ESTIMATION OF A POPULATION MEAN WITH TWO-WAY STRATIFICATION USING A SYSTEMATIC ALLOCATION SCHEME

Michael R. Chernick, Aerospace Corporation
Tommy Wright, Union Carbide Corporation

A technique of systematically allocating a sample to the strata formed by double stratification is presented. The method can proportionally allocate the sample along each variable of stratification. If there are R strata and C strata for the first and second variable of stratification respectively, the technique requires that the total sample size be at least as large as max(R,C). An unbiased estimator of the population mean is given and its variance is obtained. The technique is compared with a random allocation procedure given by Bryant, Hartley and Jessen (1960). Numerical examples are given suggesting when one technique is superior to the other.

KEY WORDS: Stratified sampling, Systematic allocation, Random allocation, Two-way stratification.

1. Introduction

Bryant, Hartley, and Jessen (1960) have developed a method for applying two-way stratification when the total number of observations is required to be less than the number of strata which would be formed by the usual double stratification. They have shown that in many situations the procedure can improve the precision of the estimates over single stratification with either variable chosen for stratification.

The authors have found that in data validation respondent surveys many variables are candidates for stratification and since the relationship between these variables and the response is not well understood, two-way or multi-way stratification with proportional allocation along each variable seems appropriate. The importance of data validation to energy data is discussed by Moses (1978).

Bryant, Hartley, and Jessen (1960) give a simple method for proportionally allocating the sample along two criteria when the total number of observations is at least the maximum of R and C, where R is the number of strata for the first variable and C is the number for the second. An extension to multi-way stratification is given by Raghunandanan and Bryant (1971).

Bryant et al. (1960) estimate the population mean $\bar{Y}$ using two-way stratification. We use their notation which is now given. The population of size N is stratified along two variables in R rows and C columns. The RC sections formed are called cells. The ij^{th} cell refers to the collection of units belonging to the i^{th} category for the first variable and j^{th} category for the second variable where i=1, 2, ..., R and j=1, 2, ..., C.

N_{ij} is the number of units in the ij^{th} cell,

Y_{ijk} is the k^{th} unit in the ij^{th} cell for k=1, 2., ..., N_{ij},

$P_{ij} = N_{ij}/N$,

$\bar{Y}_{ij}$ is the mean of the units in the ij^{th} cell,

$$S_{ij}^2 = \sum_{k=1}^{N_{ij}} (Y_{ijk}-\bar{Y}_{ij})^2/(N_{ij}-1),$$

$$P_{i.} = \sum_{j=1}^{C} P_{ij},\ P_{.j} = \sum_{j=1}^{R} P_{ij},$$

$$\bar{Y} = \sum_{i=1}^{R}\sum_{j=1}^{C} P_{ij}\bar{Y}_{ij}.$$

n_{ij} is the number of units in the sample belonging to the ij^{th} cell,

$\bar{y}_{ij}$ is the sample mean for the ij^{th} cell,

$$n_{i.} = \sum_{j=1}^{R} n_{ij}, \text{ and } n_{.j} = \sum_{i=1}^{R} n_{ij}.$$

We shall assume independence between the variables of stratification, i.e. $P_{ij} = P_{i.}P_{.j}$ for simplicity. The results can be generalized for the dependent case. We must have $n \geq \max(R,C)$. For proportional allocation the n sample units are assigned to satisfy the following equations as closely as possible.

$$n_{i.} = nP_{i.} \text{ and } n_{.j} = nP_{.j}$$

We fix $n_{i.}$ and $n_{.j}$ for each i and j, while n_{ij} is a random variable. After n, $n_{i.}$, and $n_{.j}$ have been determined, a square is constructed having n lines (s=1, 2, ..., n) and n arrays (t=1, 2, ..., n) forming n^2 squares. For i=1, ..., R and j=1, ..., C we combine $n_{i.}$ adjacent lines to form the i^{th} row and $n_{.j}$ adjacent arrays to form the j^{th} column.

The procedure of Bryant et al. (1960) shall be referred to as random allocation. In the first line, one array is selected at random and marked. In the second line, for the remaining n-1 arrays, another array is selected at random and marked. This procedure is continued until an array has been selected for each line. The size of the sample chosen from a given cell is the number of marks within that cell.

When $P_{ij} = P_{i.}P_{.j}$ for each i and j the estimator $\bar{y} = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}\bar{y}_{ij}/n$ was given. The estimator is unbiased and its variance is given in Bryant et al. (1960). Two properties of the random allocation method are: i) every square has probability 1/n of being marked; ii) the $n_{i.}$'s and $n_{.j}$'s satisfy the conditions for proportional allocation and $E(n_{ij}) = nP_{i.}P_{.j} = nP_{ij}$.

2. Systematic Allocation

Here, we propose an alternative to random allocation which has properties i) and ii) of the previous section.

We use the n×n square matrix described in Section 1. The squares are numbered moving from left to right from 1 to n^2 starting with the first line. On the first line a square is selected with probability 1/n and is marked. If the g^{th} square is chosen on the first line, then the allocation is systematically determined by marking those squares which have the numbers

$$g + \ell(n+1) \text{ for } \ell=0, 1, 2, \ldots, n-g \text{ and}$$
$$n(n-g+k) + k \text{ for } k=1, 2., \ldots, g-1 \text{ for} \quad (2.1)$$
$$g=1, 2, \ldots, n.$$

There are n possible allocations depending only on the initial choice of g. When n=10, we see that there are 10 possible allocations as contrasted to random allocation which would have 10!=3,628,800 possible allocations. Also, random allocation requires the generation of n-1 random numbers whereas systematic allocation requires just 1 random number for any sample size n.

After the sample has been allocated, a simple random sample of size n_{ij} is chosen from the ij^{th} cell for each pair (i,j). The random variable n_{ij} is determined by the allocation and can take on values between 0 and $\min(n_{i.},n_{.j})$.

The systematic allocation scheme partitions the n^2 squares into n sets of equal probability of selection. Consequently property j) of Section 1 is satisfied. Since the ij^{th} cell contains $n_{i.}n_{.j}$ unit squares

$$E_{CW}(n_{ij}) = \frac{n_{i.}n_{.j}}{n} \quad (2.2)$$

(We use the subscripts CW and BHJ to denote systematic allocation and random allocation, respectively.)

The estimator for $\bar{Y}$ under systematic allocation is

$$\bar{y} = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}\bar{y}_{ij}/n. \quad (2.3)$$

This is the same formula as for random allocation. When $n_{ij}=0$, we may define $\bar{y}_{ij}=0$. It can be shown that $E_{CW}(\bar{y}) = \bar{Y}$.

So $\bar{y}$ is unbiased under systematic allocation when $P_{i.}P_{.j} = P_{ij}$.

We now introduce some notation for use in the formulae for $Var_{CW}(\bar{y})$.

Let O_{ℓ}^{ij} = the number of ordered pairs of distinct squares in the ij^{th} cell associated with the ℓ^{th} systematic allocation for $\ell=1, 2, \ldots, n$;

$$\delta_{ij} = \sum_{\ell=1}^{n} O_{\ell}^{ij};$$

u_{ℓ}^{ij} = the number of squares in the ij^{th} cell belonging to the ℓ^{th} systematic allocation;

$$\delta_{ijij'} = \sum_{\ell=1}^{n} u_{\ell}^{ij}u_{\ell}^{ij'}, \quad \delta_{iji'j} = \sum_{\ell=1}^{n} u_{\ell}^{ij}u_{\ell}^{i'j} \text{ and}$$

$$\delta_{iji'j'} = \sum_{\ell=1}^{n} u_{\ell}^{ij}u_{\ell}^{i'j'}$$

Under systematic allocation we have

$$Var_{CW}(\bar{y}) = 1/n^2\Big[\sum_{ij} S_{ij}^2 \frac{[(N_{ij}-1)\, n_{i.}n_{.j}-\delta_{ij}]}{nN_{ij}}$$
$$+ \sum_{ij} (n_{i.}n_{.j}(n-n_{i.}n_{.j}) + n\delta_{ij})\bar{Y}_{ij}^2/n^2$$
$$+ \sum_{ij}\sum_{ij'} (n\delta_{ijij'}-n_{i.}^2 n_{.j}n_{.j'})\bar{Y}_{ij}\bar{Y}_{ij'}/n^2$$
$$+ \sum_{ij}\sum_{i'j} (n\delta_{iji'j}-n_{.j}^2 n_{i.}n_{i'.})\bar{Y}_{ij}\bar{Y}_{i'j}/n^2$$
$$+ \sum_{ij}\sum_{i'j'} (n\delta_{iji'j'}-n_{i.}n_{i'.}n_{.j}n_{.j'})$$
$$\bar{Y}_{ij}\bar{Y}_{i'j'}/n^2\Big].$$

Theorem 1: Let R=C=2. If $\min(n_{11},n_{12},n_{21},n_{22})=1$, then the joint probability distribution of n_{11}, n_{12}, n_{21}, and n_{22} is the same for random and systematic allocations.

Since the estimators are the same function of the n_{ij}'s, $Var_{BHJ}(\bar{y})$ and $Var_{CW}(\bar{y})$ can differ only if the distributions of the n_{ij}'s differ and so under the assumptions of Theorem 1, $Var_{BHJ}(\bar{y}) = Var_{CW}(\bar{y})$.

3. Systematic Allocation When R=C=n

If R=C=n and we require that each stratum for the variables of stratification be represented in the sample, then $n_{i.} = n_{.j} = 1$ for every i and j. The variance formulae under random allocation simplifies to the following:

$$Var_{BHJ}(\bar{y}) = n^{-2}\Big[\sum_{ij} \frac{(N_{ij}-1)}{nN_{ij}} S_{ij}^2 + \sum_{ij}(n-1)\bar{Y}_{ij}^2/n^2$$
$$- \sum_{ij}\sum_{ij'} \bar{Y}_{ij}\bar{Y}_{ij'}/n^2 - \sum_{ij}\sum_{i'j} \bar{Y}_{ij}\bar{Y}_{i'j}/n^2$$
$$+ \sum_{ij}\sum_{i'j'} \frac{\bar{Y}_{ij}\bar{Y}_{i'j'}}{n^2(n-1)}\Big]. \quad (3.1)$$

While under systematic allocation $\delta_{ij} = \delta_{ijij'} = \delta_{iji'j} = 0$ and so

$$\mathrm{Var}_{CW}(\bar{y}) = n^{-2}\Big[\sum_{ij}\frac{(N_{ij}-1)}{nN_{ij}}S_{ij}^2 + \sum_{ij}(n-1)\bar{Y}_{ij}^2/n^2$$

$$- \sum_{ij}\sum_{ij'}\bar{Y}_{ij}\bar{Y}_{ij'}/n^2 - \sum_{ij}\sum_{i'j}\bar{Y}_{ij}\bar{Y}_{i'j}/n^2$$

$$+ \sum_{ij}\sum_{i'j'}(n\delta_{iji'j'}-1)\bar{Y}_{ij}\bar{Y}_{i'j'}/n^2\Big]. \qquad (3.2)$$

Note: $\delta_{iji'j'} = 1$ if the ij^{th} and $i'j'^{th}$ cells are in the same systematic allocation.

$= 0$ otherwise

We note that the formulae differ only in the last term. For comparisons we need only look at the last term.

Lemma 1: When R=C=n=2 the two procedures are equivalent.

Lemma 2: If $\bar{Y}_{ij}\bar{Y}_{i'j'} = c$ for each pair (i,j) and (i',j') then $\mathrm{Var}_{BHJ}(\bar{y}) = \mathrm{Var}_{CW}(\bar{y})$

4. Concluding Remarks

In many instances systematic allocation produces the same joint distribution for the n_{ij}'s as random allocation. For these cases it is unnecessary to use random allocation, since the two estimators have identical sampling properties and the one obtained by systematic allocation requires much less effort. In some cases, the procedures are identical simply by virtue of the structure of the table and are not dependent on the population parameters $\bar{Y}_{ij}$ and S_{ij}^2 for i=1, ..., R and j=1, ..., C. Theorem 1 is one example of this type of result, and additional research in this area may be helpful.

Systematic allocation can be easily extended to higher order stratification. We believe that such methods will have advantages over the method of Raghunandanan and Bryant (1971) as it has for two-way stratification over random allocation.

In practical applications, we would recommend systematic allocation over random allocation because it is much easier to implement and unlikely (we believe) to produce estimators with significantly larger variances than random allocation. Indeed many examples exist where $\mathrm{Var}_{CW}(\bar{y}) < \mathrm{Var}_{BHJ}(\bar{y})$ (Chernick and Wright Technical Report). Future research will, hopefully, provide better guidelines for its use.

REFERENCES

Bryant, Edward, C., Hartley, H. O., and Jessen, R. J. (1960), "Design and Estimation in Two-Way Stratification," Journal of the American Statistical Association, 55, 105-124.

Chernick, M., and Wright, T. (1980), "On Multi-Way Stratification Problems," Union Carbide Nuclear Division report in preparation.

Moses, Lincoln E. (1978), "Energy Information Validation - A Status Report," Proceedings of the 1978 DOE Statistical Symposium, 33-49.

Raghunandanan, K., and Bryant, E. C. (1971), "Variance in Multi-Way Stratification," Sankhya, Series A, 33, 221-226.

*Research sponsored by the Energy Information, U.S. Department of Energy under contract W-7405-eng-26 with the Union Carbide Corporation.