UNBIASED ESTIMATES FOR STRATIFIED SUBSAMPLE DESIGN

Robert S. Cochran, Univ. Wyo. - Texas A & M Univ.

Harold F. Huddleston, U. S. D. A.

The sampling design to be discussed in this paper is an unusual but apparently effective way of combining several different ways of sampling and constructing estimates. The discussion will center around the design, estimation procedures, and some empirical results.

A. The Sampling Plan

The first step in the sampling operation is the selection of a single stage cluster sample. In application, the clusters are actually segments in the usual USDA sense. Each cluster, or segment, is divided into tracts. The tracts found in the sample of segments are then classified into various strata depending upon some characteristic of the tract.

The second stage of the sampling procedure is to draw a stratified sample of tracts from the group of tracts found in the segments sampled in the first stage. If the original sample of tracts had been drawn at random from the entire population, stratified, and then a stratified sample of tracts drawn at the second stage, the design would be classical double sampling for the stratification problem. If the tracts selected at the second stage had been selected at random from within the segments, the design would be a classical two-stage cluster sample. The actual plan is a compromise between these two operations.

B. The Estimation Formulas

Notationally, a sample of m segments is drawn at random from a population of M segments. In the sample of m segments there are found n_h tracts belonging to the h^{th} stratum. The second part of the sample calls for v_h tracts to be selected at random from the h^{th} stratum. There will be found v_{hs} tracts at the intersection of the h^{th} stratum and the s^{th} segment. Associated with each unit found in this intersection will be a characteristic, a_{hst}. These characteristics may be combined to give the following population and sample quantities:

$$a_{hs} = \sum_{t}^{v_{hs}} a_{hst} \qquad a = \sum_{h}^{H} a_h = \sum_{s}^{m} a_s$$

$$a_h = \sum_{s}^{m} a_{hs} \qquad A_s = \sum_{h}^{H} \sum_{t}^{n_h} a_{hst}$$

$$a_s = \sum_{h}^{H} a_{hs} \qquad A = \sum_{s}^{m} A_s$$

$$\hat{A}_1 = \frac{M}{m} \sum_{s}^{m} A_s = \frac{M}{m} A$$

$$\hat{A}_2 = \frac{M}{m} A = \frac{M}{m} \sum_{h}^{H} \hat{A}_h = \frac{M}{m} \sum_{s}^{m} \hat{A}_s$$

$\hat{A}_1$ is the estimate of the population total if there is a complete enumeration of all tracts in the sampled segments. $\hat{A}_2$ is based on the stratified sample of tracts.

Spelling out $\hat{A}_2$ in more detail we have

$$\hat{A}_2 = \frac{M}{m} \sum_{s}^{m} \hat{A}_s = \frac{M}{m} \sum_{s}^{m} \sum_{h}^{H} \frac{n_h}{v_h} \sum_{t}^{v_{hs}} a_{hst} \qquad (1)$$

In this form $\hat{A}_s$ is seen to be a "domain" estimate of the s^{th} segment total based upon the stratified sample that cuts across all segments.

When $\hat{A}_2$ is written as

$$\hat{A}_2 = \frac{M}{m} \hat{A} = \frac{M}{m} \sum_{h}^{H} \hat{A}_h$$

its unbiasedness is apparent since $\hat{A}_h$ is an unbiased estimate of A_h and thus $\hat{A}$ is an unbiased estimate of A.

The variance of $\hat{A}_2$ is easily obtained by applying the principle of conditional expectation. As such

$$V(\hat{A}_2) = V_1 E_2(\hat{A}_2) + E_1 V_2(\hat{A}_2) \qquad (2)$$

$$= \frac{M^2}{2} \left(1 - \frac{m}{M}\right) \sum_{s}^{M} \frac{(A_s - \frac{A}{M})^2}{M - 1} + \frac{M}{m}\left[\sum_{s}^{M} V(\hat{A}_s) + \frac{m-1}{M-1} \sum_{s \neq s'} \mathrm{Cov}(\hat{A}_s \hat{A}_{s'})\right]$$

If the sample of tracts had been drawn following traditional two-stage sampling methods, the term involving covariances would have been zero, but since the sample was drawn following strata boundaries, the covariance of the domain estimates must be present. Yates (1960) indicates that these covariances will be negative so this stratified scheme will be better than the traditional two-stage cluster operation.

The derivation of the unbiased estimate of this expression is obtained by splitting it into its components, estimating each of them, and then recombining the component estimates.

For the "between" cluster component,

$$\frac{M^2}{m} \left(1 - \frac{m}{M}\right) \frac{\sum_{s}^{M} (A_s - \frac{A}{M})^2}{M - 1},$$

the estimate is easily shown to be

$$\frac{M^2}{m}\left(1 - \frac{m}{M}\right) \left[\frac{\frac{M}{m} \sum_{s}^{m} \left(\hat{A}_s - \frac{\sum_{s}^{m} \hat{A}_s}{m}\right)^2 - \frac{M}{m} \sum_{s}^{m} v(\hat{A}_s) + \frac{v(\hat{A}_2)}{M}}{M - 1}\right] \qquad (3)$$

For the "within" term,

$$E_1 V_2(\hat{A}_2) = E_1\{\frac{M^2}{m^2} V_2(\sum_s^m \hat{A}_s)\}$$

$$= \frac{M}{m}\left[\sum_s^M V(\hat{A}_s) + \frac{m-1}{M-1} \sum_{s\neq s'} Cov(\hat{A}_s \hat{A}_{s'})\right]$$

the estimate is simply

$$\frac{M^2}{m^2} \nu(\sum_s^m \hat{A}_s). \qquad (4)$$

Thus

$$\nu(\hat{A}_2) = \frac{M^2}{m}\left(1 - \frac{m}{M}\right) \frac{\sum_s^m \left[\hat{A}_s - \frac{\sum_s^m \hat{A}_s}{m}\right]^2}{m-1}$$

$$+ \frac{M}{m} \sum_s^m \nu(\hat{A}_s) + \frac{M(M-1)}{m(m-1)}\left[\nu(\sum_s^m \hat{A}_s) - \sum_s^m \nu(\hat{A}_s)\right] \qquad (5)$$

Since

$$\sum_s^m \hat{A}_s = \hat{A} = \sum_h^H \frac{n_h}{v_h} \sum_s^m \sum_t^{v_{hs}} a_{hst}$$

is a stratified estimate of A, $(\sum_s^m \hat{A}_s)$ can be computed by

$$\nu(\sum_s^m \hat{A}_s) = \sum_h^H \frac{n_h^2}{v_h}\left(1 - \frac{v_h}{n_h}\right)$$

$$\cdot \frac{\sum_s^m \sum_t^{v_{hs}} \left[a_{hst} - \frac{\sum_s^m \sum_t^{v_{hs}} a_{hst}}{v_h}\right]^2}{v_h - 1} \qquad (6)$$

Since $\hat{A}_s$ is a "domain" type estimate for the s^{th} segment when the stratified sample of tracts cuts across all segments, the quantity $\nu(\hat{A}_s)$ can be written as

$$\nu(\hat{A}_s) = \sum_h^H \frac{n_h^2}{v_h}\left(1 - \frac{v_h}{n_h}\right)$$

$$\cdot \frac{\sum_t^{v_{hs}} \left[a_{hst} - \frac{\sum_t^{v_{hs}} a_{hst}}{v_h}\right]^2}{v_h - 1}$$

When the number of tracts in the intersection of the h^{th} stratum and the s^{th} segment, n_{hs}, is known, these formulas can be altered to take this into account.

Some additional algebra can be used to give a more compact form for $\nu(\hat{A}_2)$ as

$$\nu(\hat{A}_2) = \frac{M^2}{m}\left(1 - \frac{m}{M}\right) \frac{\sum_s^m \left[\hat{A}_s - \frac{\sum_s^m \hat{A}_s}{m}\right]^2}{m-1}$$

$$+ \frac{M^2}{m^2}\left[\sum_h^H \frac{n_h^2}{v_h}\left(1 - \frac{v_h}{n_h}\right) \frac{(-\sum_{s\neq s'} a_{hs} a_{hs'})}{v_h(v_h - 1)}\right]^2$$

$$+ \frac{M}{m} \sum_s^m \nu(\hat{A}_s) .$$

C. Some Empirical Results

An area sample survey was conducted in Ohio in June and December, 1969, using this sampling procedure. The purpose of the survey was to estimate the inventory of major livestock characteristics, spring planted crops, farm labor and operator characteristics, and the harvest of fall planted crops. The area sample survey data was collected on a subunit or tract basis for each of the segments. A tract being the land under the control of an operator. The tracts were then regrouped into nine strata based on previously established criteria for a survey which would have different emphasis then the initial June Survey. However, a number of survey characteristics were common to both surveys. The data for six of these characteristics are given in the following tables. In table 1, column 1, are the estimates referred to as $\hat{A}_1$. These are the usual single stage cluster sample estimates for the data obtained in June. In column 2 are the estimates of the June totals based upon the stratified subsample and computed using the $\hat{A}_2$ formula. In column 3 are the estimates of the December totals based upon the stratified subsample and also computed using the $\hat{A}_2$ formula. The December data has been obtained on the same units as the column 2 June estimates.

In table 2 the estimates of the variances of the estimates in table 1, column 2 and column 3 are presented in total as well as broken into three components. The first component is computed between estimates of segment totals. The second component is the usual within segment component. The third component reflects the covariance of estimates of segment totals. The relative size of these components follows the usual pattern of most of the variance estimates being obtained from the between component.

Table 1

Estimates of Characteristic Totals

Characteristic	Based on Complete June Survey	Based on Sample from June Survey	Based on December Data - June Sample
	(000)	(000)	(000)
1	2,530	3,141	3,290
2	2,252	2,891	3,236
3	1,017	1,703	1,232
4	11,847	6,824	7,411
5	1,012	1,288	1,304
6	45	46	95
M	107,858		
m	350		
Σn_h	2,377		
Σv_h		487	487

Table 2

Estimates of Variance Components

	Between		Within		Covariance		Total	
Characteristic	June	December	June	December	June	December	June	December
	(000,000)	(000,000)	(000,000)	(000,000)	(000,000)	(000,000)	(000,000)	(000,000)
1	245,123	272,023	175	240	-2,282	-1,921	243,016	270,342
2	105,499	141,990	79	114	-2,186	-2,092	103,392	140,012
3	193,538	86,396	389	174	- 482	- 506	193,445	86,064
4	17,077,493	13,526,947	33,185	24,748	-2,659	-4,367	17,108,019	13,547,328
5	13,414	17,577	20	31	- 429	- 491	13,005	17,117
6	655	1,298	1	2	0	- 4	656	1,296

D. Reference

Yates, F. (1960). Sampling methods for censuses and surveys. Charles Griffin and Co., London, third edition.