number of SR-PSU's, $n_{2s}$ the number of number of second-stage sampling units (SSUs) in the s-th SR-PSU, and $n_{3si}$, the number of third-stage sampling units (TSUs) in the i-th SSU. $n'_1$ is the number of the sampled NS-PSUs.

**Step 2** Using above definitions on $x_{aSR}$, $y_{aSR}$, $x_{acNS}$, $y_{acNS}$, and $z_c$ with the triple summations, we move the summations arising from sampling in front of those summations of ratio estimation for age-sex-colors and race-residences. Summing up over a and c for age-sex-color and race-residences, only the summations from sampling remain. (5.4) can be written as

$$\mathrm{var}(X') \approx \mathrm{var}\Big(\sum_s^{L_1} . \sum_i^{n_{s2}} \sum_j^{n_{s3i}} ( B_{+sij} + C_{+sij}) + \sum_s^{n'_1} \sum_i^{n_{s2}} \sum_j^{n_{s3i}} (B'_{++sij} - C'_{++sij} + D'_{++sij} - D_{++sij})\Big), \quad (5.5)$$

which can be expressed as

$$\mathrm{var}(X') = \mathrm{var}\Big(\sum_s^{L_1} \sum_i^{n_{s2}} \sum_j^{n_{s3i}} t_{sij} + \sum_s^{n'_1} \sum_i^{n_{s2}} \sum_j^{n_{s3i}} t'_{sij}\Big) \quad (5.6)$$

with $t_{sij}$ and $t'_{sij}$ are so defined. The six terms in (5.5) are:

(1) $\sum_a^A \gamma_a x_{aSR} = \sum_s^{L_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}}\sum_a^A B_{asij} = \sum_s^{L_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}} B_{+sij}$,

$B_{asij} = \gamma_a w_{sij} x_{asSRij}$;

(2) $\sum_a^A \frac{\tilde{X}_a}{\tilde{Y}_a} \gamma_a y_{aSR} = \sum_s^{L_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}}\sum_a^A C_{asij} = \sum_s^{L_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}} C_{+sij}$,

$C_{asij} = \frac{\tilde{X}_a}{\tilde{Y}_a} \gamma_a w_{sij} y_{asSRij}$;

For the second term,

(3) $\sum_a^A \sum_c^C \gamma_a x_{ac} = \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}}\sum_a^A\sum_c^C B'_{acsij}$

$= \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}} B'_{++sij}$, $B'_{acsij} = \gamma_a w_{sij} x_{acNSsij}$ ;

(4) $\sum_a^A \sum_c^C \gamma_a \frac{\tilde{X}_a}{\tilde{Y}_{ac}} y_{ac} = \sum_s^{n_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}}\sum_a^A\sum_c^C C'_{acsij}$

$= \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}} C'_{++sij}$, $C'_{acsij} = \gamma_a \frac{\tilde{X}_a}{\tilde{Y}_{ac}} w_{sij} y_{acNSsij}$;

(5) $\sum_a^A \sum_c^C \gamma_a \frac{\tilde{X}_a}{\tilde{Y}_a} \frac{\tilde{Y}_{ac}}{\tilde{Z}_c} z_c = \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}}\sum_a^A\sum_c^C D'_{acsij}$

$= \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}} D'_{++sij}$, $D'_{csij} = \gamma_a \frac{\tilde{X}_a}{\tilde{Y}_a} \frac{\tilde{Y}_{ac}}{\tilde{Z}_c} w_{sij} z_{csij}$;

(6) $\sum_a^A \sum_c^C \gamma_a \frac{\tilde{Y}_{ac}}{\tilde{Z}_c} z_c = \sum_s^{n'_1}\sum_i^{n_{s2}}\sum_j^{n_{s3i}}\sum_a^A\sum_c^C D_{acsij}$

$$= \sum_{s}^{n_1'} \sum_{i}^{n_{s2}} \sum_{j}^{n_{s3i}} D_{++sij}, \quad D_{acsij} = \gamma_a \frac{\tilde{Y}_{ac}}{\tilde{Z}_c} w_{sij} z_{csij};$$

**Step 3** Variance of X'. The first term is due to the SR-PSUs and the second term for NS-PSUs. we can calculate their variances separately and sum the two.

Since for the nonself-representing strata, the samples were selected by pps design for the first two stages with replacement, and the last stage with equal probability without replacement, we may use (3.16) for the variance estimation of nonself-representing strata, while the data from self-representing strata arise from only two stages as shown in (3.17). The variance of X' is

$$\text{var}(\sum_{s}^{L_1} \sum_{i}^{n_{s2}} \sum_{j}^{n_{s3i}} t_{sij}) + \text{var}(\sum_{s}^{n_1'} \sum_{i}^{n_{s2}} \sum_{j}^{n_{s3i}} t'_{sij})$$

$$= \sigma^2_{sSR} + \sigma^2_{sNS}, \text{ say.}$$

The first term is the sum of the $L_1$ self-representing PSUs, while the second is that of the $n_1'$ nonself-representing PSUs.

One may want to have the variance of the ratio X'/Y', where X' is already linearized as shown above and Y' is some other ratio estimate using (5.1) and var(X'/Y') can be obtained similarly.

We may use delta method repeatedly for the var(X'). From this example, the first application can be done for the age-sex-race categories, the second application for the color-residence cells, and finally the third application for the three-stage design. Here, there is no other assumption made except for the usual ones arising from Taylor series approximation.

It appears that the procedures presented in this paper may be one of the better methods of variance estimation in the sense that we can indeed reflect the sampling and estimation procedures on the sample variance. The behavior of this method could be further investigated by way of an empirical method.

This method may be used for discrete variables as well as for continuous ones.

## REFERENCES

Durbin, J.(1953). Some results in sampling theory when the units are selected with unequal probabilities. Journal of the Royal Statistical Society, B, 15, 262.

Hidiroglou, M. A. and Rao, J.N.K. (1983). Chi-Square Tests for the Analysis of Three Way Contingency Tables from the Canada Health Survey. Statistics Canada, Ottawa, Canada.

Shah, B.V. (1981). SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Unpublished document, RTI.

Kendall M. G. and Stuart A. S. (1968). The Advanced theory of Statistics, Vol. 3. Hafner Publishing Company, New York.

Woodruff, Ralph S..(1971). Simple Method for Approximating Variance of a Complicated Estimate. Journal of the American Statistical Association, Volume 66, June, pp411-414.