Jour. Ind. Soc. Ag. Statistics
51(2&3), 1998 : 283-292

Some Aspects of Techniques for Unbiased Estimation in Sampling from a Finite Population

Om P. Aggarwal[1]

SUMMARY

The method of enumerating the different probabilities of sampling attached to the different unit-squares (Cells) of the field in crop-cutting surveys for estimation of average yield of crops and eliminating bias introduced by assuming that they all had the same probability is discussed in the paper.

Keywords: Probability sampling, Bias, Correction factor, Sampling units, Estimator.

1. Introduction

In crop-cutting surveys for estimation of the average yield of crops, initiated by the Indian Council of Agricultural Research in 1944 under direction of Sukhatme, the procedure given for selecting a plot for crop-cutting within a field growing the crop under survey can be briefly outlined as follows:

Suppose that the field is a rectangle with dimensions $L \times B$, and that dimension L lies along the east-west line. Suppose a rectangular plot of dimension $a (\leq L)$ and $b (\leq B)$ units is to be located at random within the field. Select a pair of random numbers, say (r, s), with the help of random number tables, such that $0 \leq r \leq L - a$ and $0 \leq s \leq B - b$. Without loss of generality, and for the sake of simplicity, we shall call the dimension L as length, and B as breadth. Now locate a plot in the field in such a way that its length a and breadth b are along the length and breadth of the field, and its south-west corner is at a distance of r units along the length, and s units along the breadth from the south-west corner of the field [4].

The above method of selection of a plot implies a division of the field into $(L - a + 1) \times (B - b + 1)$ plots of size $(a \times b)$ and then selecting one of these plots giving equal probability of selection, namely $1/[(L - a + 1)(B - b + 1)]$, to each of these plots. However, since the plots are not distinct, but overlapping, the method does not provide equal probability of selection to each of the unit squares of the field. It will be shown in the next section that the unit squares in the central portion of the field would have a relatively higher probability of selection as compared with the unit-squares near the border,

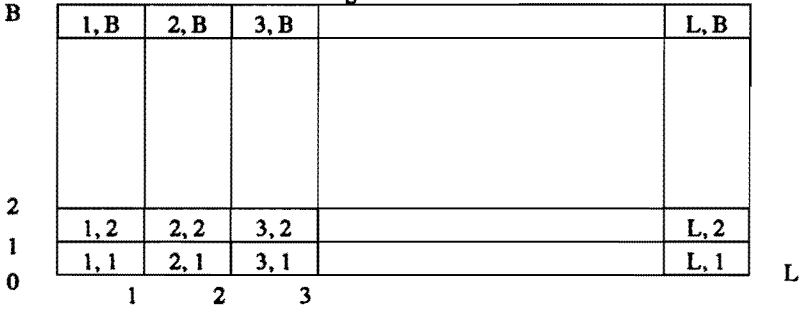
1 229 Pershing DR., Oakland, USA

because a unit-square around the central portion of the field is included in more plots than a unit square around the border. The method of enumerating the different probabilities of sampling attached to the different unit-squares of the field, hereafter called "cells", and "eliminating the bias" introduced by assuming that they all had the same probability is the purpose of this note.

2. *Determination of the Number of Plots in which a Given Cell is Included*

Let the field, assumed to be rectangular of dimensions L × B units, be divided into L × B cells by drawing parallel lines along the length and breadth of the field shown in the following diagram:

Diagram 1



Let us number the columns 0, 1, 2, ..., L to the right and the rows 0, 1, 2, ..., B upward in the above diagram. We shall call a cell (i, j) when it lies between the $(i-1)^{th}$ and the i^{th} column; i = 1, 2, ..., L and between the $(j-1)^{th}$ and the j^{th} rows; j = 1, 2, ..., B.

By following the procedure of selecting a plot outlined in the previous paragraph, the sample plot located in the field by the pair (r, s) of random numbers consists of a rectangular block of (a × b) cells lying between the r^{th} and the $(r + a)^{th}$ column. (r = 0, 1, 2, ... L-a) and between the s^{th} and the $(s + b)^{th}$ row, (s = 0, 1, 2, ... B-b). We shall say that this plot is determined by the pair (r, s).

It will be noted that if a > 1, the cell (1, 1) is includable in only one plot, namely the one determined by the pair (0, 0), while the cell (2, 1) is includable in 2 plots, namely those determined by the pairs (0, 0) and (l, 0). Similarly, if a > 1, b >1, the cell (2, 2) is includable in exactly 4 plots, namely those determined by the pairs (0, 0); (1, 0); (0,1) and (1, 1). In general, the number of plots in which a particular cell occurs is given by the following theorem in which m = min (a, L – a + 1) and n = min (b, L – b + 1). Obviously $m \geq 1, n \geq 1$.

Theorem 1. The cell (i, j) is includable in f_{ij} plots where

$$f_{ij} = g_i h_j, \text{ for all i and j} \tag{1}$$

with g_i and h_j assuming the following values:

$g_i = 1$ for all i, if $m = 1$; and if $m \geq 2$

$$g_i = \begin{cases} i & \text{for } 1 \leq i \leq m-1 \\ m & \text{for } m \leq i \leq L-m+1 \\ L-i+1 & \text{for } L-m+2 \leq i \leq L \end{cases} \tag{2}$$

and where h_j is obtained by replacing in (2) above m, g_i, i, a, and L by n, h_j, j, b and B respectively. For brevity, the term f_{ij} will be called the frequency of he cell (i, j).

Proof. For the sake of simplicity, at first consider only the first row of the L cells of the type (i, 1) given in Diagram 1. These cells can be renumbered as indicated in the following diagram by dropping the number 1.

Diagram 2

1	2	3	...	i	...	L
0	1	2	3			L

The problem is to determine the number of plots of size a × b in which a given cell can be included by following the selection procedure given in the paragraph 1. At first we assume, for simplicity, that $b = 1$.

We state here 2 lemmas that will be useful in the proof of Theorem 1. Their proof is straightforward and is not given here.

Let $m = \min(a, L - a + 1)$ where a and L are positive integers and $a \leq L$. Then:

Lemma 1. $m = 1$ if and only if $a = 1$ or $a = L$ (3)

Lemma 2. $m > 1$ if and only if $1 < a < L$ (4)

It is easily seen that in the two extreme cases where $a = 1$ (the smallest possible value), or $a = L$ (the largest possible value), the number of plots in which each cell of the diagram 2 can be included is only 1. Since Lemma 1 states that these cases are equivalent to $m = 1$, it proves that when $m = 1$, the frequency of each cell of the diagram 2 is 1. Furthermore, if the frequency of each cell of the diagram 2 is 1, it can only happen when $a = 1$ or $a = L$.

Let us now consider the frequency of the cells in the case where $1 < a < L$. As Lemma 2 states, this case is equivalent to the case $m > 1$.

We note that the cell 1 can be included in only one plot since it can be selected only when the selected random number, r, is equal to 0. The same is true of the last cell, L, since it can be selected only when $r = L - a$ (the largest possible value). However the cell 2 is includable in 2 plots, namely when $r = 0$ or $r = 1$. The same is true of the last but one cell, L - 1, since that cell is also includable in only two plots, namely when $r = L - a - 1$ or $r = L - a$.

Continuing with the same argument, it will be noted that:

(i) The frequency of the cells, starting with cell 1, and going from 1 to 2, 2 to 3, and so on, as well as starting with the last cell and counting back from L to L-1 to L-2 as so on, continues to increase by one (but see below):

(ii) If $a \leq L - a + 1$, the highest possible frequency of any cell is a. This frequency is reached by the a^{th} cell from the beginning and from the end, and it remains the same for all the cells in between.

(iii) However, if $a > L - a + 1$, the maximum frequency $(L - a + 1)$ is reached by the $(L - a + 1)^{th}$ cell from the beginning and from the end, and then it remains the same $(L - a + 1)$ for all the cells in between.

By consideration of (ii) and (iii) above, we note that the maximum frequency reached by the middle cell or cells is $m = \min (a, L - a + 1)$ and the frequency pattern of all the cells is seen to be:

1, 2, 3, ..., (m – 1), m, m, m, ..., m, (m – 1), (m – 2), ..., 3, 2, 1

The g_i is obviously the i^{th} term of this series, and is given by (2) of Theorem 1.

Obviously a similar sequence of the cell frequencies is obtained by considering the first column of the B cells of diagram 1. The j^{th} term of the sequence, say h_j, is obtained by replacing in (2) above the m, g_i, i, a, and L by n, h_j, j, b and B respectively.

Now if we consider the frequencies of the different cells of the diagram 1 when plots of size $a \times b$ are to be located, it will be seen at once, by following the same procedure as in the beginning of this proof, that the frequency for the cell (i, j) is given by:

$$f_{ij} = g_i h_j$$

where g_i is the frequency of the cell (i, 1), and h_j is the frequency of the cell (1, j). This completes the proof of Theorem 1.

3. *Sum of the Frequencies of All the Cells Comprising a Plot Located by a Particular Pair of Random Numbers*

If the pair of random numbers selected for locating the plot in the field is (r, s), $0 \le r \le L-a, 0 \le s \le B-b$, the plot located will comprise of the $a \times b$ cells in the rectangle starting from the cell $(r+1, s+1)$ through the cell number $(r+a, s+1)$ along the length, and from the cell $(r+1, s+1)$ through the cell number $(r+1, s+b)$ along the breadth. In an attempt to obtain a correction factor to be used as a weight to compensate for the unequal probabilities with which the different portions of the field were sampled by this method of locating the plot, expressions were obtained for the sum of the frequencies of the different cells comprising a plot located by a given pair of random numbers (r, s). It was argued that since the probabilities of the individual cells being sampled were not equal, the plot yield from a plot consisting of cells having relatively smaller probability (border plots) should be given an appropriately higher weight by multiplying its yield with a factor greater than one, and the yield from a plot consisting of cells having relatively larger probability (central plots) should be given a lower weight by multiplying its yield with a factor less than one. We shall first obtain expressions for the sum of the frequencies of the different cells comprising the plot (r, s) - meaning the plot located by the pair of random numbers (r, s) - in this section.

Let f(r,s) denote the sum of the frequencies of all the cells in the plot located with the random number pair (r, s). Then:

$$f(r,s) = \sum_{i,j} f_{ij} \tag{5}$$

where the summation extends over all values of the pair (i, j) with $(r+1) \le i \le (r+a)$ and $(s+1) \le j \le (s+b)$.

Since from (1), $f_{ij} = g_i\, h_j$, it is seen that:

$$f(r,s) = S(L,a,r).\, S(B,b,s) \tag{6}$$

where
$$S(L,a,r) = \sum_{i=r+1}^{r+a} g_i; \quad r = 0, 1, \ldots, L-a \tag{7}$$

and
$$S(B,b,s) = \sum_{j=s+1}^{s+b} h_j; \quad s = 0, 1, \ldots, B-b \tag{8}$$

We proceed first to obtain the summation in (7).

We note that:
$$g_i = g_{L-i+1} \text{ for all } i \tag{9}$$

Actually, we can divide up all the L frequencies into the following three blocks:

First block : 1, 2, 3, ..., m–1

Second or middle block : m, m, m, ..., m

Third block: m –1, m – 2, ..., 2, 1

The number of terms in the first and third blocks is (m – 1) each and in the second block [L – 2 (m – 1)].

It will be useful to have an expression for the total of all the L frequencies g_i. Whether m = min (a, L – a + 1) = a or L – a + 1, this is simply given by:

$$\sum_{i=1}^{L} g_i = (m-1)\,m + m\,[L-2\,(m-1)] = m\,(L-m+1)$$

$$= a\,(L-a+1) \qquad (11)$$

It will be convenient to consider the following three cases separately, corresponding to three suitable ranges on L. For the cases I and II below, m = a, and for case III, m = L – a + 1.

The case when m = a = L – a + 1, i.e. when L = 2a – 1, is included in case II.

Case I : L > 3 (a – 1). This ensures the middle block of at least 'a' terms each equal to a.

Case II : 2 (a – 1) < L ≤ 3 (a – 1). This ensures the middle block of at least one but not more than (a – 1) terms each equal to a.

Case III : a ≤ L ≤ 2 (a – 1). This corresponds to m = L – a + 1 and provides the middle block of (2a – L) terms each equal to (L – a + 1).

For the sake of simplicity, we shall write only S, instead of S(L, a, r).

Case I : L > 3(a – 1)

Depending upon r, the summation involves either the terms from the first 2 blocks or from the second block alone, or from the second and third blocks. The following sub-cases cover all the possible values of r.

Sub-Case 1 : $0 \le r \le a-2$

S consists of the sum of the last (a – 1 – r) terms from the first block and the first (r + 1) terms from the second block. Thus

$$S = a^2 - (1/2)(a-r-1)(a-r) \qquad (12)$$

Sub-Case 2 : $a-1 \le r \le L-2a+1$

S consists of all the a terms each equal to a, hence

$$S = a^2 \tag{13}$$

Sub-Case 3 : $L-2a+2 \le r \le L-a$

S consists of the last $(L - a + 1 - r)$ terms each equal to a from the second block and the first $(2a + r - L - 1)$ terms from the third block. Obviously,

$$S = a^2 - (1/2)(2a+r-L-1)(2a+r-L) \tag{14}$$

Case II : $2(a-1) < L \le 3(a-1)$

In this case, depending upon r, the summation will consist of the terms from either the first two blocks, or from all three blocks, or from only second and third blocks. Thus the following three sub-cases arise:

Sub-Case 1 : $0 \le r \le L-2a+1$

This results in the same situation as the sub-case 1 of Case I, and consequently the expression for S is the same as in that sub-case.

Sub Case 2 : $L-2a+2 \le r \le a-2$

Here the summation extends over the last $(a - 1 - r)$ terms from the first block, all the $L - 2 (a - 1)$ terms of the middle block, and the first $(2a + r - L - 1)$ terms of the third block. It is easily seen that

$$S = a^2 - (\tfrac{1}{2})(a-r-1)(a-r) - (\tfrac{1}{2})(2a+r-L-1)(2a+r-L) \tag{15}$$

Sub-Case 3 : $a-1 \le r \le L-a$

This results again in the same situation as the sub-case 3 of Case I, and consequently in the same expression for S as in that sub-case.

Case III : $a \le L \le 2(a-1)$

In this case, $M = L - a + 1$. We separate the following sub-cases corresponding to the different values of r.

Sub-Case 1 : $r = 0$

The summation involves all the terms of the first and second blocks and entirely leaves out the third block consisting of $(L - a)$ terms. We make use of result (11) and obtain:

$$S = a(L-a+1) - (1/2)(L-a)(L-a+1) \tag{16}$$

Sub-Case 2 : $1 \le r \le m-2$

The summation consists of the last $(m - 1 - r)$ terms of the first block, all the $(2a - L)$ terms of the second block, and finally the first r terms of the third block. Thus, again utilizing the result (11), we get:

$$S = a(L - a + 1) - (1/2)(L - a - r)(L - a - r + 1) - (1/2)r(r + 1) \quad (17)$$

Sub-Case 3 : $r = m - 1 = L - a$

Here the summation involves all the terms of the second and third blocks, and entirely leaves out the terms of the first block. Consequently, S is the same as in sub-case 1 above, and is given by (16).

It is easily seen that (16) is a special case of (17) when $r = 0$ or $L - a$. Thus (17) covers all the three sub-cases of Case III.

Further, in the trivial case when $L = a$, the only value for r is 0, and we get $S = a$.

All the expressions obtained above for the values of the function S are given on the next page in a tabular form. It should be remembered that L and a are positive integers with $a \leq L$ while r is a non-negative integer such that $0 \leq r \leq L - a$.

It can be algebraically verified that all the different values of S (L,a,r) for different L, a, and r, given in the table for the various cases can be represented by the following elegant and compact form:

$$S(L, a, r) = a^2 - (1/2)\alpha(\alpha - 1) - (1/2)\beta(\beta - 1) \quad (18)$$

where $\alpha = \max(a - r, 0)$ and $\beta = \max(2a + r - L, 0)$

Following the same procedure used for the derivation of S(L, a, r), it is clear that the summation (8) will be given by

Table 1: Values of S (L, a, r) for different L, a, and r

Values of L and a	Values of r in the range $0 \leq r \leq L - a$	Values of S(L, a, r)
$L > 3(a - 1)$	$0 \leq r \leq a - 2$	$a^2 - (1/2)(a - r - 1)(a - r)$
	$a - 1 \leq r \leq L - 2a + 1$	a^2
	$L - 2a + 2 \leq r \leq L - a$	$a^2 - (1/2)(r - L + 2a - 1)(r - L + 2a)$
$2(a - 1) < L \leq 3(a - 1)$	$0 \leq r \leq L - 2a + 1$	$a^2 - (1/2)(a - r - 1)(a - r)$
	$L - 2a + 2 \leq r \leq a - 2$	$a^2 - (1/2)(a - r - 1)(a - r) - (1/2)(2a + r - L - 1)(2a + r - L)$
	$a - 1 \leq r \leq L - a$	$a^2 - (1/2)(r - L + 2a - 1)(r - L + 2a)$
$a \leq L \leq 2(a - 1)$	$0 \leq r \leq L - a$	$a(L - a + 1) - (1/2)(L - a - r)(L - a - r + 1) - (1/2)r(r + 1)$

$$S(B, b, s) = b^2 - (1/2)\alpha(\alpha - 1) - (1/2)\beta(\beta - 1) \quad (19)$$

where $\alpha = \max(b - s, 0)$ and $\beta = \max(2b + s - B, 0)$

4. Correction Factor Suggested

In an unpublished note, the author had pointed out that since the probabilities with which the individual cells had been sampled were not equal, the plot-yield from the plots (around the borders of a field) comprising of cells having smaller probability, should be given an appropriately higher weight by multiplying their yields with a factor greater than one, and the plot-yield from the plots (near the center of a field) comprising cells having larger probability should be given a lower weight by multiplying their yield with a factor less than one. The suggested correction factor or weight was arrived at by the following considerations.

If equal chance were available to each one of the $L \times B$ cells to be sampled, the probability of a cell being included in a plot of size $a \times b$ cells would have been $Q = (ab)/(LB)$. By extending the argument to the case when the chances given to each one of the $L \times B$ cells were not equal, we stated that probability of a cell being sampled in the plot comprising $a \times b$ cells would now be:

$$P = f(r, s) \Big/ \sum_{r=0}^{L-a} \sum_{s=0}^{B-b} f(r, s) \tag{20}$$

where $f(r, s)$ is given by (5).

By utilizing (6), (11), (18) and (19), it is seen that (20) reduces to:

$$P = \frac{a^2 - \frac{1}{2}\alpha(\alpha-1) - \frac{1}{2}\beta(\beta-1)}{a(L-a+1)} \times \frac{b^2 - \frac{1}{2}\alpha'(\alpha'-1) - \frac{1}{2}\beta'(\beta'-1)}{b(B-b+1)} \tag{21}$$

where α, β, α', and β' are given in (18) and (19).

Since unequal probability sampling was regarded as a "biased" procedure, it was suggested that the bias could be removed by dividing the plot yield by the unequal probability (21) and multiplying it by the equal probability $Q = ab/LB$. This factor w, where

$$w = Q/P \tag{22}$$

seemed to satisfy the requirement that $w > 1$ for the border plots and $w < 1$ for the central plots. This factor w, along with the compact formulae (18) and (19) was given in the notes of the seminar at Iowa State [2].

5. Relationship to Sampling with Unequal Probabilities

Although the suggested "correction factor" was meant to remove the "bias" introduced by unequal probabilities of sampling, it is only a step away from

the generalization of sampling from equal to unequal probabilities. For example, if a sample $y_1, y_2, ..., y_n$ has been obtained by assigning unequal probabilities to the different sampling units of the finite population and this results in a situation that the i^{th} unit of the population, say u_i, will be included in a sample of size n with probability $P(u_i)$, the correction factor w would suggest that if the population mean were to be estimated by

$$\bar{y} = \sum_{i=1}^{n} y_i / n$$

it would be a biased estimator. In order to remove this bias, it would be necessary to multiply each selected y_i (i = 1, ..., n) by a proper correction factor w_i, which in this case works out as Q/P_i, where Q = probability of a unit of the population to be included in the sample with equal probabilities = n/N, and $P_i = P(u_i)$. Thus, we obtain an unbiased estimator of the population mean as:

$$\frac{1}{n}\sum_{i=1}^{n} y_i \, Q/P(u_i) = \frac{1}{N}\sum_{i=1}^{n} y_i / P(u_i) \qquad (23)$$

This estimator is known in the literature as the Horvitz-Thompson estimator [3].

ACKNOWLEDGEMENTS

I am grateful to Dr. P.V. Sukhatme for kindling my interest in this problem in the first place, and later by persuading me to write-up this paper in view of its importance in the history of the development of the theory and practice of sampling from finite populations with unequal probabilities. I would also like to thank Vikas Aggarwal for his help in typing the paper and for his helpful comments.

REFERENCES

[1] Johnson, N.L. and Smith, H., Jr., 1969. *New Developments in Survey Sampling.* Wiley-Intersciences, 528-561, New York.

[2] Aggarwal, O.P., 1949-50. Crop-cutting experiments on random sampling basis for the estimation of yield of crops in India (Seminar). *Annual Report of the Statistical Laboratory of Iowa State College*, 11.

[3] Horvitz, D.G. and Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe, *J. Amer. Statist. Assoc.*, 47, 663-685.

[4] Sukhatme, P.V., and Sukhatme, B.V., 1970. *Sampling Theory of Surveys with Application, Second Revised Edition*, FAO, Rome.