A MULTIVARIATE METHODOLOGY FOR ANALYZING DATA FROM STRATIFIED MULTI-STAGE SAMPLING SURVEYS

Robert E. Hausman and Aridaman K. Jain, Bell Laboratories

1. INTRODUCTION

Techniques for estimating means and their variances from stratified multistage sampling survey data are well-known [1,2]. Often, we would like to perform multivariate analyses of such data. This would require estimates of variances and covariances of individual observations. Estimation of components of variances in the case of simple random sampling is discussed extensively in the literature [3,4,5]. However, there is a very limited discussion of the estimation of variance components in the case of unequal probability sampling [6,7]. In this paper, we discuss concepts related to multivariate analyses and develop estimators for variances and covariances for multistage sampling with unequal probabilities. Application of the estimators discussed here to stratified multistage sampling surveys is illustrated by showing the correspondence with a large Bell System survey.

The Bell System survey discussed here is the Measured Impairment Survey (MIS) which was aimed at estimating the quality of the Bell System network as a function of central office measurements. A sampling of customer telephone calls was performed in three stages. In the first stage, 27 central offices were chosen from ten geographic strata with probabilities proportional to the number of loops (telephone lines) in the central office. In the second stage, a sample of loops was selected from each office with the probabilities proportional to one plus the number of outgoing toll calls in a recent month. In the third stage, only one outgoing call (toll or local) was chosen from each sampled loop for study.

This paper discusses the estimation of means, variances and covariances from multistage sampling survey data such as in the MIS. These estimates can be used to perform several types of multivariate analysis. Section 2 gives some notation and a linear model which describes individual measurements as a sum of the contributions of different sampling stages. Section 3 gives the partition of the total variability into within and between first-stage units. Section 4 discusses the estimation of population parameters for an individual stratum. Section 5 discusses how the estimators developed in Section 4 can be used in various multivariate analyses. Section 6 gives five estimates of the unknown sizes of the second stage units. Section 7 discusses how the estimators for individual strata are combined to obtain estimators for the whole population. Section 8 gives a summary.

2. NOTATION AND MODEL

In this section we develop the notation used in the subsequent sections. While the definitions below are in general terms, the corresponding specific definitions for the MIS are indicated in parentheses to facilitate understanding of the statistical methodology. Let

q = stratum number, $q = 1,2,\ldots,L$

M_q = total number of first-state units (central offices) in stratum q;

m_q = number of first stage units in the sample from stratum q, $m_q \leq M_q$;

N_{qi} = total number of second-stage units (loops) in first-stage unit i in the stratum q;

$N_{q\cdot}$ = total number of second-stage units in stratum q;

$W_{qi} = N_{qi}/N_{q\cdot}$ = probability of selection of first-stage unit i in stratum q on a single draw;

n_{qi} = number of second-stage units in the sample from first-stage unit i in stratum q, $n_{qi} \leq N_{qi}$;

C_{qij} = total number of third-stage units (telephone calls - outgoing toll and outgoing local) for second-stage unit j in first-stage unit i in stratum q;

A_{qij} = first estimate (all outgoing and incoming telephone calls in an 8-hour period, called peg count) of the total number of third-stage units, subject to a constant of proportionality, for second-stage unit j in first-stage unit i in stratum q;

T_{qij} = second estimate (1 + outgoing toll telephone calls in a month) of the total number of third-stage units, subject to a constant of proprotionality, for second-stage unit j in first-stage unit i in stratum q;

$T_{qi\cdot}$ = total of the second estimate of the number of third-stage units in first-stage unit i in stratum q;

$Z_{qij} = T_{qij}/T_{qi\cdot}$ = probability of selection of second-stage unit j in first-stage unit i in stratum q on a single draw;

$X_{qij\ell k}$ = measurement on variable k ($k = 1,2,\ldots,K$) on third-stage unit ℓ ($\ell = 1,2,\ldots,C_{qij}$) in second-stage unit (q,i,j);

$$X_{qij\cdot k} = \sum_{\ell=1}^{C_{qij}} X_{qij\ell k};\quad \bar{X}_{qij\cdot k} = X_{qij\cdot k}/C_{qij};$$

$$X_{qi\cdot\cdot k} = \sum_{j=1}^{N_{qi}} X_{qij\cdot k};\quad C_{qi\cdot} = \sum_{j=1}^{N_{qi}} C_{qij};$$

$$\bar{X}_{qi\cdot\cdot k} = X_{qi\cdot\cdot k}/C_{qi\cdot};$$

$$X_{q\cdot\cdot\cdot k} = \sum_{i=1}^{M_q} X_{qi\cdot\cdot k};\quad C_{q\cdot\cdot} = \sum_{i=1}^{M_q} C_{qi\cdot};$$

$$\bar{X}_{q\cdot\cdot\cdot k} = X_{q\cdot\cdot\cdot k}/C_{q\cdot\cdot};$$

$$X_{\cdots k} = \sum_{q=1}^{L} X_{q\cdots k}; \quad C_{\cdots} = \sum_{q=1}^{L} C_{q\cdot\cdot};$$

and

$$\bar{X}_{\cdots k} = X_{\cdots k} / C_{\cdots} .$$

A linear model which describes an individual measurement as a sum of the contributions of different sampling stages is as follows:

$$X_{qij\ell k} = \bar{X}_{\cdots k} + (\bar{X}_{q\cdots k} - \bar{X}_{\cdots k}) + (\bar{X}_{qi\cdot\cdot k} - \bar{X}_{q\cdots k}) + (\bar{X}_{qij\cdot k} - \bar{X}_{qi\cdot\cdot k}) + \varepsilon_{qij\ell k} ,$$

where

$\varepsilon_{qij\ell k}$ = variation among third-stage units in measurement on variable k for second-stage unit (q,i,j).

As indicated earlier, only one third-stage unit was chosen in MIS (i.e., $\ell = 1$). In this case we cannot separate out the effects of second and third stage units. A partition of the variance of $X_{qij\ell k}$ into contributions due to the first and second stages is discussed in the next section.

3. PARTITION OF VARIATION IN STRATUM q

In stratum q, the population variance within first-stage units $(V_{qk}(W))$ and the population variance between first-stage units $(V_{qk}(B))$ for for variable k can be defined as follows

$$V_{qk}(W) = \frac{\sum_{i=1}^{M_q} \sum_{j=1}^{N_{qi}} \sum_{\ell=1}^{C_{qij}} \left(X_{qij\ell k} - \bar{X}_{qi\cdot\cdot k}\right)^2}{\sum_{i=1}^{M_q} \sum_{j=1}^{N_{qi}} C_{qij}}$$

and

$$V_{qk}(B) = \frac{\sum_{i=1}^{M_q} C_{qi\cdot} \left(\bar{X}_{qi\cdot\cdot k} - \bar{X}_{q\cdots k}\right)^2}{\sum_{i=1}^{M_q} \sum_{j=1}^{N_{qi}} C_{qij}} .$$

The total population variance for variable k is the sum of the above two components:

$$V_{qk}(T) = V_{qk}(W) + V_{qk}(B) .$$

This breakdown of variance into within and between components can also be extended to covariances. The population covariance within first-stage units $(C_{qkk'}(W), k \neq k')$ and the population covariance between first-stage units $(C_{qkk'}(B), k \neq k')$ for variables k and k′ are as follows:

$$C_{qkk'}(W) = \frac{\sum_{i=1}^{M_q} \sum_{j=1}^{N_{qi}} \sum_{\ell=1}^{C_{qij}} \left(X_{qij\ell k} - \bar{X}_{qi\cdot\cdot k}\right)\left(X_{qij\ell k'} - \bar{X}_{qi\cdot\cdot k'}\right)}{\sum_{i=1}^{M_q} \sum_{j=1}^{N_{qi}} C_{qij}}$$

and

$$C_{qkk'}(B) = \frac{\sum_{i=1}^{M_q} C_{qi\cdot} \left(\bar{X}_{qi\cdot\cdot k} - \bar{X}_{q\cdots k}\right)\left(\bar{X}_{qi\cdot\cdot k'} - \bar{X}_{q\cdots k'}\right)}{\sum_{i=1}^{M_q} \sum_{j=1}^{N_{qi}} C_{qij}} .$$

The total population convariance for variables k and k′ is the sum of the above two components:

$$C_{qkk'}(T) = C_{qkk'}(W) + C_{qkk'}(B) .$$

In a multivariate setting, we can use the preceding formulas to define covariance matrices (and hence correlation matrices) at three levels: overall (T), within (W), and between (B).

Suppose that we wish to perform a multivariate analysis which requires the covariance matrix of the variables involved. We must first determine which of the matrices described above should be used. The proper choice depends on the purpose of the analysis.

To illustrate the partition of the total variation into within and between components, we discuss a small example. Consider a bivariate population with two first-stage units, each of which has two second-stage units, each of which has just one third-stage unit. In this case

$M_q = 2$;

$N_{qi} = 2$ for i = 1, 2;

$C_{qij} = 1$ for i = 1, 2 and j = 1, 2;

and

k = 2.

The population values for the two variables are shown in Table 1 and the means, variances and correlations are given in Table 2.

Suppose we were interested in the correlation between the two variables. By ignoring the office structure, we can use the four observations (telephone calls) to compute an overall correlation $\rho_T \simeq -0.84$. This number, however, does not tell the whole story. For example, let the first variable correspond to some physical characteristic of the central office building and the second variable relate to the quality of service. If we were

interested in designing new office buildings, then we should use the between offices correlation $\rho_B = -1.00$.

On the other hand, suppose the first variable corresponds to some physical attribute of the line itself. If we were interested in determining the optimal value of that variable for future lines, then we should probably use the within-office correlation $\rho_W = +1.00$ in our analysis.

Table. 1. Population Values for the Two Variables

Variable \ Stage	First	1		2	
	Second	1	2	1	2
1		5	6	2	3
2		30	40	70	80

Table 2. Means, Variances and Correlations for the Two Variables

(i) Means

$$\bar{X}_{q1\cdot\cdot 1} = 5.5,\ \bar{X}_{q2\cdot\cdot 1} = 2.5,\ \bar{X}_{q\cdot\cdot\cdot 1} = 4.0$$

$$\bar{X}_{q1\cdot\cdot 2} = 35,\ \bar{X}_{q2\cdot\cdot 2} = 75,\ \bar{X}_{q\cdot\cdot\cdot 2} = 55$$

(ii) Variances and Covariances

$$\underset{\sim}{T}_q = \begin{bmatrix} 2.5 & -27.5 \\ -27.5 & 425.0 \end{bmatrix}$$

$$\underset{\sim}{W}_q = \begin{bmatrix} 0.25 & 2.5 \\ 2.5 & 25.0 \end{bmatrix}$$

$$\underset{\sim}{B}_q = \begin{bmatrix} 2.25 & -30.0 \\ -30 & 400.0 \end{bmatrix}$$

(iii) Correlations

$$\rho_T = -0.84$$

$$\rho_W = +1.00$$

$$\rho_B = -1.00$$

4. ESTIMATORS OF POPULATION PARAMETERS IN STRATUM q

In this section we develop some estimators of population means, variances and covariances.

4.1 Estimators of Means for First-Stage Unit 1 in Stratum q

We consider the case in which only one third-stage unit is selected from each second-stage unit, as in the case of MIS. An estimate of the total of the measurements of variable k for all third-stage units in second-stage unit (q, i, j) is given by

$$\hat{X}_{qij\cdot k} = C_{qij} X_{qijlk}\ , \qquad (1)$$

where a hat is used to denote an estimator of a population quantity. In a more general case in which two or more third-stage units are selected, $\hat{X}_{qij\cdot k}$ would be based on the mean of the sampled third-stage units in second-stage unit (q, i, j).

Next, an estimate of $X_{qi\cdot\cdot k}$, the total of the measurements of variable k for all third-stage units in first-stage unit i in stratum q, based on second stage unit (q, i, j) is

$$C_{qij} X_{qijlk} / Z_{qij}\ ,$$

where Z_{qij} is the probability of selection of second-stage unit (q, i, j). By averaging over n_{qi} second-stage units in the sample, an estimate of $X_{qi\cdot\cdot k}$ is given by

$$\hat{X}_{qi\cdot\cdot k} = \frac{1}{n_{qi}} \sum_{j=1}^{n_{qi}} C_{qij} X_{qijlk} / Z_{qij}\ . \qquad (2)$$

The estimator given by equation (2) can be computed only if the $C_{qij} (j = 1,2,\ldots,N_{qi})$ are known. Here, we will consider the case in which the C_{qij} are unknown even for the sampled second-stage units, as in the case of the MIS. In this case, it is necessary to estimate the C_{qij}. Let $\hat{C}_{qij}$ denote an estimate of C_{qij}. Five choices for estimating C_{qij} are discussed in Section 6. By replacing C_{qij} by its estimate $\hat{C}_{qij}$, we obtain

$$\hat{X}_{qi\cdot\cdot k} = \frac{1}{n_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij} X_{qijlk} / Z_{qij} \qquad (3)$$

and

$$\hat{C}_{qi\cdot} = \frac{1}{n_{qi}} \sum_{j=1}^{n_{qi}} C_{qij} / Z_{qij}\ , \qquad (4)$$

and an estimate of $\bar{X}_{qi\cdot\cdot k}$, the mean of variable k for first-stage unit i in stratum q, is

$$\hat{\bar{X}}_{qi\cdot\cdot k} = \hat{X}_{qi\cdot\cdot k} / \hat{C}_{qi\cdot} \qquad (5)$$

4.2 Estimators of Means for Stratum q

Estimates of $X_{q\cdot\cdot\cdot k}$, $C_{q\cdot\cdot}$, and $\bar{X}_{q\cdot\cdot\cdot k}$ are given by

$$\hat{X}_{q\cdot\cdot\cdot k} = \frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi} W_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij} X_{qijlk} / Z_{qij}, \qquad (6)$$

$$\hat{C}_{q\cdot\cdot} = \frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi} W_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij} / Z_{qij}, \qquad (7)$$

and $\hat{\bar{X}}_{q\cdot\cdot\cdot k} = \hat{X}_{q\cdot\cdot\cdot k} / \hat{C}_{q\cdot\cdot}$.

By substituting the values of $\hat{X}_{q\cdot\cdot\cdot k}$ and $\hat{C}_{q\cdot\cdot}$,

we obtain

$$\hat{\bar{X}}_{q\cdots k} = \frac{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}X_{qij\ell k}/Z_{qij}}{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}} . \tag{8}$$

4.3 Estimators of Variances and Covariances

The population variance of $X_{qij\ell k}$ within stratum q is

$$V\left(X_{qij\ell k}\right) = E\left(X_{qij\ell k}-X_{q\cdots k}\right)^2 . \tag{9}$$

In the last subsection, we derived a ratio estimator of the mean $\bar{X}_{q\cdots k}$. Since the population variance is also a mean, a similar ratio estimator of the population variance is given by

$$\hat{V}(X_{qij\ell k}) = \frac{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}\left(X_{qij\ell k}-\bar{X}_{q\cdots k}\right)^2/Z_{qij}}{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}} . \tag{10}$$

The numerator of the right hand side of equation (10) can be decomposed into three parts as follows:

$$\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}\left(X_{qij\ell k}-\bar{X}_{q\cdots k}\right)^2/Z_{qij}$$

$$= \frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}\left(X_{qij\ell k}-\hat{\bar{X}}_{qi\cdot\cdot k}\right)^2/Z_{qij}$$

$$+ \frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\left(\hat{\bar{X}}_{qi\cdot\cdot k}-\hat{\bar{X}}_{q\cdots k}\right)^2\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}$$

$$+ \left(\hat{\bar{X}}_{q\cdots k}-\bar{X}_{q\cdots k}\right)^2\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij} . \tag{11}$$

By substituting the above three parts for the numerator in equation (10), we obtain

$$\hat{V}(X_{qij\ell k}) = \frac{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}\left(X_{qij\ell k}-\hat{\bar{X}}_{qi\cdot\cdot k}\right)^2/Z_{qij}}{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}}$$

$$+ \frac{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\left(\hat{\bar{X}}_{qi\cdot\cdot k}-\hat{\bar{X}}_{q\cdots k}\right)^2\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}}{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}}$$

$$+ \left(\hat{\bar{X}}_{q\cdots k}-\bar{X}_{q\cdots k}\right)^2 . \tag{12}$$

The estimator of the variance given in equation (12) is the sum of three components. The first component is an estimate of the variance within first-stage units in stratum q. The second component is an estimate of the variance between first-stage units in stratum q. The third component cannot be computed because $\bar{X}_{q\cdots k}$ is unknown. However, if we ignore the bias, it can be estimated by $\hat{V}(\hat{\bar{X}}_{q\cdots k})$, which is given later. Since the sample size in MIS is quite large, it is expected that the bias will be small.

Similarly, an estimate of the population covariance between $X_{qij\ell k}$ and $X_{qij\ell r}$ can be shown to be the sum of (i) the covariance within first-stage units, (ii) the covariance between first-stage units, and (iii) the product of the differences between population means and their estimates. Equation (13) gives the formula for an estimate of the population covariance:

$$\hat{Cov}(X_{qij\ell k},X_{qij\ell r}) = \frac{1}{m_q}\left\{\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\left(X_{qij\ell k}-\hat{\bar{X}}_{qi\cdot\cdot k}\right)\cdot\left(X_{qij\ell r}-\hat{\bar{X}}_{qi\cdot\cdot r}\right)\frac{\hat{C}_{qij}}{Z_{qij}}\right\} \Big/ \left[\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}\right]$$

$$+ \frac{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\left(X_{qi\cdot\cdot k}-\hat{\bar{X}}_{q\cdots k}\right)\left(X_{qi\cdot\cdot r}-\hat{\bar{X}}_{q\cdots r}\right)\sum_{j=1}^{n_{qi}}\frac{\hat{C}_{qij}}{Z_{qij}}}{\frac{1}{m_q}\sum_{i=1}^{m_q}\frac{1}{n_{qi}W_{qi}}\sum_{j=1}^{n_{qi}}\hat{C}_{qij}/Z_{qij}}$$

$$+ (\hat{\bar{X}}_{q\cdots k}-\bar{X}_{q\cdots k})(\hat{\bar{X}}_{q\cdots r}-\bar{X}_{q\cdots r}) . \tag{13}$$

The last term on the right-hand side can be estimated by $\hat{Cov}(\hat{\bar{X}}_{q\cdots k},\hat{\bar{X}}_{q\cdots r})$.

As indicated earlier, the third component on the right hand side of Equations (12) and (13) can be estimated by the variance and covariance of the means for stratum q. These estimators are given by the following expressions:

$$\hat{V}(\hat{\bar{X}}_{q\cdots k}) = \frac{\frac{1}{m_q(m_q-1)} \sum_{i=1}^{m_q} \left[\frac{1}{n_{qi}W_{qi}} \sum_{j=1}^{n_{qi}} \left(X_{qijlk} - \hat{\bar{X}}_{q\cdots k}\right) \frac{\hat{C}_{qij}}{Z_{qij}}\right]^2}{\left[\frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi}W_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij}/Z_{qij}\right]^2} \tag{14}$$

and

$$\hat{C}ov(\hat{\bar{X}}_{q\cdots k}, \hat{\bar{X}}_{q\cdots r}) = \frac{1}{m_q(m_q-1)} \sum_{i=1}^{m_q} \left\{\left(\frac{1}{n_{qi}W_{qi}}\right)^2 \cdot \left[\sum_{j=1}^{n_{qi}} \left(X_{qijlk} - \hat{\bar{X}}_{q\cdots k}\right) \frac{\hat{C}_{qij}}{Z_{qij}}\right] \cdot \left[\sum_{j=1}^{n_{qi}} \left(X_{qijlr} - \hat{\bar{X}}_{q\cdots r}\right) \frac{\hat{C}_{qij}}{Z_{qij}}\right]\right\} \Big/ \left[\frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi}W_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij}/Z_{qij}\right]^2 . \tag{15}$$

5. APPLICATION TO MULTIVARIATE METHODS

In this section, we discuss how the ideas and estimators developed in the preceding sections can be used in various multivariate analyses.

5.1 Multivariate Normal

The estimators described in Section 4 can be immediately applied to standard multivariate normal analyses in which the analysis requires only the sample mean vector and the sample covariance matrix. For example, principal components, linear discriminant functions, canonical correlations, and regression coefficients can be obtained for the within, between, or overall level. The choice of which level to use depends on the purpose of the analysis.

5.2 Extensions to Logistic Regression

Some multivariate analyses require more than a mean vector and covariance matrix. For many of these analyses, however, we believe that the ideas presented in the previous sections can still be useful, even if the specific estimators are not. In particular, we suggest weighted analyses of observations of second-stage units. We recommend that the observation of second-stage unit (q, i, j) be assigned the weight

$$\frac{\hat{C}_{qij}}{m_q n_{qi} W_{qi} Z_{qij}} .$$

The MIS includes a number of dichotomous variables. Logistic regression allows us to analyze the dependence of a dichotomous variable on other variables. In order to perform such an analysis, we suggest two modifications of the usual procedure.

First, a logistic regression typically gives weights $\hat{P}_{qij}(1-\hat{P}_{qij})$ to the observations, where $\hat{P}_{qij}$ is the estimated probability of the dependent variable taking on value 1 given the values of the independent variables for that observation. In order to make the weights proportional to both $\hat{P}_{qij}(1-\hat{P}_{qij})$ and those recommended in the beginning of this subsection, we suggest weights proportional to

$$\frac{\hat{P}_{qij}(1-\hat{P}_{qij})\hat{C}_{qij}}{m_q n_{qi} W_{qi} Z_{qij}} .$$

Second, since we are primarily interested in the variation within first-stage units, it is necessary to exclude the variation between first-stage units. We recommend that this be accomplished by introducing additional independent variables, one indicator variable for each first-stage unit. The variances and covariances of the variables of interest are then conditioned on these indicator variables.

6. ESTIMATES OF THE SIZE OF SECOND-STAGE UNITS

Before we can apply the estimators of the last section, it is necessary to estimate C_{qij}. Here, we discuss five estimates of C_{qij} associated with the MIS.

The first estimate of C_{qij} (subject to a constant of proportionality) is A_{qij}, which is the total number of all outgoing and incoming telephone calls during an 8-hour period.

The second estimate of C_{qij} (subject to a constant of proportionality) is obtained by multiplying the relative sizes of first-stage and second-stage units in the sample as follows:

$$\hat{C}_{qij} = W_{qi} Z_{qij} .$$

It may be recalled that

$$W_{qi} = \frac{N_{qi}}{N_{q\cdot}} \quad \text{and} \quad Z_{qij} = \frac{T_{qij}}{T_{qi\cdot}} .$$

The third estimate C_{qij} is obtained by taking a weighted geometric mean of the first and second estimates given above. This estimate is

$$\hat{C}_{qij} = (A_{qij})^{\alpha} (W_{qi} Z_{qij})^{1-\alpha} ,$$

where $0 \leq \alpha \leq 1$. The first two estimates are special cases of the third estimator, since they correspond to $\alpha = 1$ and $\alpha = 0$, respectively.

The fourth estimate of C_{qij} is designed to give equal weight to each third-stage unit in the sample. This estimate is given by

$$\hat{C}_{qij} = n_{qi} W_{qi} Z_{qij} .$$

The last estimate of C_{qij} is designed to produce the mean over the population of second-stage units. In this case,

$$\hat{C}_{qij} = 1 .$$

7. AGGREGATION OVER STRATA

So far we have discussed estimators for an individual stratum. Estimators for individual strata can be combined to obtain estimators for the whole population:

$$\hat{X}_{\ldots\ldots k} = \sum_{q=1}^{L} \hat{X}_{q\ldots k} ,$$

$$\hat{C}_{\ldots} = \sum_{q=1}^{L} \hat{C}_{q\ldots} ,$$

and

$$\hat{\bar{X}}_{\ldots\ldots k} = \hat{X}_{\ldots\ldots k} / \hat{C}_{\ldots} .$$

By substituting the expressions given in equations (6) and (7) in Section 4 for $\hat{X}_{q\ldots k}$ and $\hat{C}_{q\ldots}$, we obtain

$$\hat{\bar{X}}_{\ldots\ldots k} = \frac{\sum_{q=1}^{L} \frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi} W_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij} X_{qijlk} / Z_{qij}}{\sum_{q=1}^{L} \frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi} W_{qi}} \sum_{j=1}^{n_{qi}} \hat{C}_{qij} / Z_{qij}} . \qquad (16)$$

An estimate of the variance of $\hat{\bar{X}}_{\ldots\ldots k}$ is given by

$$\hat{V}(\hat{\bar{X}}_{\ldots\ldots k}) = \frac{\sum_{q=1}^{L} \frac{1}{m_q (m_q - 1)} \sum_{i=1}^{m_q} \left[\frac{1}{n_{qi} W_{qi}} \sum_{j=1}^{n_{qi}} \left(X_{qijlk} - \hat{\bar{X}}_{\ldots\ldots k} \right) \frac{\hat{C}_{qij}}{Z_{qij}} \right]^2}{\left[\sum_{q=1}^{L} \frac{1}{m_q} \sum_{i=1}^{m_q} \frac{1}{n_{qi} W_{qi}} \sum_{j=1}^{n_{qi}} \frac{\hat{C}_{qij}}{Z_{qij}} \right]^2} . \qquad (17)$$

8. SUMMARY

Techniques for estimating means and their variances from multistage sampling survey data are well-known. In this paper, we have discussed some concepts related to multivariate analyses of such data. We have also discussed the estimation of variances and covariances which provide a basis for many multivariate analyses. Application of the estimators discussed here to multistage sampling surveys has been illustrated by showing the correspondence with MIS.

The total variation was broken down into that within and between first-stage units. Which of these two components of variation should be used depends on the application.

Estimators of means, variances and covariances in a specified stratum have been developed for the three-stage sampling scheme in the MIS. These estimators were combined over strata to obtain estimators for the whole population.

In MIS, sizes of second-stage units are not known. Five estimators of these sizes were discussed. Software for implementing the estimators discussed in this paper has been developed.

9. ACKNOWLEDGEMENTS

The authors would like to thank S. R. Dalal, J. D. Healy, J. M. Landwehr, M. G. Linnell, J. E. McRae, J. T. Page and A. K. Severinsen of Bell Laboratories for several helpful discussions.

REFERENCES

[1] Cochran, W. G. (1977), Sampling Techniques, Third Edition, John Wiley and Sons.

[2] Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), Sample Survey Methods and Theory, Vols. I and II, John Wiley and Sons.

[3] Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," JASA, 72, 320-338.

[4] Searle, S. R. (1971), "Topics in Variance Component Estimation", Biometrics, 27, 1-76.

[5] Seeger, P. (1970), "A Method of Estimating Variance Components in Unbalanced Designs", Technometrics, 12, 207-218.

[6] Bean, J. A. and Schnack, G. A. (1977), "Application of Balanced Repeated Replication to the Estimation of Variance Components, Proceedings of the Social Statistics Section, American Statistical Association, Washington, D. C., 938-942.

[7] Folsom, R. E., Bayless, D. L., and Shah, B. V. (1971), "Jackknifing for Variance Components in Complex Survey Designs", Proceedings of the Social Statistics Section, American Statistical Association, Washington, D. C. 36-39.