SOME MODEL BASED VARIANCE ESTIMATORS FOR PROPORTIONS FROM CLUSTER SAMPLE SURVEYS WHEN THE DENOMINATOR IS KNOWN

Jai W. Choi
National Center for Health Statistics

1 INTRODUCTION

This paper presents the variance estimators of proportions from weighted or unweighted data arising from cluster sample plans, when the denominator of proportion is known. The variance estimators discussed in this paper are summarized in Table 1. The three methods are applied to the data from four estimation plans.

Often multistage cluster sampling methods are used in a large survey. For instance, National Health Interview Survey and other types of national surveys conducted by the National Center for Health Statistics (NCHS) fall in this category.

The data collected in these surveys are usually inflated to show the estimate of the U.S. population. A weight or inflation factor is attached to each of sample persons in NCHS data tape. This weight is known as the inverse of his or her inclusion probability in the sample, representing a post-stratified group to which this person belongs. The prime motive of this paper is to investigate the variance of such NCHS data.

Section 2 discusses the variance for data from one- or two-stage cluster sample plan. Section 3 introduces the variance for the data, not only clustered, but also weighted. Finally, Section 4 comments on these results and suggests further studies on this topic.

Table 1
12 Types of Variances presented in this paper

	Original sample		Weighted sample	
Method	One-stage clustered	Two-stage clustered	One-stage clustered	Two-stage clustered
Model 1	$var_1(p_h^*)$	$var_1(p_h)$	$var_1(\hat{\pi}_h)$	$var_1(\hat{\pi}_h)$
Model 2	$var_2(p_h)$	$var_2(p_h)$	$var_2(\hat{\pi}_h)$	$var_2(\hat{\pi}_h)$
Design	$var_3(p_h)$	$var_3(p_h)$	$var_3(\hat{\pi}_h)$	$var_3(\hat{\pi}_h)$

* Estimates of population proportion.
Note that letter p is used for clustered data and π for data, weighted and clustered.

2. VARIANCE ESTIMATOR FOR CLUSTERED DATA

Assume that a sample of "a" psu's is selected from A psu's in the fisrt stage, and a sample of b_i elements is selected from B_i elements in the i-th selected psu for the second-stage sampling. It is assumed that the sampling is done randomly with replacement. The notations used for one-stage clustered sample data are shown in Table 2.

For the estimation purpose, define $y_{ijh} = 1$ if the (i,j)-th element falls into the h-th category and $y_{ijh} = 0$ otherwise.

Table 2
Notations for one-stage clustered sample data

	Population	Sample
Cluster:	A	a
Element:	B_i	b_i
cluster:	$i = 1 \ldots A$	$i = 1 \ldots a$
elements:	$j = 1 \ldots B_i$	$j = 1 \ldots b_i$
cells:	$h = 1 \ldots r$	
all counts:	$Y = \sum_i^A B_i$	$n = \sum_i^a b_i$
cell counts:	Y_h^*	$n_h = \sum_i^a \sum_j^{b_i} y_{ijh}$
cell prop.:	$\pi_h = Y_h / Y$	$p_h = n_h / n$
cell vector:	$\boldsymbol{\pi} = (\pi_1 \ldots \pi_r)$	$\mathbf{p} = (p_1 \ldots p_r)$

* The sums are indicated by dropping the subscript.

The basic parameters of interest is $\boldsymbol{\pi}$ ($\sum \pi_h = 1$ and $\pi_h > 0$) for the r multivariate response categories. An unbiased estimate p for $\boldsymbol{\pi}$ is given in Table 2 and its variance is obtained as follows.

Suppose $P(y_{ijh} = 1) = E(y_{ijh}) = \pi_h$ and

$$E(y_{ijh}, y_{i'j'h'}) = \begin{cases} \pi_h \pi_{h'} & \text{if } i \neq i' \\ \delta_{ihh'} & \text{if } i = i',\ j \neq j', \\ 0 & \text{if } i = i',\ j = j' \text{ and } h \neq h' \\ \pi_h & \text{if } i = i',\ j = j' \text{ and } h = h' \end{cases}$$

where $\delta_{ihh'}$ is the probability that $y_{ijh} = y_{i'j'h'} = 1$ for $i = i'$ $j \neq j'$, and all h and h'. $\delta_{ihh'}$ often depends on the size of cluster i and categories h and h'. Using these conditions, it can be shown that

$$E(p_h^2) = \frac{1}{n^2}\left(n\pi_h + \sum_i^a \delta_{ihh} b_i(b_i - 1) + (n^2 - n - G)\pi_h^2\right),$$

$$E(p_h, p_{h'}) = \frac{1}{n^2}\left(\sum_i^a \delta_{ihh} b_i(b_i - 1) + (n^2 - n - G)\pi_h \pi_{h'}\right),$$

where $G = \sum_i^a b_i(b_i - 1)$.

a1. Model 1 results for one-stage clustered sample data

Define Model 1 as follows. Suppose that there exists a set of parameters $(\theta_{ihh'})$ for $0 < \theta_{ihh'} \leq 1$ and $(\delta_{ihh'})$ with

$$\delta_{ihh'} = \begin{cases} \theta_{ihh'} \pi_h + (1 - \theta_{ihh'}) \pi_h^2 & \text{for } h = h' \\ (1 - \theta_{ihh'}) \pi_h \pi_{h'} & \text{for } h \neq h', \end{cases} \quad (2.1)$$

which is the probability that the one member of a pair falls into cell h and the other in cell h', when this pair came from the i-th cluster. Here $\theta_{ihh'}$ is considered as a type of pairwise intra- or inter-cluster homogeneities for the members in the i-th cluster. In this model, we also assume that $\theta_{ihh'} = \theta_{ihh} = 1$ for perfect correlation.

Using this definition (2.1), we can write the variance and covariance of p shown in Table 2 as

$$\text{var}_1(p_h) = \frac{\pi_h(1-\pi_h)}{n}\left(1 + \frac{\sum_i \theta_{ihh'}\, b_i(b_i-1)}{n}\right), \qquad (2.2)$$

$$\text{cov}_1(p_h, p_{h'}) = -\frac{\pi_h\pi_{h'}}{n}\left(1 + \frac{\sum_i^a \theta_{ihh'}\, b_i(b_i-1)}{n}\right).$$

b1. Model 2 results for one-stage clustered sample data

Define Model 2 as

$$d_{ihh'} = \begin{cases} \rho_{ihh}\,\pi_h(1-\pi_h) + \pi_h^2 & \text{for } h = h' \\ \rho_{ihh'}\sqrt{\pi_h(1-\pi_h)\pi_{h'}(1-\pi_{h'})} + \pi_h\pi_{h'} & \text{for } h \neq h' \end{cases} \qquad (2.3)$$

where we also consider the positively correlated data so that $0 \leq \rho_{ihh'} \leq 1$. $\rho_{ihh'}$ is the intra- or inter-cluster homogeneity. (2.3) is the probability when $y_{ijh} = y_{ij'h'} = 1$ occurs.

If two members in the cluster is perfectly correlated, the happening of off-diagonal is impossible, and thus the second row concurs the first row and we can have only the diagonal elements, which add up to one. On the other hand if there is no correlation, the sum of the $d_{ihh'}$ over all h and h' is also one.

Note that the first row of Model 1 is the same as that of Model 2. The latter is more general since there is no need for the assumption of $\theta_{ihh'} = \theta_{ihh} = 1$ for perpectly correlated data. One should note that for binomial data, the two models are equivalent.

Using Model 2, we can show

$$\text{var}_2(p_h) = \frac{\pi_h(1-\pi_h)}{n}\left(1 + \frac{\sum_i^a \rho_{ihh'}\, b_i(b_i-1)}{n}\right) \qquad (2.4)$$

$$\text{cov}_2(p_h, p_{h'}) = -\frac{\pi_h\pi_{h'}}{n}\left(1 + T_{hh'}\,\frac{\sum_i^a \rho_{ihh'} b_i(b_i-1)}{n}\right)$$

where $\rho_{ihh'}$ is the homogeneity parameters in Model 2, and

$$T_{hh'} = \sqrt{\frac{(1-\pi_h)(1-\pi_{h'})}{\pi_h\pi_{h'}}}. \qquad (2.5)$$

We may test the appropriateness of a model. Let the number of total pairs be X and those in cell (h,h') be $X_{hh'}$ h, h' = 1, ..., r, for all clusters of different sizes. One may choose the model of a smaller value of

$$\sum_h^r \sum_{h'}^r \frac{(X_{hh'} - X\,p_{hh'})^2}{X\,p_{hh'}} \qquad (2.6)$$

where $p_{hh'}$ is the probability Model 1 or 2. When r parameters, r-1 of π_h's and θ in the model $p_{hh'}$ are replaced by their consistent estimates, the statistic (2.6) may be tested with chi-squared distribution with $r^2 - r - 1$ degrees of freedom.

c1. A Design-Based estimator from one-stage clustered sample data

We may also derive the variance and covariance via the classical concepts applied to categorical variables. Cochran (1967, pp300-310), Hansen, Hurwitz, and Madow (1964, Vol. 2 p144), and Kendall and Stuart (1968, Vol 3 p192) describe variance of quantitative data arising from multistage sample surveys.

In the following developments, no model is assumed. The units are selected with equal probabilities and elements by simple random sample. For this particular sample plan, the variance estimator for p is given by

$$\text{var}_3(p_h) = \frac{1-f_1}{a\bar{b}^2} S_{1h}^2 + \frac{1}{a^2\bar{b}^2}\sum_i^A \frac{b_i^2(1-f_{2i})}{b_i} S_{2ih}^2 \qquad (2.7)$$

$$\text{cov}_3(p_h, p_{h'}) = \frac{1-f_1}{a\bar{b}^2} S_{1hh'} + \frac{1}{a^2\bar{b}^2}\sum_i^A \frac{b_i^2(1-f_{2i})}{b_i} S_{2ihh'}$$

where $\bar{b} = \frac{1}{a}\sum_i^a b_i$, $f_1 = \frac{a}{A}$, $f_{2i} = \frac{b_i}{B_i}$,

$$S_{1h}^2 = \frac{\sum_i^A (Y_{ih} - \bar{Y}_h)^2}{A-1}, \quad S_{2ih}^2 = \frac{\sum_j^{B_i}(Y_{ijh} - \bar{Y}_{ih})^2}{B_i - 1},$$

$$Y_{ih} = \sum_j^{B_i} y_{ijh}, \quad \bar{Y}_{ih} = \frac{Y_{ih}}{B_i}, \text{ and } \bar{Y}_h = \frac{1}{A}\sum_i^A Y_{ih}.$$

An unbiased estimate of (2.7) is given by replacing the parameters by unbiased sample estimates:

$$s_{1h}^2 = \frac{\sum_i^a (b_i\, n_{ih} - \bar{n}_h)^2}{a-1}, \quad s_{2ih}^2 = \frac{\sum_j^{b_i}(y_{ijh} - \bar{n}_{ih})^2}{b_i - 1},$$

$$n_{ih} = \sum_j^{b_i} y_{ijh}, \quad \bar{n}_{ih} = \frac{n_{ih}}{b_i}, \text{ and } \bar{n}_h = \frac{1}{a}\sum_i^a n_{ih}.$$

$S_{1hh'}$ and $S_{2ihh'}$ can be similarly obtained.

If $f_{2i} = 1$ for all i, that is $b_i = B_i$, (2.7) becomes that appropriate the simple random sampling of the psu's.

If $f_1 = 1$, that is a = A, the formula is that for proportional stratified random sampling, since psu's may then be regarded as strata, all of which are sampled. When f_1 and f_2 are negligible, (2.7) becomes the with-replacement case.

The second term of (2.7) can be written in terms

of proportions (Cochran, 1967 p248). The variance can also be expressed in terms of usual correlation coefficient ρ. It can be shown this resulting variance takes the same form as that of model-based results shown in (2.3) except for constant factor (Cochran, 1967 p242).

a2. Model 1 results for two-stage clustered sample data

We will use the notations shown in Table 3 for two stage cluster sample situations.

Suppose that the sampling of units is done with equal probabilities and that of elements by simple random sample with replacement.

Define $y_{ijkh} = 1$ if the (i,j,k)-th element falls into the h-th response category and = 0 otherwise.

Table 3

Notations for two-stage clustered sample data

	Population	Sample
1st-stage	A	a clusters
2nd-stage	* B_i	b_i clusters
element	M_{ij}	m_{ij} elements
Subscripts		
1st-stage	i = 1... A	i = 1... a
2nd-stage*	j = 1... B_i	j = 1... b_i
elements	k = 1... M_{ij}	k = 1... m_{ij}
cells	h = 1... r	
all counts:	$Y = \sum_i^A \sum_j^{B_i} M_{ij}$	$n = \sum_i^a \sum_j^{b_i} m_{ij}$
cell counts:	Y_h	$n_h = \sum_i^a \sum_j^{b_i} \sum_k^{m_{ij}} y_{ijkh}$
cell prop.:	$\pi_h = Y_h / Y$	$p_h = n_h / n$
cell vector:	$\pi = (\pi_1 \dots \pi_r)$	$p = (p_1 \dots p_r)$

* These are used for elements in one-stage case.

Suppose that $P(y_{ijkh} = 1) = \pi_h$, and

$$E(y_{ijkh}, y_{i'j'k'h'}) = \begin{cases} \pi_h\pi_{h'} & \text{if } i \neq i', \\ \delta_{phh'} & \text{if } i=i',\ j \neq j' \\ \delta_{shh'} & \text{if } i=i',\ j=j',\ k \neq k' \\ \pi_h & \text{if } i=i',\ j=j',\ k=k' \end{cases} \quad (2.8)$$

where δ is the probability when $y_{ijkh} = y_{ij'k'h'} = 1$ occurs for the various combinations of subscripts.

Then, it can be shown that

$$E(p_h^2) = \frac{1}{n^2}(y\pi_h + F\delta_{shh} + (H-F)\delta_{phh} + (n^2 - n - H)\pi_h^2)$$

$$E(p_h p_{h'}) = \frac{1}{n^2}(F\delta_{shh'} + (H-F)\delta_{phh'} + (n^2 - n - H)\pi_h\pi_{h'}), \quad (2.9)$$

where $F = \sum_i^a \sum_j^{b_i} m_{ij}(m_{ij} - 1)$, and $H = \sum_i^a b_i(b_i - 1)$.

We may use Model 1 twice first for the members in the segment and secondly for those in the psu excluding those pairs already counted as a segment pair so that any one pair can not be counted twice.

Specifically, write the pairwise probability of the members in the psu as

$$\delta_{phh'} = \begin{cases} \theta_p\pi_h + (1 - \theta_p)\pi_h^2 & \text{for } h = h' \\ (1 - \theta_p)\pi_h\pi_{h'} & \text{for } h \neq h', \end{cases} \quad (2.10)$$

where $\theta_p = \theta_{pihh'}$. θ_p is an average of homogeneities for the members over all psu's and categories with $0 \leq \theta_p \leq 1$, and the pairwise probability for the members in the segment as

$$\delta_{shh'} = \begin{cases} \theta_s\pi_h + (1 - \theta_s)\pi_h^2 & \text{for } h = h' \\ (1 - \theta_s)\pi_h\pi_{h'} & \text{for } h \neq h', \end{cases} \quad (2.11)$$

where $\theta_s = \theta_{sijhh'}$. θ_s is an average of homogeneities for the members in the segment over all segments and categories with $0 \leq \theta_s \leq 1$.

Using these definitions for $\delta_{shh'}$ and $\delta_{phh'}$, the variance and covariance of p is given by

$$\text{var}_1(p_h) = \frac{\pi_h(1 - \pi_h)}{y}\left(1 + \theta_s\frac{F}{y} + \theta_p\frac{H - F}{y}\right),$$

$$\text{cov}_1(p_h, p_{h'}) = \frac{-\pi_h\pi_h}{y}\left(1 + \theta_s\frac{F}{y} + \theta_p\frac{H - F}{y}\right) \quad (2.12)$$

b2. Model 2 results for two-stage clustered sample data

Model 2 given in (2.3) can also be applied to the first and second stage clusters and we can write the pairwise probability for members in psu as

$$d_{phh'} = \begin{cases} \rho_{phh}\,\pi_h(1 - \pi_h) + \pi_h^2 & \text{for } h = h' \\ \rho_{phh'}\sqrt{\pi_h(1-\pi_h)\pi_{h'}(1-\pi_{h'})} + \pi_h\pi_{h'} & \text{for } h \neq h', \end{cases} \quad (2.13)$$

where $\rho_{phh'}$ is the pairwise intra- or inter-cluster homogeneity for the members in the psu with condition, $0 \leq \rho_{phh'} \leq 1$, and that for the members in the segment as

$$d_{shh'} = \begin{cases} \rho_{shh}\,\pi_h(1 - \pi_h) + \pi_h^2 & \text{for } h = h' \\ \rho_{shh'}\sqrt{\pi_h(1-\pi_h)\pi_{h'}(1-\pi_{h'})} + \pi_h\pi_{h'} & \text{for } h \neq h'. \end{cases} \quad (2.14)$$

where $\rho_{shh'}$ is the pairwise inter- or intra-cluster homogeneity for the members in the segment with the condition of $0 \leq \rho_{shh'} \leq 1$.

Using these probabilities, we can write the variance and covariance of unbiased estimate p as

$$\text{var}_2(p_h) = \frac{\pi_h(1 - \pi_h)}{y}\left(1 + \rho_{shh}\frac{F}{y} + \rho_{phh}\frac{H-F}{y}\right) \quad (2.15)$$

$$\text{cov}_2(p_h p_{h'}) = \frac{-\pi_h\pi_{h'}}{y}\left(1 + T_{hh'}\left(\rho_{shh'}\frac{F}{y} + \rho_{phh'}\frac{H-F}{y}\right)\right)$$

where ρ_{shh} and ρ_{phh} are the pairwise intra-cluster homogeneities in the psu's and segments, respectilvely; similarly, $\rho_{shh'}$ and $\rho_{phh'}$ are for the inter-cluster homogeneities; (2.12) differs from (2.15) only by $T_{hh'}$ shown in (2.5).

c2. A design-based estimator for two-stage clustered sample data

The units are selected with equal probabilities in each stage and elements by simple random sample. Then

$$\text{var}_3(\hat{p}_h) = \frac{a^2(1-f_1)}{n^2 a} S_{1h}^2 + \frac{1}{n^2}\sum_i^A b_i(1-f_{2i}) S_{2ih}^2 + \frac{1}{n^2}\sum_i^a\sum_j^{b_i} m_{ij}(1-f_{3ij}) S_{3ijh}^2, \quad (2.16)$$

where $S_{1h}^2 = \frac{1}{A-1}\sum_{i=1}^A (b_i\bar{Y}_{ih} - \sum_i^A \frac{b_i \bar{Y}_{ih}}{A})^2$,

$$S_{2ih}^2 = \frac{1}{B_i - 1}\sum_{j=1}^{B_i}(m_{ij}\bar{Y}_{ijh} - \sum_j^{B_i}\frac{m_{ij}\bar{Y}_{ih}}{B_i})^2,$$

$$S_{3ijh}^2 = \frac{1}{M_{ij}-1}\sum_{k=1}^{M_{ij}}(Y_{ijkh} - \bar{Y}_{ijh})^2,$$

$$\bar{Y}_{ih} = \frac{Y_{ih}}{B_i}, \text{ and } \bar{Y}_{ijh} = \frac{Y_{ijh}}{M_{ij}}.$$

$\text{cov}_3(p_h, p_{h'})$ can be similarly obtained.

If $B_i = B$, $b_i = b$, $M_{ij} = M$, and $m_{ij} = m$, then $N = ANM$ and $n = abm$ and (2.16) reduces to a simpler form.

An unbiased estimate of (2.16) is obtained from replacing the parameters with unbiased sample estimates:

$$\bar{y}_{ih} = \frac{1}{b_i}\sum_j^{b_i} y_{ijh}, \text{ and } \bar{y}_{ijh} = \frac{1}{m_{ij}}\sum_k^{m_{ij}} y_{ijkh}.$$

The estmate of proportion and its variance for one-stage cluster can be obtained by replacing a by 1, b by a, m by b, j by i, k by j, f_1 by 1, f_{2i} by f_1, f_{3ij} by f_{2i}, and deleting the subscript i.

3. VARIANCE ESTIMATORS FOR CLUSTERED AND WEIGHTED SAMPLE DATA

Here we consider the weighted data instead. The individual weights are known and approximately the inverse of probability to include this person in the sample, often representing the post-stratified group to which he belongs.

Table 4

Notations for one-stage weighted data

	Population	Weighted data
all counts:	$Y = \sum_i^A B_i$	$y = \sum_i^a\sum_j^{b_i} w_{ij}$
cell counts:	Y_h	$y_h = \sum_i^a\sum_j^{b_i} w_{ij} y_{ijh}$
cell prop.:	$\pi_h = Y_h / Y$	$\hat{\pi}_h = y_h / y$
cell vector:	$\pi = (\pi_1 \dots \pi_r)$	$\hat{\pi} = (\hat{\pi}_1 \dots \hat{\pi}_r)$

Let the weight be the inverse of its selection probability ξ_{ij} for the (i,j)-th element. Denote its weight by $w_{ij} = 1/\xi_{ij}$. It can be shown that the variance and covariance of $\hat{\pi}$ are given by

$$\text{var}(\hat{\pi}_h) = \frac{1}{y^2}\left[\pi_h(1-\pi_h)\sum_i^a\sum_j^{b_i} w_{ij}^2 + \sum_i^a\sum_{j\neq j}^{b_i} w_{ij}w_{ij'}(\delta_{ihh'} - \pi_h^2)\right]$$

$$\text{cov}(\hat{\pi}_h, \hat{\pi}_{h'}) = \frac{1}{y^2}\left[-\pi_h\pi_{h'}\sum_i^a\sum_j^{b_i} w_{ij}^2 + \sum_i^a\sum_{j\neq j}^{b_i} w_{ij}w_{ij'}(\delta_{ihh'} - \pi_h\pi_{h'})\right] \quad (3.1)$$

a1'. Model 1 results for oen-stage clustered and weighted sample data

Replacing $\delta_{ihh'}$ with Model 1 shown in (2.1), (3.1) reduces to

$$\text{var}_1(\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{y}\frac{G_{hh}}{y},$$

$$\text{cov}_1(\hat{\pi}_h, \hat{\pi}_{h'}) = -\frac{\pi_h\pi_h}{y}\frac{G_{hh'}}{y}, \quad (3.2)$$

where $G_{hh'} = \sum_i^a \theta_{ihh'}\sum_{j\neq j'}^{b_i} w_{ij}w_{ij'} + \sum_i^a\sum_j^{b_i} w_{ij}^2$.

When the sampling of units is done with equal probabilities and that of the elements by simple random sample, $w_{ij} = AB_i/ab_i$. Under this particular survey plan, we can express $G_{hh'}$ as

$$G_{hh'} = \sum_i^a \frac{A^2B_i^2}{a^2 b_i}(\theta_{ihh'}(b_i - 1) + 1). \quad (3.3)$$

b1'. Model 2 results for one-stage clustered and weighted sample data

Replacing $\delta_{ihh'}$ with Model 2 shown in (2.3), (3.1) reduces to

$$\text{var}_2(p_h) = \frac{\pi_h(1-\pi_h)}{y}\frac{G_{hh}}{y},$$

$$\text{cov}_2(p_h, p_{h'}) = \frac{-\pi_h\pi_h}{y}\frac{G_{hh'}}{y}, \quad (3.4)$$

where $G_{hh} = \sum_i^a \rho_{ihh'}\sum_{j\neq j'}^{b_i} w_{ij}w_{ij'} + \sum_i^a\sum_j^{b_i} w^2{}_{ij}$,

$$G_{hh'} = T_{hh'}\sum_i^a \rho_{ihh'}\sum_{j\neq j'}^{b_i} w_{ij}w_{ij'} + \sum_i^a\sum_j^{b_i} w^2{}_{ij},$$

and $T_{hh'}$ is given in (2.5). G_{hh} and $G_{hh'}$ can be reduced to simpler froms as done in (3.3).

c1'. A design-based estimator for one-stage clustered and estimated data

In the following developments, we use the notations given in Tables 2 and 4 and no model assumptions are made.

When units are selected with equal probabilities and elements by simple random sample, $w_{ij} = AB_i/ab_i$.

Then, an unbiased estimate for population proportion is given by

$$\hat{\pi}_h = \frac{\frac{A}{a}\sum_i^a \frac{B_i}{b_i}\sum_j^{b_i} y_{ijh}}{\sum_i^A B_i} \qquad (3.5)$$

for h = 1, ..., r, with

$$\text{var}_3(\hat{\pi}_h) \simeq \frac{1-f_1}{a\bar{B}^2} S_{1h}^2 + \frac{1}{aA\bar{B}^2}\sum_i^A \frac{B_i^2(1-f_{2i})}{b_i} S_{2ih}^2 \qquad (3.6)$$

where $f_1 = a / A$, $f_{2i} = b_i / B_i$,

$$S_{1h}^2 = \frac{\sum_i^A (Y_{ih} - \bar{Y}_h)^2}{A-1}, \quad S_{2ih}^2 = \frac{\sum_j^{B_i}(Y_{ijh} - \bar{Y}_{ih})^2}{B_i - 1},$$

$$Y_{ih} = \sum_j^{B_i} y_{ijh}, \quad \bar{Y}_{ih} = \frac{Y_{ih}}{B_i}, \text{ and } \bar{Y}_h = \frac{1}{A}\sum_i^A Y_{ih}.$$

The unbiased estimate of (3.6) can be obtained from replacing the parameters with sample estimates:

$$s_{1h}^2 = \frac{\sum_i^a (B_i \bar{y}_{ih} - \hat{\bar{y}}_h)^2}{a-1}, \quad s_{2i}^2 = \frac{\sum_j^{b_i}(y_{ijh} - \bar{y}_{ih})^2}{b_i - 1},$$

$$\bar{y}_{ih} = \frac{y_{ih}}{b_i}, \quad \bar{y}_h = \frac{y_h}{a}, \text{ and } \hat{\bar{y}}_{ih} = \frac{1}{a}\sum_i^a B_i \bar{y}_{ih};$$

and similarly for $\text{cov}_3(\hat{\pi}_h, \hat{\pi}_{h'})$.

a2'. Model 1 result for two-stage clustered and weighted sample data

The one-stage case previously discussed can easily be extended to two-stage situation.

The notations used in this section are shown in Tables 3 and 5.

Let $w_{ijk} = 1/\xi_{ijk}$ be the weight for the (i,j,k)-th element, where ξ_{ijk} is the inclusion probability of this element in the sample.

Table 5
Notations for two-stage weighted data

	Population	Weighted data
all counts:	$Y = \sum_i^A \sum_j^{B_i} M_{ij}$	$y = \sum_i^a \sum_j^{b_i} \sum_k^{m_{ij}} w_{ijk}$
cell counts:	Y_h	$y_h = \sum_i^a \sum_j^{b_i} \sum_k^{m_{ij}} w_{ij} y_{ijkh}$
cell prop.:	$\pi_h = Y_h / Y$	$\hat{\pi}_h = y_h / y$
cell vector:	$\boldsymbol{\pi} = (\pi_1 \ldots \pi_r)$	$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1 \ldots \hat{\pi}_r)$

The variance estimator for π is given by

$$\text{var}(\hat{\pi}_h) = \frac{1}{y^2}\left[\pi_h(1-\pi_h)\sum_i^a \sum_j^{b_i} \sum_\kappa^{m_{ij}} w_{ijk}^2 + (\delta_{ishh'} - \pi_h^2)\sum_i^a \sum_j^{b_i} \sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ijk}w_{ijk'} + (\delta_{iphh'} - \pi_h^2)\sum_i^a \sum_{j\neq j'}^{b_i} \sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ijk}w_{ij'k'}\right],$$

$$\text{cov}(\hat{\pi}_h, \hat{\pi}_{h'}) = \frac{1}{y^2}\left[-\pi_h\pi_{h'}\sum_i^a \sum_j^{b_i} \sum_k^{m_{ij}} w_{ijk}^2 + (\delta_{ishh'} - \pi_h\pi_{h'})\sum_i^a \sum_j^{b_i} \sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ijk}w_{ijk'} + (\delta_{iphh'} - \pi_h\pi_{h'})\sum_i^a \sum_{j\neq j'}^{b_i} \sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ijk}w_{ij'k'}\right] \qquad (3.7)$$

When $\delta_{ishh'}$ and $\delta_{iphh'}$ are replaced by (2.10) and (2.11), the Model 1 result of (3.7) is given by

$$\text{var}_1(\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{y^2}\left[\sum_i^a \sum_j^{b_i} m_{ij}w_{ij}^2 + \theta_s \sum_i^a \sum_j^{b_i} m_{ij}(m_{ij}-1)w_{ij}^2 + \theta_p \sum_i^a \sum_{j\neq j'}^{b_i} \sum_{k\neq k'}^{m_{ij}} w_{ij}\, w_{ij'}\right]$$

$$\text{cov}_1(\hat{\pi}_h\hat{\pi}_{h'}) = \frac{-\pi_h\,\pi_{h'}}{y^2}\left[\sum_i^a \sum_j^{b_i} m_{ij}\, w_{ij}^2 + \theta_s \sum_i^a \sum_j^{b_i} m_{ij}(m_{ij}-1)w_{ij}w_{ij} + \theta_p \sum_i^a \sum_{j\neq j'}^{b_i} \sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ij}\, w_{ij'}\right] \qquad (3.8)$$

where $\theta_s = \theta_{sijhh'}$ and $\theta_p = \theta_{pihh'}$.

If $w_{ijk} = w$, $m_{ij} = m$, and $b_i = b$, (3.8) reduces to

$$\text{var}_1(\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{y} w(1+\theta_s(m-1)+\theta_p\, m(b-1))$$

$$\text{cov}_1(\hat{\pi}_h, \hat{\pi}_{h'}) = \frac{-\pi_h\,\pi_{h'}}{y} w(1+\theta_s(m-1)+\theta_p m(b-1)) \qquad (3.9)$$

where w is the weighting effects and $1 + \theta_s(m-1) + \theta_p m(b-1)$ is the design effect from two-stage cluster sample plan. When units are selected with equal probabilities, and the elements by simple random sample, we may write $w_{ijk} = AB_iM_{jk}/ab_im_{jk}$, which in turn gives a simple form of (3.7).

b2'. Model 2 results for two-stage clustered and weighted sample data

Replacing $\delta_{ishh'}$ and $\delta_{iphh'}$ with (2.13) and (2.14), respectively, (3.7) reduces to

$$\mathrm{var}_2(\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{y^2}\left[\sum_i^a\sum_j^{b_i}\sum_\kappa^{m_{ij}} w_{ijk}^2 + \rho_s \sum_i^a\sum_j^{b_i}\sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ijk}w_{ijk'} + \rho_p \sum_i^a\sum_{j\neq j'}^{b_i}\sum_{\kappa\neq\kappa'}^{m_{ijj'}} w_{ijk}w_{ij'k'}\right],$$

$$\mathrm{cov}_2(\hat{\pi}_h,\hat{\pi}_{h'}) = \frac{-\pi_h\pi_h}{y^2}\left[\sum_i^a\sum_j^{b_i}\sum_\kappa^{m_{ij}} w_{ijk}^2 + \rho_s \sum_i^a\sum_j^{b_i}\sum_{\kappa\neq\kappa'}^{m_{ij}} w_{ijk}w_{ijk'}T_{hh'} + \rho_p \sum_i^a\sum_{j\neq j'}^{b_i}\sum_{\kappa\neq\kappa'}^{m_{ijj'}} w_{ijk}w_{ij'k'}T_{hh'}\right], \quad (3.10)$$

where $T_{hh'}$ is given in (2.5). If w_{ijk} are replaced by AB_iM_{ij}/ab_im_{ij}, (3.10) can be simplified as done in (3.3).

c2'. A design-based estimator for two-stage clustered and estimated data

Kendall and Stuart (1968 Vol 3 p190) show the variance for quantitative data.

If the sampling is done with equal probabilities for each stages and with simple random sample for elements, the weight can be expressed as

$$w_{ijk} = AB_iM_{ij}/ab_im_{ij}.$$

An unbiased estimate of population proportion is given by

$$\hat{\pi}_h = \frac{\frac{A}{a}\sum_i^a \frac{B_i}{b_i}\sum_j^{b_i}\frac{M_{ij}}{m_{ij}}\sum_k^{m_{ij}} y_{ijkh}}{Y} \quad \text{with}$$

$$\mathrm{var}_3(\hat{\pi}_h) \sim \frac{A^2}{Y^2}\frac{1-f_1}{a}S_{1h}^2 + \frac{A}{Y^2a}\sum_i^A \frac{B_i^2(1-f_{2i})}{b_i}S_{2ih}^2 + \frac{A}{Y^2a}\sum_i^a\frac{B_i}{b_i}\sum_j^{b_i}\frac{M_{ij}^2(1-f_{3ij})}{m_{ij}}S_{3ijh}^2, \quad (3.11)$$

where Y is given Table 5,

$$S_{1h}^2 = \frac{1}{A-1}\sum_{i=1}^A (Y_{ih}-\bar{Y}_h)^2, \quad S_{2ih}^2 = \frac{1}{B_i-1}\sum_{j=1}^{B_i}(Y_{ijh}-\bar{Y}_{ih})^2,$$

$$S_{3ijh}^2 = \frac{1}{M_{ij}-1}\sum_{k=1}^{M_{ij}}(Y_{ijkh}-\bar{Y}_{ijh})^2.$$

An unbiased estimate of (3.11) is obtained from replacing the parameters with unbiased estimates:

$$s_{1h}^2 = \frac{1}{a-1}\sum_{i=1}^a (B_i\bar{y}_{ih} - \frac{1}{a}\sum_i^a B_i\bar{y}_{ih})^2,$$

$$s_{2ih}^2 = \frac{1}{b_i-1}\sum_{j=1}^{b_i}(M_{ij}\bar{y}_{ijh} - \frac{1}{b_i}\sum_j^{b_i}M_{ij}\bar{y}_{ijh})^2,$$

$$s_{3ijh}^2 = \frac{1}{m_{ij}-1}\sum_{k=1}^{m_{ij}}(y_{ijkh}-\bar{y}_{ijh})^2,$$

$$\bar{y}_{ih} = \frac{1}{b_i}\sum_j^{b_i} y_{ijh}, \quad \text{and} \quad \bar{y}_{ijh} = \frac{1}{m_{ij}}\sum_k^{m_{ij}} y_{ijkh}.$$

Similarly, $\mathrm{cov}_3(\hat{\pi}_h,\hat{\pi}_{h'})$ can be obtained.

If $B_i = B$, $b_i = b$, $M_{ij} = M$, $m_{ij} = m$, (3.11) reduces to a simpler from.

The estimate of proportion and its variance for one-stage cluster case can be obtained by setting $A = a = 1$ and replacing B_i by A, b_i by a, M_{ij} by b_i, f_1 by 1, f_{2i} by f_1, f_{3ij} by f_{2i}, j by i, k by j, and deleting i.

Other sampling plans can easily be reflected in the variance form. Kendall and Stuart (1968, vol.3 p198) show a general form of such variance for quantitative data.

4 COMMENTS

All model-based variance estimators show a common form of variance and covariance, that is the variance of multinomial variates multiplied by the factor G. G is design effects for unweighted data. For clustered and weighted data, it combines the design effect with that of weight. The four design-based estimators for proportions can be comparable with the corresponding model-based results when the latters are properly adjusted for particular features of design.

Above results from the three approaches can also be compared with those from other methods, such as BRR, Jackknife, Bootstrap, Linear approximation, and Williams method, through the simulation or other algebraic solutions.

One of the two approaches may be used to test appropriateness for the use of a model. One is shown in (2.6) and the other may be accomplished by comparing the result from the model assumption to that arising from design-based (no-model) estimator.

Further studies on these and other types of proportions are needed to fully understand the characteristics of these variances, including numerical examples as well as their asymptotic properties.

REFERENCES

Cochran, William G. (1967). Sampling Techniques. 2nd Ed., John Wiley and Sons, Inc. New York.

Hansen, Morris H.; Hurwitz, William N.;and Madow, William G. (1964). Sample Survey Methods and Theory. John Wiley and Sons, Inc. New York.

Kendall, Maurice G. and Stuart, Alan (1968). The Advanced Theory of Statistics. Hafner Publishing Co., New York.