ESTIMATION OF TRANSITION PROBABILITIES FROM SAMPLE SURVEY DATA AND USES

J.C. Koop and M.P. Singh
Dominion Bureau of Statistics
Ottawa

1. Introduction

Consider a universe where the units are grouped into a number of class intervals or strata according to the variate values of an important characteristic of each of these units. Hereafter the variate value of such a characteristic will be termed the stratification variable. Due to changes in the stratification variable each group of units assigned to a particular stratum (out of a total of L) at a point or epoch of time before a survey is now distributed in a particular manner over L+1 strata, hereafter to be termed transitional strata, at the time of the survey, the (L+1)th stratum consisting of units having one or more special characteristics, e.g. going out of business in the context of a farm economy. In this paper we shall define the special characteristic for those units found in each of the (L+1)th strata as one of having zero variate values for all variables of interest to the survey.

At a point or epoch of time before a survey let N_h be the number of units in the hth stratum (h = 1, 2, ..., L). This is known to the sample surveyor. At the time of the survey the stratification variable of some of the units of the hth stratum changes, and N_{hj} out of N_h units are found in the j^{th} transitional stratum (j = 1, 2, ..., L+1); the sample surveyor has evidence of this change only from his sample. Let x_{hjk} be a variate value of interest of the k^{th} unit of the j^{th} transitional stratum of the h^{th} original stratum (h = 1, 2, ..., L; j = 1, 2, ..., L+1; k = 1, 2, ..., N_{hj}). The problem is to estimate

$$T = \sum_{h=1}^{L} \sum_{j=1}^{L+1} \sum_{k=1}^{N_{hj}} x_{hjk}, \qquad (1)$$

on the basis of a stratified sample of n_1, n_2, ..., n_L units drawn by simple random sampling without replacement from each respective original stratum.

$$\text{If } \bar{\bar{x}}_{hj} = \sum_{k=1}^{N_{hj}} x_{hjk}/N_{hj}, \text{ for all } (h,j),$$

is the mean of the j^{th} transitional stratum of the h^{th} original stratum, then (1) may be rewritten as

$$T = \sum_{h=1}^{L} \sum_{j=1}^{L} N_{hj} \bar{\bar{x}}_{hj}, \qquad (1.1)$$

since $\bar{\bar{x}}_{hj} = 0$ for h = 1, 2, ..., L and j = L+1. Further with respect to the original stratification if $\bar{\bar{x}}_h = \sum_{j=1}^{L} N_{hj} \bar{\bar{x}}_{hj}/N_h$, then (1.1) may also be rewritten as

$$T = \sum_{h=1}^{L} N_h \bar{\bar{x}}_h. \qquad (1.2)$$

2. Estimation Problems

The stratified sample shows that n_{hj} out of n_h units, drawn from the h^{th} original stratum, are found in the j^{th} transitional stratum in regard to which

$$0 \le n_{hj} \le n_h, \text{ for all } (h,j). \qquad (2)$$

The universe total T may be estimated
(i) by standard theory, effectively ignoring the information provided by (2) regarding the transition of units from stratum to stratum, or
(ii) by taking into account these transitions.

Estimation by standard theory. In this case the usual estimator of T is

$$\hat{T} = \sum_{h=1}^{L} N_h \bar{x}_h, \qquad (3)$$

where $\bar{x}_h$ is the mean of the sample units in stratum h. Alternatively in (1.1) if unbiased estimates of N_{hj} and $\bar{\bar{x}}_{hj}$ are substituted, viz.

$$\hat{N}_{hj} = \frac{n_{hj}}{n_h} N_h, \qquad (4)$$

and

$$\bar{\bar{x}}_{hj} = \sum_{k=1}^{n_{hj}} x_{hjk}/n_{hj}, \qquad (5)$$

for all relevant (h,j), then it is not difficult to see that the resulting estimator does not involve the n_{hj}'s and is identical to (3). The variance of this estimator is

$$V(\hat{T}) = \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right), \qquad (6)$$

where

$$S_h^2 = \sum_{j=1}^{L+1} \sum_{k=1}^{N_{hj}} (x_{hjk} - \bar{\bar{x}}_h)^2/(N_h - 1) \quad \text{for all h.}$$

Estimation taking into account transition probabilities. The probability of transition of a unit in stratum h to the j^{th} transitional stratum is $p_{hj} = N_{hj}/N_h$, and on the basis of sample data an unbiased estimate of this transition probability is $\hat{p}_{hj} = n_{hj}/n_h$ for all (h,j). Thus we have a matrix, with L+1 rows and L columns, of the estimated transition probabilities

$$(\hat{p}_{hj}). \tag{7}$$

This matrix multiplied by the column vector of the number of units in the L strata at the point of epoch of time before the survey, i.e., $\{N_1, N_2, \ldots, N_L\}$, yields a vector of L+1 unbiased estimates of the number of units in the L+1 strata at the time of the survey, the number in the j^{th} stratum being

$$\hat{N}_{.j} = \sum_{h=1}^{L} \hat{p}_{hj} N_h, \tag{8}$$

which is simply the sum over h of the estimate given by (4). By classical theory the variance of $\hat{N}_{.j}$, for all j, is given by

$$V(\hat{N}_{.j}) = \sum_{h=1}^{L} N_h^2 \frac{p_{hj}(1-p_{hj})}{n_h} \left(\frac{N_h-n_h}{N_h-1}\right). \tag{9.1}$$

Also

$$\mathrm{Cov}(\hat{N}_{.j}, \hat{N}_{.j'}) = -\sum_h N_h^2 p_{hj}p_{hj'} \left(\frac{N_h-n_h}{N_h-1}\right)/n_h \quad \text{for all } j\neq j'. \tag{9.2}$$

On the basis of the estimated stratum numbers given by (8), a biased but consistent estimator of T is

$$T' = \sum_{j=1}^{L} \hat{N}_{.j} \left(\sum_{h=1}^{L} n_{hj} \bar{x}_{hj} \Big/ \sum_{h=1}^{L} n_{hj}\right). \tag{10}$$

The loss in efficiency of the original stratification is to some extent regained by restratification through pooling each set of L transitional strata relating to the same range of x-values; this is the justification for constructing an alternative estimator T' given by (10).

We shall derive briefly an expression for the bias of T'. Substituting the expression for $\hat{N}_{.j}$ in (10), and denoting the random variables specified in (2) by $\underset{\sim}{n}_{hj}$ we find

$$E(T') = E\{E(T'|\underset{\sim}{n}_{hj})\}$$

$$= E\left\{\sum_{j=1}^{L}\left(\sum_{h=1}^{L}\frac{N_h}{n_h}n_{hj}\right)\left(\frac{\sum_{h=1}^{L} n_{hj}\bar{\bar{x}}_{hj}}{\sum_{h=1}^{L} n_{hj}}\right)\right\}$$

$$= \sum_{j=1}^{L}\sum_{h=1}^{L}\frac{N_h\bar{\bar{x}}_{hj}}{n_h}E\left(\frac{n_{hj}^2}{\sum_{h=1}^{L} n_{hj}}\right)$$

$$+ \sum_{j=1}^{L}\sum_{h\neq h'}\frac{N_h\bar{\bar{x}}_{h'j}}{n_h}E\left(\frac{n_{hj}n_{h'j}}{\sum_{h=1}^{L} n_{hj}}\right). \tag{11}$$

In order to evaluate the expectations of ratios of functions of random variables found in (11) we need the following result due to one of us. If X and Y are pairs of random variables with X assuming all values but not zero, then

$$E\left(\frac{Y}{X}\right) = \frac{E(Y)}{E(X)} - \frac{\mathrm{Cov}(Y,X)}{E^2(X)} + \frac{1}{E^2(X)}E\left\{\frac{Y}{X}(X-E(X))^2\right\}. \tag{12}$$

With this result, we find

$$E\left(\frac{n_{hj}^2}{\sum_h n_{hj}}\right) = \frac{E(n_{hj}^2)}{E(\sum_h n_{hj})} - \frac{\mathrm{Cov}(n_{hj}^2, \sum_h n_{hj})}{E^2(\sum_h n_{hj})} + \frac{1}{E^2(\sum_h n_{hj})}E\left\{\frac{n_{hj}^2}{\sum_h n_{hj}}\left(\sum_h (n_{hj}-E(n_{hj}))^2\right)\right\}, \tag{13}$$

and

$$E\left(\frac{n_{hj}n_{h'j}}{\sum_h n_{hj}}\right) = \frac{E(n_{hj}n_{h'j})}{E(\sum_h n_{hj})} - \frac{\mathrm{Cov}(n_{hj}n_{h'j}, \sum_h n_{hj})}{E^2(\sum_h n_{hj})} + \frac{1}{E^2(\sum_h n_{hj})}E\left\{\frac{n_{hj}n_{h'j}}{\sum_h n_{hj}}\left(\sum_h (n_{hj}-E(n_{hj}))^2\right)\right\}. \tag{14}$$

Further details of the derivation of E(T') are found in the appendix. Considering only the leading terms of (13) and (14) we find

$$E(T') = T + \frac{1}{n}\left[\sum_{j=1}^{L}\left\{\sum_{h>h'} N_hN_{h'}p_{hj}p_{h'j}(\bar{\bar{x}}_{hj}-\bar{\bar{x}}_{h'j})\left(\frac{n_h}{N_h}-\frac{n_{h'}}{N_{h'}}\right)\Big/\left(\sum_{h=1}^{L}\alpha_h p_{hj}\right)\right.\right.$$

$$\left.\left. + \sum_{h=1}^{L} N_{hj}\bar{\bar{x}}_{hj}(1-p_{hj})\left(\frac{N_h-n_h}{N_h-1}\right)\Big/\left(\sum_{h=1}^{L}\alpha_h p_{hj}\right)\right\}\right], \tag{15}$$

where $\alpha_h = n_h/n$ is the proportion of the sample allocated to stratum h. Now $\bar{\bar{x}}_{hj}$ is the mean of all the N_{hj} units of the h^{th} original stratum which have moved to the j^{th} transitional stratum. Similarly $\bar{\bar{x}}_{h'j}$ is the mean of all the $N_{h'j}$ units of the h'^{th} original stratum which have moved to the j^{th} transitional stratum. Since by this procedure of restratification the range of x-values in each of the j^{th} transitional strata is controlled, the differences $\bar{\bar{x}}_{hj}-\bar{\bar{x}}_{h'j}$, for all h>h',

could not be considerable. Further it is unlikely that the signs of such differences are the same unless the class interval of the stratification variable is wide. Indeed (15) suggests that the bias, $E(T')-T$, can be reduced by narrowing the width of the transitional strata so as to make the $\bar{\bar{x}}$'s for each j as equal as possible, in which situation this bias will be negligible when n is large.

The first steps in the derivation of the variance of T' are as follows:

$$V(T') = V\{E(T'|\underset{\sim}{n}_{hj})\} + E\,V(T'|\underset{\sim}{n}_{hj})\}$$

$$= V\{\sum_{j=1}^{L}\Big(\sum_{h=1}^{L}\frac{N_h}{n_h}n_{hj}\Big)\Big(\frac{\sum_{h=1}^{L}n_{hj}\bar{\bar{x}}_{hj}}{\sum_h n_{hj}}\Big)\}$$

$$+ E\{\sum_{j=1}^{L}\Big(\sum_{h=1}^{L}\frac{N_h}{n_h}n_{hj}\Big)^2\frac{1}{(\sum_h n_{hj})^2}\cdot$$

$$\sum_{h=1}^{L} n_{hj}^2\frac{S_{hj}^2}{n_{hj}}\Big(1-\frac{n_{hj}}{N_{hj}}\Big)\}, \qquad (16)$$

where $S_{hj}^2 = \sum_{k=1}^{N_{hj}}(x_{hjk}-\bar{\bar{x}}_{hj})^2/(N_{hj}-1)$ for all (h,j). The first part of the variance function $V\{\ \}$ may be ascribed to movement of some of the units of the original strata to their respective transitional strata resulting in the $\bar{\bar{x}}_{hj}$'s. This variance function involves the products of the estimates of the transition probabilities n_{hj}/n_h and the corresponding transitional stratum weights $n_{hj}/\sum_h n_{hj}$, and also involves the relevant cross-product terms. It can be shown that the variances and covariances of all these product and cross-product terms are of order $1/n_h^2$ for all h so that there is control on this part of the variance through restratification. The second part of the variance function is due to the variation of the x's in the transitional strata. We again remark that this source of variation can be controlled by narrowing the width of the transitional strata.

The variance of the unbiased estimator given by (6) can be rewritten as

$$V(\hat{\bar{T}}) = \sum_{h=1}^{L} N_h^2\Big(1-\frac{n_h}{N_h}\Big)\frac{1}{n_h}\sum_{j=1}^{L}\Big(\frac{N_{hj}}{N_h-1}\Big)(\bar{\bar{x}}_{hj}-\bar{\bar{x}}_h)^2$$

$$+ \sum_{h=1}^{L}\frac{N_h^2}{n_h}\Big(1-\frac{n_h}{N_h}\Big)\Big(\frac{N_{h,L+1}}{N_h-1}\Big)\bar{\bar{x}}_h^2$$

$$+ \sum_{h=1}^{L} N_h^2\Big(1-\frac{n_h}{N_h}\Big)\frac{1}{n_h}\sum_{j=1}^{L}\Big(\frac{N_{hj}-1}{N_h}\Big)S_{hj}^2. \qquad (17)$$

The expression speaks for itself. There are reasons to believe that $V(\hat{\bar{T}})$ can be greater than $V(T')$, because of the middle term in (17).

The variance expression for T' is developed to this point just to gain an insight into the features commented upon in the foregoing account. An exact expression for it will involve all the transition probabilities, and is not useful for estimation purposes because of its complexity.

The problem of variance estimation can be resolved by redesigning the survey so as to have two or more independent stratified samples, each yielding an independent estimate of T. When the total sample consists of two independent replicates, then the variance of the mean $\bar{T}=(T_1'+T_2')/2$ is given by

$$v(\bar{T}) = (T_1'-T_2')^2/4, \qquad (18)$$

despite the complexity of (16).

3. Remarks

The problem has been considered in the context of one-stage stratified sampling partly because it is simple, the strata corresponding to the "states" in the theory of Markov chains, and partly because the problem was originally seen in relation to this sampling design. The problem of estimation can be considered almost along the same lines for more ramified sampling designs.

Acknowledgement. We are grateful to Dr. I.P. Fellegi for some critical remarks which have improved the presentation of this paper. To our colleagues in the Agriculture Division of this Bureau, we are also grateful for many enlightening discussions on the practical problems of estimation in livestock and poultry sample surveys, which have provided the impulse for the estimation problem considered in this paper.

Reference

Koop, J.C. On the derivation of expected value and variance of ratios without the use of infinite series expansions. To appear in Metrika.

Appendix

Derivation of E(T'). For evaluation of terms involved in (13) and (14) we note the following.

$$E(n_{hj}) = n_h p_{hj},\quad E(n_{hj}^2) = n_h^2 p_{hj}^2 + n_h p_{hj} q_{hj} f_h,$$

$$E(n_{hj}n_{h'j}) = n_h n_{h'} p_{hj} p_{h'j},$$

where $q_{hj}=1-p_{hj}$, $f_h = \Big(1-\frac{n_h-1}{N_h-1}\Big)$,

$$\mathrm{Cov}\Big(n_{hj}^2, \sum_{h=1}^{L} n_{hj}\Big) = E(n_{hj}^3) - E(n_{hj}^2)\,E(n_{hj})$$

$$\doteq n_h^2 p_{hj}^2\,(3-q_{hj}f_j) - n_h p_{hj}\,(2-3q_{hj}f_h),$$

and

$$\mathrm{Cov}(n_{hj}n_{h'j}, \sum_{h=1}^{L} n_{hj}) = \mathrm{Cov}(n_{hj}n_{h'j}, n_{hj}) + \mathrm{Cov}(n_{hj}n_{h'j}, n_{h'j})$$

$$= E(n_{h'j})\, V(n_{hj}) + E(n_{hj})\, V(n_{h'j})$$

$$= n_h n_{h'} p_{hj} p_{h'j} (f_h q_{hj} + f_{h'} q_{h'j}).$$

Further denoting $E\{\frac{n_{hj}^2}{\sum_h n_{hj}} (\sum_{h=1}^{L} (n_{hj} - n_h p_{hj}))^2\}$ by $R_{hj}^{(1)}$ and $E\{\frac{n_{hj}n_{h'j}}{\sum_h n_{hj}} (\sum_{h=1}^{L} (n_{hj} - n_h p_{hj}))^2\}$ by $R_{hj}^{(2)}$, we get from (11), (13) and (14)

$$E(T') = \frac{1}{n} \sum_{j=1}^{L} \{ \sum_{h=1}^{L} N_h \bar{\bar{x}}_{hj} \frac{n_h p_{hj}^2}{\sum_h \alpha_h p_{hj}} + \sum_{h \neq h'} N_h \bar{\bar{x}}_{h'j} \frac{n_{h'} p_{h'j} p_{hj}}{\sum_h \alpha_h p_{hj}} \}$$

$$+ \frac{1}{n} \sum_{j=1}^{L} \sum_{h=1}^{L} N_h \bar{\bar{x}}_{hj} \frac{p_{hj} q_{hj} f_h}{\sum_h \alpha_h p_{hj}}$$

$$- \frac{1}{n^2} \sum_{j=1}^{L} \sum_{h=1}^{L} \frac{N_h \bar{\bar{x}}_{hj}}{(\sum_h \alpha_h p_{hj})^2} \{ n_h p_{hj}^2 (3 - q_{hj} f_h) - p_{hj}(2 - 3 q_{hj} f_j) - \frac{R_{hj}^{(1)}}{n_h} \}$$

$$- \frac{1}{n^2} \sum_{j=1}^{L} \sum_{h \neq h'} \frac{N_h \bar{\bar{x}}_{h'j}}{(\sum_h \alpha_h p_{hj})^2} \cdot$$

$$\{ n_h p_{h'j} p_{hj} (f_h q_{hj} + f_{h'} q_{h'j}) - \frac{R_{hj}^{(2)}}{n_h} \},$$

where $\alpha_h = n_h/n$ is the proportion of sample allocated to stratum h. Ignoring the contributions of third and fourth term for large sample size and adding and subtracting

$$\sum_{h \neq h'}^{L} (N_h \bar{\bar{x}}_{hj}\, p_{hj})(n_{h'} p_{h'j}) / \sum_{h=1}^{L} n_h p_{hj}$$

to the first term we can write

$$E(T') = \sum_{j=1}^{L} \sum_{h=1}^{L} N_{hj} \bar{\bar{x}}_{hj}$$

$$+ \frac{1}{n} \sum_{j=1}^{L} \{ \sum_{h=h'}^{L} N_{hj} (\bar{\bar{x}}_{h'j} - \bar{\bar{x}}_{hj}) \frac{n_{h'} p_{h'j}}{\sum_{h=1}^{L} \alpha_h p_{hj}}$$

$$+ \sum_{h=1}^{L} N_{hj} \bar{\bar{x}}_{hj} \frac{q_{hj} f_j}{\sum_{h=1}^{L} \alpha_h p_{hj}} \},$$

which can also be expressed in the form given in (15).