Ann. Inst. Statist. Math.
34 (1982), Part A, 91-103

ESTIMATION IN STRATIFIED RANDOM SAMPLING: ADJUSTMENT FOR CHANGES IN STRATA COMPOSITION*

ARIDAMAN K. JAIN

(Received June 30, 1981; revised Sept. 4, 1981)

Abstract

Panel studies are widely used to collect data on consumer expenditures, labor force participation and other demographic variables. Frequently, the population changes significantly over the period of interest. We consider the case when a stratified random sample is drawn with probability proportional to initial size (pps) and the value of one of the stratification factors changes over time. Information on such changes, which result in the movement of primary units from their original strata to new strata, is readily available for the sampled units but not for the nonsampled units. As a result of changes in the composition of strata, the stratified sample no longer corresponds to a pps selection within each stratum. A method of estimation based on the original selection of the sample but which incorporates the subsequent changes is proposed. It is shown that these estimates are approximately unbiased.

1. Introduction

Panel studies are widely used to collect data on consumer expenditures, labor force participation and other demographic variables. After the initial selection, the sampled units may be used for several surveys over a period of several years. During this period, there may be significant and frequent changes in the population. Since changes would continue to occur in the future, a single brand new selection of sample units is not a satisfactory solution. Moreover, continued use of the original sample has several advantages over a brand new sample: (i) it is much cheaper, (ii) it provides a more precise comparison of the results of successive surveys, and (iii) it avoids the delay in the availability of results due to the start-up of a new sample.

The problems of changing probabilities within fixed strata and

* An earlier version of this paper was presented at the 1978 ASA Meetings (Jain [3]).

changing strata for primary units in panel studies have been discussed by several authors. Keyfitz [4] presented a procedure for adjusting a sample, consisting of one primary unit from each of a number of strata, chosen with probabilities proportional to an earlier measure of size to a sample chosen with probabilities proportional to a later measure of size. His procedure for maximizing the retention of old units was as follows: (i) if the selection probability of an old sample unit is not decreased, retain the old unit; (ii) if the selection probability of an old sample unit is decreased, the probability of replacing this unit by a new unit is the relative decrease in its selection probability.

Kish and Hess [5] proposed a reduction in the expected number of changes required by Keyfitz's procedure by retaining the old units for which the reduction in probability was small. Fellegi [2] presented a generalization of Keyfitz's procedure for the case when two units are selected without replacement with probability proportional to size from each stratum. Kish and Scott [6] extended the Kish and Hess procedure of maximum retention of initial units to the case of changing strata.

Here we discuss an estimation method in which the sample units undergo extensive testing before they can be used for data collection. We consider the case when primary units within a stratum are selected with probability proportional to initial size (pps) and the value of one of the stratification factors may change over time. Information on such changes, which result in the movement of primary units from their original strata to new strata, is readily available for the sampled units but not for the nonsampled units. As a result of changes in the composition of strata, the stratified sample no longer corresponds to a pps selection within each stratum.

We propose a method of estimation based on the original sample but which incorporates the subsequent changes in the sample. Major advantages of this procedure, which yields approximately unbiased estimators, are that it does *not* require

(i) information on changes in the nonsampled units,

(ii) replacement of any units in the original sample by new units.

The proposed procedure differs from those discussed in the literature, cited above, in that it does not require the replacement of any units in the original sample by new units.

The method discussed here was motivated by a telephone company measurement plan based on a panel of plant facilities where the type of facility is a principal factor for stratification. However, the techniques discussed apply to panel studies in general. For example, a consumer panel may be a stratified random sample with family income as a stratification factor. A significant change in family income of a

panel member would result in a change in strata composition. Section 2 gives some background on the panel and notation. Section 3 contains the development of the estimator of the population proportion of elements that possess a specified characteristic. Section 4 includes a proof of the approximate unbiasedness of the proportion estimator. Section 5 gives a summary and points out two possible generalizations of the estimation procedure discussed here.

2. Background and notation

We will discuss the panel design for the case when strata are combinations of two factors. But it can be generalized to the case of more than two factors. In each stratum, sampling was done in two stages. Again, the methodology discussed here can be generalized to sampling with more than two stages. The following method was used for the selection of the original sample: (i) first-stage units were drawn with replacement with probability proportional to size (in this case, the number of second-stage units within the first-stage unit) and (ii) second-stage units were selected by simple random sampling. Each second-stage unit consists of a cluster of elements, each of which is examined to determine whether it possesses a specified characteristic (say C). In the telephone company measurement plan, possession of characteristic C corresponds to an error in the recording of a telephone call. The proportion of elements that possess characteristic C is the quantity of primary interest.

While the notation below is in general terms, the corresponding specific definitions for the telephone company measurement plan are indicated in parentheses to facilitate understanding of the estimation procedure. Let

i=value of stratification factor 1 (geographical area), $i=1, 2, \cdots, I$;

j=*original* value of stratification factor 2 (switching machine type), $j=1, 2, \cdots, J$;

u=*current* value of stratification factor 2, $u=1, 2, \cdots, J$;

M_{ij}=total number of first-stage units (switching machines) in stratum (i, j);

m_{ij}=number of first-stage units in the sample from stratum (i, j), $m_{ij} \leqq M_{ij}$;

M_{iju}=number of first-stage units in stratum (i, j) out of M_{ij} that are currently of type u, $\sum_{u=1}^{J} M_{iju}=M_{ij}$;

m_{iju}=number of first-stage units in the sample from stratum (i, j) out of m_{ij} that are currently of type u, $\sum_{u=1}^{J} m_{iju}=m_{ij}$;

$M_{i.}=\sum_{j=1}^{J} M_{ij}$;

$m_{i.}=\sum_{j=1}^{J} m_{ij}$;

v_{ijq}=number of second-stage units (number of telephone lines) for first-stage unit q in stratum (i, j) at the time of the original sample selection, $q=1, 2, \cdots, M_{ij}$;

v_{ijq}^{N}=current value of v_{ijq};

w_{ijq}=number of times first-stage unit (i, j, q) has been selected[1] in the sample;

$v_{ij.}=\sum_{q=1}^{M_{ij}} v_{ijq}$;

l_{ijq}=number of second-stage units (telephone lines) selected from first-stage unit[2] q in stratum (i, j), $q=1, 2, \cdots, m_{ij}$;

X_{ijqr}=total number of elements (telephone calls) in selected second-stage unit (i, j, q, r), $r=1, 2, \cdots, l_{ijq}$;

Y_{ijqr}=number of elements in selected second-stage unit (i, j, q, r) which have characteristic C, $Y_{ijqr} \leqq X_{ijqr}$;

$X_{ijq.}=\sum_{r=1}^{v_{ijq}^{N}} X_{ijqr}$;

$Y_{ijq.}=\sum_{r=1}^{v_{ijq}^{N}} Y_{ijqr}$;

$p_{ijq}=Y_{ijq.}/X_{ijq.}$=proportion of elements in first-stage unit (i, j, q) that have characteristic C.

As described above, the panel of facilities consists of a two-stage sample from strata that are combinations of switching machine type and geographical area. After the initial selection, the process of modernization results in the replacement of some of the older types of first-stage units[3] (i.e., switching machines) by newer types of units. Since switching machine type is a stratification factor, a change in switching machine type results in the movement of a unit from one stratum to another stratum and consequently it changes strata composition and selection probabilities. The next section describes an estimation procedure that does not require information on changes in the nonsampled units.

---

[1] It may be recalled that the sampling of first-stage units is with replacement. Therefore, the number of selections of a sample first-stage unit can be more than one. The number of second-stage units selected from a first-stage unit is proportional to the number of selections of the first-stage unit. The second-stage units are selected by simple random sampling.

[2] It is assumed that the M_{ij} first-stage units in stratum (i, j) are reordered after sampling, so that the first m_{ij} first-stage units correspond to those in the sample.

[3] In the subsequent discussion, a first-stage unit will be referred to simply as a unit.

3. Estimators for population proportions

Because of the replacement of some units, our interest has shifted from original strata (i, j), from which the panel units were selected, to new strata (i, u) which correspond to the current breakdown of the population of units. The estimates for the new strata (i, u) are derived by first subdividing each original stratum (i, j) into *domains* of current switch type $(u=1, 2, \cdots, J)$ and then summing over j for each geographical area i. We discuss the development of estimators for the proportion of elements with characteristic C for units within a stratum as well as for higher levels of aggregation.

3.1. *Estimators for unit q in original stratum* (i, j)

Estimators for the total number of elements $(X_{ijq\cdot})$ and the number of elements which have characteristic C $(Y_{ijq\cdot})$ for unit[4] q in original stratum (i, j) are as follows:

$$\hat{X}_{ijq\cdot}=\frac{v_{ijq}^{N}}{l_{ijq}}\sum_{r=1}^{l_{ijq}}X_{ijqr} \tag{1}$$

and

$$\hat{Y}_{ijq\cdot}=\frac{v_{ijq}^{N}}{l_{ijq}}\sum_{r=1}^{l_{ijq}}Y_{ijqr}\ , \tag{2}$$

where a hat is used to denote an estimator of a population quantity. Then an estimator for the proportion of elements that have characteristic C in unit (i, j, q) is

$$\hat{p}_{ijq}=\frac{\hat{Y}_{ijq\cdot}}{\hat{X}_{ijq\cdot}}\ . \tag{3}$$

A variance estimator for $\hat{p}_{ijq}$ (denoted by v) is

$$v(\hat{p}_{ijq})=\frac{\sum_{r=1}^{l_{ijq}}[Y_{ijqr}-\hat{p}_{ijq}X_{ijqr}]^2}{l_{ijq}(l_{ijq}-1)X_{ijq\cdot}^2}\ . \tag{4}$$

Formula (4) is based on Theorem 2.5 in Cochran [1].

3.2. *Estimators for domain u in original stratum* (i, j)

In the subsequent discussion, we will replace a subscript by a dot to indicate either summation over telephone lines (X and Y) or aver-

[4] As in the case of the first-stage units, it is assumed that the v_{ijq} second-stage units in first-stage unit (i, j, q) are reordered after sampling, so that the first l_{ijq} second-stage units correspond to those in the sample.

aging for the proportion of elements with characteristic C. Since the subscript in the fourth place is not needed to denote the line number (we have already summed over lines), it will be used to denote the current switch type (u). Let

$$X_{ijqu}=\begin{cases} X_{ijq\cdot}\,, & \text{if the current switch type of unit } (i,\ j,\ q) \text{ is } u\,,\\ 0\,, & \text{otherwise;}\end{cases}$$

$$Y_{ijqu}=\begin{cases} Y_{ijq\cdot}\,, & \text{if the current switch type of unit } (i,\ j,\ q) \text{ is } u\,,\\ 0\,, & \text{otherwise;}\end{cases}$$

$$X_{ij\cdot u}=\sum_{q=1}^{M_{ij}} X_{ijqu}\ ;$$

$$Y_{ij\cdot u}=\sum_{q=1}^{M_{ij}} Y_{ijqu}\ ;$$

$$p_{ij\cdot u}=Y_{ij\cdot u}/X_{ij\cdot u}\ .$$

If $m_{iju}=0$, we define

$$\hat{X}_{ij\cdot u}=\hat{Y}_{ij\cdot u}=\hat{p}_{ij\cdot u}=v(\hat{p}_{ij\cdot u})=0\ . \tag{5}$$

If $m_{iju}\geqq 1$, estimators for $X_{ij\cdot u}$ and $Y_{ij\cdot u}$ are given as follows:

$$\hat{X}_{ij\cdot u}=\frac{1}{w_{ij\cdot}}\sum_{q=1}^{m_{ij}}\frac{v_{ij\cdot}}{v_{ijq}}X_{ijqu}\cdot w_{ijq} \tag{6}$$

and

$$\hat{Y}_{ij\cdot u}=\frac{1}{w_{ij\cdot}}\sum_{q=1}^{m_{ij}}\frac{v_{ij\cdot}}{v_{ijq}}X_{ijqu}\cdot\hat{p}_{ijq\cdot}\cdot w_{ijq}\ . \tag{7}$$

Then

$$\hat{p}_{ij\cdot u}=\hat{Y}_{ij\cdot u}/\hat{X}_{ij\cdot u} \tag{8}$$

and

$$v(\hat{p}_{ij\cdot u})=\frac{1}{w_{ij\cdot}(w_{ij\cdot}-1)\hat{X}^2_{ij\cdot u}}\sum_{q=1}^{m_{ij}}\left[\left\{\frac{v_{ij\cdot}}{v_{ijq}}X_{ijqu}\right\}^2 w_{ijq}\cdot(\hat{p}_{ijq\cdot}-\hat{p}_{ij\cdot u})^2\right]\ , \tag{9}$$

where

$$v_{ij\cdot}=\sum_{q=1}^{M_{ij}} v_{ijq} \qquad \text{and} \qquad w_{ij\cdot}=\sum_{q=1}^{m_{ij}} w_{ijq}\ .$$

If $m_{iju}=1$, then there is only *one* sample unit in stratum $(i,\ j)$ that

is currently of type u. In this case there is a value of q (say q_1) such that $p_{ij\cdot u}=p_{ijq_1}$. and it is not possible to use formula (9). When $m_{iju}=1$, we suggest the use of the following formula:

$$v(\hat{p}_{ij\cdot u})=v(\hat{p}_{ijq_1\cdot}) \ .$$

In Section 4 we show that $\hat{X}_{ij\cdot u}$ and $\hat{Y}_{ij\cdot u}$ are unbiased estimators of $X_{ij\cdot u}$ and $Y_{ij\cdot u}$, respectively. Therefore, it seems reasonable to expect that, in most cases, $\hat{p}_{ij\cdot u}=\hat{Y}_{ij\cdot u}/\hat{X}_{ij\cdot u}$ is an approximately unbiased estimator of $p_{ij\cdot u}=Y_{ij\cdot u}/X_{ij\cdot u}$.

3.3. *Estimators for new stratum* (i, u)

Now for each combination of domain (u) and geographical area (i), the estimators for population totals are obtained by summation over the original strata corresponding to the original switch type (j) as follows:

$$\hat{X}_{i\cdot\cdot u}=\sum_{j=1}^{J}\hat{X}_{ij\cdot u} \tag{10}$$

and

$$\hat{Y}_{i\cdot\cdot u}=\sum_{j=1}^{J}\hat{Y}_{ij\cdot u}=\sum_{j=1}^{J}\hat{p}_{ij\cdot u}\hat{X}_{ij\cdot u} \ . \tag{11}$$

Then, the estimator for the proportion of elements in new stratum (i, u) with characteristic C is given by

$$\hat{p}_{i\cdot\cdot u}=\hat{Y}_{i\cdot\cdot u}/\hat{X}_{i\cdot\cdot u}=\sum_{j=1}^{J}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)\hat{p}_{ij\cdot u} \ . \tag{12}$$

An estimator for the variance of $\hat{p}_{i\cdot\cdot u}$ which is derived in Subsection 3.4 is

$$\begin{aligned} v(\hat{p}_{i\cdot\cdot u})=&\sum_{j=1}^{J}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)^2 v(\hat{p}_{ij\cdot u}) \\ &+\sum_{j=1}^{J}(\hat{p}_{ij\cdot u})^2\cdot\left[\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^2_{i\cdot\cdot u}}+\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^4_{i\cdot\cdot u}}\cdot\sum_{j'=1}^{J}v(\hat{X}_{ij'\cdot u})\right] \\ &+\sum_{j=1}^{J}v(\hat{p}_{ij\cdot u})\cdot\left[\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^2_{i\cdot\cdot u}}+\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^4_{i\cdot\cdot u}}\sum_{j'=1}^{J}v(\hat{X}_{ij'\cdot u})\right] , \end{aligned} \tag{13}$$

where

$$v(\hat{X}_{ij\cdot u})=\frac{1}{(w_{ij\cdot})(w_{ij\cdot}-1)}\sum_{q=1}^{m_{ij}}\left[\left(\frac{v_{ij\cdot}}{v_{ijq}}\right)X_{ijqu}-\hat{X}_{ij\cdot u}\right]^2 w_{ijq} \ .$$

The above expression for $v(\hat{p}_{i\cdot\cdot u})$ can be simplified to the following ex-

pression when the variation in $\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u}$ is expected to be small relative to the variation in $\hat{p}_{ij\cdot u}$:

$$v(\hat{p}_{i\cdot\cdot u})=\sum_{j=1}^{J}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)^2 v(\hat{p}_{ij\cdot u})\ . \tag{14}$$

It may be noted that the variance estimator given by formula (14) tends to underestimate the true variance on account of the fact that the variation in the ratios $\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u}$ is being ignored. It is assumed in the above that $\hat{X}_{i\cdot\cdot u}$ is not zero. If it were, then $\hat{p}_{i\cdot\cdot u}$ is not estimable.

It may also be noted that if none of the sample units is replaced by one of type u, then

$$\hat{X}_{i\cdot\cdot u}=\hat{X}_{iu\cdot u}\ ,$$

$$\hat{Y}_{i\cdot\cdot u}=\hat{Y}_{iu\cdot u}\ ,$$

$$\hat{p}_{i\cdot\cdot u}=\hat{p}_{iu\cdot u}\ ,$$

and

$$v(\hat{p}_{i\cdot\cdot u})=v(\hat{p}_{iu\cdot u})\ .$$

In this case, the variance estimator given by formula (14) would not tend to underestimate the true variance.

3.4. *Derivation of an estimator for* $V(\hat{p}_{i\cdot\cdot u})$

This subsection contains the derivation of an estimator for the true variance of $\hat{p}_{i\cdot\cdot u}$, $V(\hat{p}_{i\cdot\cdot u})$, given by equation (13) in the last subsection. The development in this subsection does not ignore the variability in $\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u}$. Recall that

$$\hat{p}_{i\cdot\cdot u}=\sum_{j=1}^{J}\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\hat{p}_{ij\cdot u}\ . \tag{15}$$

Based on some empirical evidence from the telephone company measurement plan that motivated the method discussed here, we assume that $\hat{p}_{ij\cdot u}$, which is the fraction of $\hat{X}_{ij\cdot u}$ elements that have characteristic C, is statistically independent of $\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u}$. Under this independence assumption, we have

$$\begin{aligned} V(\hat{p}_{i\cdot\cdot u})=&\sum_{j=1}^{J}\left(\mathrm{E}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)\right)^2\cdot V(\hat{p}_{ij\cdot u})+\sum_{j=1}^{J}[\mathrm{E}\,(\hat{p}_{ij\cdot u})]^2\cdot V\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)\\ &+\sum_{j=1}^{J}V\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)\cdot V(\hat{p}_{ij\cdot u})\ . \end{aligned} \tag{16}$$

The following estimator of $V(\hat{p}_{i\cdot\cdot u})$ is obtained by replacing the expected

values of the statistics $\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u}$ and $\hat{p}_{ij\cdot u}$ by the respective statistics and the true variances $V(\hat{p}_{ij\cdot u})$ and $V(\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u})$ by their respective estimators:

$$(17)\qquad v(\hat{p}_{i\cdot\cdot u})=\sum_{j=1}^{J}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)^2\cdot v(\hat{p}_{ij\cdot u})+\sum_{j=1}^{J}(\hat{p}_{ij\cdot u})^2\cdot v\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)$$
$$+\sum_{j=1}^{J}v\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)\cdot v(\hat{p}_{ij\cdot u})\ .$$

The estimator $v(\hat{p}_{ij\cdot u})$ was given earlier by formula (9) and an estimator of $V(\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u})$ is presented below.

$V(\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u})$ can be computed as the sum of the average conditional variance and the variance of the conditional average:

$$V\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)=\mathrm{E}\left[V\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\,\middle|\,\hat{X}_{i\cdot\cdot u}\right)\right]+V\left[\mathrm{E}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\,\middle|\,\hat{X}_{i\cdot\cdot u}\right)\right]\ .$$

Conditional on $\hat{X}_{i\cdot\cdot u}$, $\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u}$ is an unbiased estimate of the unknown binomial proportion $X_{ij\cdot u}/X_{i\cdot\cdot u}$. Therefore,

$$V\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\,\middle|\,\hat{X}_{i\cdot\cdot u}\right)=\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\left(1-\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\right)\frac{1}{\hat{X}_{i\cdot\cdot u}}$$

and

$$\mathrm{E}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\,\middle|\,\hat{X}_{i\cdot\cdot u}\right)=\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\ .$$

Now taking the expected value of the conditional variance and variance of the conditional expected value, we obtain

$$V\left[\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right]=\mathrm{E}\left[\frac{X_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\left(1-\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\right)\cdot\frac{1}{\hat{X}_{i\cdot\cdot u}}\right]+V\left[\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\right]$$
$$=\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\left(1-\frac{X_{ij\cdot u}}{X_{i\cdot\cdot u}}\right)\mathrm{E}\left(\frac{1}{\hat{X}_{i\cdot\cdot u}}\right)$$
$$\doteq\frac{V(\hat{X}_{ij\cdot u})}{X_{i\cdot\cdot u}}\left[\frac{1}{X_{i\cdot\cdot u}}+\frac{V(\hat{X}_{i\cdot\cdot u})}{X_{i\cdot\cdot u}^3}\right]$$
$$=\frac{V(\hat{X}_{ij\cdot u})}{X_{i\cdot\cdot u}^2}+\frac{V(\hat{X}_{ij\cdot u})V(\hat{X}_{i\cdot\cdot u})}{X_{i\cdot\cdot u}^4}\ .$$

By replacing the population quantities $X_{i\cdot\cdot u}$ and $X_{ij\cdot u}$ by their corresponding estimators and the true variances by their corresponding estimators, we have the following estimator of $V(\hat{X}_{ij\cdot u}/\hat{X}_{i\cdot\cdot u})$:

$$v\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)=\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^2_{i\cdot\cdot u}}+\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^4_{i\cdot\cdot u}}\sum_{j=1}^{J} v(\hat{X}_{ij\cdot u})\ . \tag{18}$$

Recall that

$$\hat{X}_{ij\cdot u}=\frac{1}{w_{ij\cdot}}\sum_{q=1}^{m_{ij}}\frac{v_{ij\cdot}}{v_{ijq}}X_{ijqu}w_{ijq}\ ,$$

as given in Subsection 3.2. Therefore, we take, as our estimator for the variance of $\hat{X}_{ij\cdot u}$,

$$v(\hat{X}_{ij\cdot u})=\frac{1}{w_{ij\cdot}(w_{ij\cdot}-1)}\sum_{q=1}^{m_{ij}}\left[\frac{v_{ij\cdot}}{v_{ijq}}X_{ijqu}-\hat{X}_{ij\cdot u}\right]^2 w_{ijq}\ . \tag{19}$$

Now putting together all the pieces, we obtain

$$\begin{aligned} v(\hat{p}_{i\cdot\cdot u})=&\sum_{j=1}^{J}\left(\frac{\hat{X}_{ij\cdot u}}{\hat{X}_{i\cdot\cdot u}}\right)^2 v(\hat{p}_{ij\cdot u})\\ &+\sum_{j=1}^{J}(\hat{p}_{ij\cdot u})^2\cdot\left[\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^2_{i\cdot\cdot u}}+\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^4_{i\cdot\cdot u}}\sum_{j'=1}^{J} v(\hat{X}_{ij'\cdot u})\right]\\ &+\sum_{j=1}^{J} v(\hat{p}_{ij\cdot u})\cdot\left[\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^2_{i\cdot\cdot u}}+\frac{v(\hat{X}_{ij\cdot u})}{\hat{X}^4_{i\cdot\cdot u}}\sum_{j'=1}^{J} v(\hat{X}_{ij'\cdot u})\right]\ , \end{aligned} \tag{20}$$

where $v(\hat{X}_{ij\cdot u})$ is given by equation (19). This equation was given earlier as equation (13).

3.5. *Other estimators*

Estimators for the total number of elements and the number of elements with characteristic C for a level of one factor (i.e., when combined over the levels of the other factor) and for the whole population can be derived by appropriate summations over the strata values. Then the estimators for proportions of elements that have characteristic C can be obtained by taking appropriate ratios.

4. Approximate unbiasedness of $\hat{p}_{ij\cdot u}$

In this section we show that $\hat{X}_{ij\cdot u}$ and $\hat{Y}_{ij\cdot u}$ are unbiased estimators of $X_{ij\cdot u}$ and $Y_{ij\cdot u}$, respectively. Therefore, it seems reasonable to expect that, in most cases, $\hat{p}_{ij\cdot u}=\hat{Y}_{ij\cdot u}/\hat{X}_{ij\cdot u}$ is an approximately unbiased estimator of $p_{ij\cdot u}$. There are m_{iju} units in the sample in original stratum (i, j) which are currently of switch type u. Let us assume that the m_{ij} units in the sample in original stratum (i, j) are reordered, so that the first m_{iju} units are currently of switch type u. Similarly, let the M_{ij} units in the population in original stratum (i, j) be reordered

so that the first M_{iju} units are currently of switch type u. Then the probability of selection of a unit in (i, j, u) is $v_{ijqu}/v_{ij\cdot u}$, where

$$v_{ij\cdot u}=\sum_{q=1}^{M_{iju}} v_{ijqu}$$

and

$$v_{ijqu}=\begin{cases} v_{ijq} & \text{if unit } q \text{ in stratum } (i, j) \text{ is currently of switch type } u, \\ 0 & \text{otherwise.} \end{cases}$$

An unbiased estimator of the total number of elements $X_{ij\cdot u}$ in the M_{iju} units of type u is

$$\hat{X}_{ij\cdot u}=\left[\sum_{q=1}^{m_{iju}} \left\{X_{ijq}\Big/\left[\frac{v_{ijqu}}{v_{ij\cdot u}}\right]\right\} w_{ijq}\right]\Big/\sum_{q=1}^{m_{iju}} w_{ijq}\,.$$

But it is quite difficult to determine $v_{ij\cdot u}$. However, we know that the probability of selection of a unit of type u is $v_{ij\cdot u}/v_{ij\cdot}$. Since m_{iju} of the m_{ij} sample units are of type u, an unbiased estimate of $v_{ij\cdot u}/v_{ij\cdot}$ is $\sum_{q=1}^{m_{iju}} w_{ijq}\Big/\sum_{q=1}^{m_{ij}} w_{ijq}$. By substituting this estimate for $v_{ij\cdot u}/v_{ij\cdot}$ in the above expression for $\hat{X}_{ij\cdot u}$, we obtain the following expression which is algebraically equivalent to the estimator given by formula (6):

$$\begin{aligned}\hat{X}_{ij\cdot u}&=\left(\sum_{q=1}^{m_{iju}} w_{ijq}\cdot X_{ijq}\cdot\frac{v_{ij\cdot}}{v_{ijqu}}\right)\cdot\left(\frac{\sum_{q=1}^{m_{iju}} w_{ijq}}{\sum_{q=1}^{m_{ij\cdot}} w_{ijq}}\right)\Big/\sum_{q=1}^{m_{iju}} w_{ijq}\\ &=\frac{1}{w_{ij\cdot}}\sum_{q=1}^{m_{iju}} w_{ijq}\cdot X_{ijq}\cdot\frac{v_{ij\cdot}}{v_{ijqu}}\,.\end{aligned}$$

Thus

$$\mathrm{E}\,(\hat{X}_{ij\cdot u})=\mathrm{E}\left(\frac{1}{w_{ij\cdot}}\sum_{q=1}^{m_{iju}} w_{ijq}\cdot X_{ijq}\cdot\frac{v_{ij\cdot}}{v_{ijqu}}\right)\,.$$

We take this expected value in two steps. First, we compute the expected value conditional on m_{iju} and then we take the expectation over m_{iju}. The conditional expected value is

$$\begin{aligned}\mathrm{E}\,(\hat{X}_{ij\cdot u}\,|\,m_{iju})&=\frac{v_{ij\cdot}}{v_{ij\cdot u}}\cdot\frac{\sum_{q=1}^{m_{iju}} w_{ijq}}{w_{ij\cdot}}\cdot\mathrm{E}\left(X_{ijq}\cdot\frac{v_{ij\cdot u}}{v_{ijqu}}\right)\\ &=\frac{v_{ij\cdot}}{v_{ij\cdot u}}\cdot\frac{\sum_{q=1}^{m_{iju}} w_{ijq}}{w_{ij\cdot}}\cdot X_{ij\cdot u}\,.\end{aligned}$$

Now taking the expected values over m_{iju}, we obtain

$$\text{(21)} \qquad \mathrm{E}(\hat{X}_{ij\cdot u}) = \frac{v_{ij\cdot}}{v_{ij\cdot u}} \cdot \frac{v_{ij\cdot u}}{v_{ij\cdot}} \cdot X_{ij\cdot u} = X_{ij\cdot u} \ .$$

Similarly,

$$\hat{Y}_{ij\cdot u} = \frac{1}{w_{ij\cdot}} \sum_{q=1}^{m_{iju}} w_{ijq} \cdot X_{ijq} \cdot \hat{p}_{ijq} \frac{v_{ij\cdot}}{v_{ijqu}} \ ,$$

and the expected value is

$$\text{(22)} \qquad \mathrm{E}(\hat{Y}_{ij\cdot u}) = \sum_{q=1}^{M_{iju}} X_{ijq} p_{ijq} = Y_{ij\cdot u} \ .$$

Recall that

$$\hat{p}_{ij\cdot u} = \frac{\hat{Y}_{ij\cdot u}}{\hat{X}_{ij\cdot u}} \ .$$

Since equations (21) and (22) show that $\hat{X}_{ij\cdot u}$ and $\hat{Y}_{ij\cdot u}$ are unbiased estimates of $X_{ij\cdot u}$ and $Y_{ij\cdot u}$, respectively, it is reasonable to expect that, in most cases, $\hat{p}_{ij\cdot u} = \hat{Y}_{ij\cdot u} / \hat{X}_{ij\cdot u}$ is an approximately unbiased estimator of $p_{ij\cdot u} = Y_{ij\cdot u} / X_{ij\cdot u}$.

5. Summary and generalizations

In panel studies, the sampled units are used for several surveys over a period of years. The problem of changing strata and the resultant changes in selection probabilities after the initial selection is not solved by a single brand new selection of sampled units. Here we have discussed a method of estimation based on the original selection of the sample but which incorporates the subsequent changes in the value of one of the stratification factors. This method does not require (i) information on changes in the nonsampled units, nor (ii) replacement of some units in the original sample by new units.

The estimation technique described above assumes that the modernization process does not result in the rearrangement of sampling units (i.e., a sampling unit stays intact, only its type is changed). One generalization would be to the case of more complex changes (e.g., three units rearranged into two units).

Another generalization would be to develop a scheme for periodic updates so that only information on changes since the last update would be required. The current estimation technique requires the knowledge of all changes in sampled units since the original selection.

Acknowledgement

I would like to thank Professor Tore Dalenius of Brown University for several helpful discussions. I would also like to thank the referee for helpful comments.

BELL LABORATORIES

REFERENCES

[1] Cochran, W. G. (1977). *Sampling Techniques*, third edition, John Wiley & Sons.

[2] Fellegi, I. P. (1966). Changing the probabilities of selection when two units are selected with PPS without replacement, *Proceedings of the Social Statistics Section, Amer. Statist. Ass.*, Whashington, D.C., 434-442.

[3] Jain, A. K. (1978). Estimation from a stratified random sample under changes in strata composition, *Proceedings of the Section on Survey Research Methods, Amer. Statist. Ass.*, Washington, D.C., 642-646.

[4] Keyfitz, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in the probabilities, *J. Amer. Statist. Ass.*, **46**, 105-109.

[5] Kish, L. and Hess, I. (1959). Some sampling techniques for continuing survey operations, *Proceedings of the Social Statistics Section, Amer. Statist. Ass.*, Washington, D.C., 139-143.

[6] Kish, L. and Scott, A. (1971). Retaining units after changing strata and probabilities, *J. Amer. Statist. Ass.*, **66**, 461-470.