# A NOTE ON SAMPLING WITH VARYING PROBABILITIES

BY M. S. CHIKKAGOUDAR,
*Karnatak University, Dharwar*

## INTRODUCTION

In case of sampling with varying probabilities and without replacement Horvitz and Thompson (1952) provided an estimator of a population total and the variance estimator respectively as

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{P(i)} \qquad \ldots(1)$$

and

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n} \frac{1-P(i)}{P^2(i)} y_i^2 + \sum_{i \neq j=1}^{n} \frac{P(i)-P(ij)\,P(j)}{P(ij)\,P(i)\,P(j)} y_i y_j \quad \ldots(2)$$

where $P(i)$ and $P(ij)$ are respectively the inclusion probabilities of the unit $(i)$ and the pair of units $(ij)$ in a sample of size $n$. It is well known that for samples of sizes greater than two these estimators are of less practical use because of the tedious algebra involved in calculating the probabilities $P(i)$ and $P(ij)$ in such cases. As an alternative we can use the above estimators (1) and (2) in conjunction with Sen's (1955) method of sampling, which is a generalization of Midzuno's (1950) method and which consists in selecting only the first few, say $r$, units with varying probabilities and the remaining $(n-r)$ units with equal probabilities, the selection being without replacement throughout. For this purpose we require the probabilities $P(i)$ and $P(ij)$ for the above sampling plan. Also, if third and higher moments of $\hat{Y}$ and variance of the estimator $\hat{V}(\hat{Y})$ given in (2) are to be found the inclusion probabilities of more than two units in the sample are essential. In this paper a general formula is derived in a compact form for obtaining the probability $P(j_1 j_2 \ldots j_m)$ of including a specified set $(j_1 j_2 \ldots j_m)$ of $m$ units in a sample of size $n$, drawn according to Sen's method and some interesting particular cases of this probability are given. The summation of $P(j_1\, j_2 \ldots j_m)$ over all the population units excepting the $(m-1)$ units $(j_1\, j_2 \ldots j_{m-1})$ is shown to be $(n-m+1)\; P(j_1\, j_2 \ldots j_{m-1})$ which is in keeping with the result

$$\sum_{j(=i)}^{N} P(i_j)=(n-1)\; P(i) \qquad \ldots(3)$$

given by Horvitz and Thompson (1952).

INCLUSION PROBABILITY

Consider a population of size $N$ with the $i$-th unit having the initial selection probability $P_i$ for $i=1, 2, \ldots, N$. Let $Q(i_1\, i_2 \ldots i_r)$ be the probability of selecting a set of $r$ units $(i_1\, i_2 \ldots i_r)$ from the population with initial selection probabilities $p_{i_1},\, p_{i_2}, \ldots, p_{i_r}$ respectively. Also, let $P(n, r)$ be the probability of selecting a set of specified units $(i_1 \ldots i_{n-m}\, j_1 \cdot j_m)$ containing the set $(j_1 \ldots j_m)$ when a sample of size $n$ is drawn such that any $r$ units of the above set are first selected with unequal probabilities and the remaining $(n-r)$ units with equal probabilities, the selection being without replacement throughout.

Then as is shown by Sen (1955)

$$Q(i_1\, i_2 \ldots i_r)=\sum_{j=1}^{r} P_{i_j}\; Q_{i_j}\; (i_1 \ldots i_{j-1}\; i_{j+1} \ldots i_r) \qquad \ldots(4)$$

where $Q i_j$ indicates that the unit $i_j$ is eliminated as a possible selection and

$$P(n, r)=\frac{1}{\binom{N-r}{n-r}} \sum_{(n;\, r)} Q(i_1\, i_2 \ldots i_r) \qquad \ldots(5)$$

where $\sum_{(n;\, r)}$ denotes the summation over all the combinations of $r$ units out of $(i_1 \ldots i_{n-m}\, j_1 \ldots j_m)$. The first $r$ units selected with varying probabilities may contain $K$ units of the set $(j_1\, j_2 \ldots j_m)$ and $(r-k)$ units of the set $(i_1\, i_2 \ldots i_{n-m})$, where $K$ may be any number from 0 to $m$. Hence the probability $P(n, r)$ can be written as

$$P(n, r)=\frac{1}{\binom{N-r}{n-r}} \sum_{k=0}^{m} \sum_{(m;\, k)} \sum_{(n-m;\, r-k)} Q(i_1 \ldots i_{r-k}\, j_1 \ldots j_k) \ldots(6)$$

where $Q(i_1 \ldots i_{r-k}\, j_1 \ldots j_k)$, as defined earlier, is the probability of selecting the set of $r$ units $(i_1 \ldots i_{r-k}\, j_1 \ldots j_k)$ with respective selection probabilities $p_{i_1}, \ldots, p_{i_{r-k}}, p_{j_1} \ldots p_{j_k}$.

Summing the equation (6) on both sides over all the $\binom{N-m}{n-m}$ samples of size $n$ containing the set $(j_1 j_2 \ldots j_m)$, we get

$$P(j_1 \ldots j_m) = \frac{1}{\binom{N-r}{n-r}} \sum_{(N-m;\, n-m)} \sum_{k=0}^{m} \sum_{(m;\, k)} \sum_{(n-m;\, r-k)} Q(i_1 \ldots i_{r-k}\, j_1 \ldots j_k)$$

$$= \frac{1}{\binom{N-r}{n-r}} \sum_{k=0}^{m} \binom{N-m-r+k}{n-m-r+k} \left[ \sum_{(m;\, k)} \sum_{(N-m;\, r-k)} Q(i_1 \ldots i_{r-k}\, j_1 \ldots j_k) \right] \quad \ldots(7)$$

But the quantity inside the square brackets is nothing but the probability $P[k]$ that exactly $k$ units of $(j_1 j_2 \ldots j_m)$ are included in a sample of size $r$, drawn with varying probabilities and without replacement. Let $Q_r(j_1 \ldots j_p)$ be the probability of inclusion of the set $(j_1 \ldots j_p)$ in a sample of size $r$ where all the $r$ units are drawn with varying probabilities and without replacement. The distinction between $Q_r(j_1 \ldots j_p)$ and $P(j_1 \ldots j_p)$ should be carefully noted. Also let

$$S_p = \sum_{(m;\, p)} Q_r(j_1 \ldots j_p) \text{ for } p = 1, 2, \ldots, m \quad \ldots(8)$$

and

$$S_o = 1.$$

In the above notations, the probability $P[k]$ is given by

$$P[k] = \sum_{p=0}^{m-k} (-1)^p \binom{k+p}{p} S_{k+p} \quad \ldots(9)$$

which follows from the eqn. (3·1), p. 96 of Feller (1960).

Substituting from (9) in (7) we obtain

$$P(j_1 \ldots j_m) = \frac{1}{\binom{N-r}{n-r}} \sum_{k=0}^{m} \binom{N-m-r+k}{n-m-r+k} \sum_{p=0}^{m-k} (-1)^p \binom{k+p}{p} S_{k+p} \quad \ldots(10)$$

The coefficient of $S_u$ in the right-hand side of the above equation can be found to be

$$\frac{1}{\binom{N-r}{n-r}} \sum_{v=0}^{u} (-1)^{u-v} \binom{N-m-r+v}{n-m-r+v} \binom{u}{u-v}$$

This coefficient simplifies to

$$\frac{\binom{N-m-r}{N-n-u}}{\binom{N-r}{n-r}}$$

after the evaluation of the summation by equating the coefficients of $x^{(n-m-r+u)}$ on two sides of the identity

$$(1-x)^{-(N-n+1)} \times (1-x)^{u} = (1-x)^{-(N-n-u+1)}.$$

Hence the inclusion probability given in (10) becomes

$$P(j_1 \ldots j_m) = \frac{1}{\binom{N-r}{n-r}} \sum_{u=0}^{m} \binom{N-m-r}{N-n-u} S_u \qquad \ldots(11)$$

since $u$ can take the values from 0 to $m$.

Special cases :

1. $r$, some general value
   *i.e.*, Sen's method of sampling

(*i*) $m=1$

$$P(i_1) = \frac{(n-r)}{(N-r)} + \frac{(N-n)}{(N-r)} Q_r(j_1) \qquad \ldots(12)$$

(*ii*) $m=2$,

$$P(j_1 j_2) = \frac{1}{(N-r)(N-r-1)} [(n-r)(n-r-1) + (n-r)(N-n) \{Q_r(j_1) + Q_r(j_2)\} + (N-n)(N-n-1) Q_r(j_1 j_2)] \qquad \ldots(13)$$

(*iii*) $m=n$,

$$P(j_1 \ldots j_n) = \frac{1}{\binom{N-r}{n-r}} \sum_{(n;\, r)} Q_r(j_1 \ldots j_r) \qquad \ldots(14)$$

which is same as that given by Sen (1955).

2. $r=2$

(*i*) $m=1$,

$$P(j_1) = \frac{(n-2)}{(N-2)} + \frac{(N-n)}{(N-2)} Q_2(j_1) \qquad \ldots(15)$$

(*ii*) $m=2$,

$$P(j_1 j_2) = \frac{1}{(N-2)(N-3)} [(n-2)(n-3) + (n-2)(N-n) \{ Q_2(j_1) + Q_2(j_2) \} + (N-n)(N-n-1) Q_2(j_1 j_2)] \qquad (16)$$

These formulae (15) and (16) for $P(j_1)$ and $P(j_1 j_2)$ can be shown to be equivalent to the corresponding formulae given by J. Rao (1961) in his eqns. (6) and (9) respectively.

(3) $r=1$ *i.e.*, Midzuno's method of sampling

(*i*) $m=1$,

$$P(j_1)=\frac{1}{N-1}\left[\ (n-1)+(N-n)pj_1\ \right] \qquad ...(17)$$

(*ii*) $m=2$,

$$P(j_1 j_2)=\frac{(n-1)}{(N-1)}\left[\frac{n-2}{N-2}+\frac{N-n}{N-2}\ (pj_1+pj_2)\ \right] \qquad ...(18)$$

It can be seen that $p(j_1)$ and $p(j_1 j_2)$ given by eqns. (17) and (18) are same as those given by Horvitz and Thompson (1952) in their eqns. (19) and (20) respectively

(4) $r=o$, $m=$same general value *i.e.*, equal probability sampling

$$P(j_1 \cdots j_m)=\frac{\binom{N-m}{n-m}}{\binom{N}{n}}$$

as is expected to be.

The summation of the inclusion probability $P(j_1 \cdots j_m)$ over all the values of $j_m$ (say) from 1 to $N$ excepting the values $(j_1 \cdots j_{m-1})$ can be shown to be equivalent to $(n-m+1)\ P(j_1 \cdots j_{m-1})$. The eqn. (11) can be written as

$$P(j_1 \cdots j_m)=\frac{1}{\binom{N-r}{n-r}} \sum_{u=0}^{m} \binom{N-r-m}{N-n-u}\left[\sum_{(m-1;\,u-1)} Q_r(j_1 \cdots j_{u-1} j_m)\right.$$

$$\left. + \sum_{(m-1;\,u)} Q_r(j_1 \cdots j_{u-1} j_u)\right] \qquad ...(20)$$

where $Q_r(j_1 \cdots j_{u-1} j_m)$ is the inclusion probability of a set of any $u$ units from $(j_1 j_2 \cdots j_m)$ containing the unit $(j_m)$ and $Q_r(j_1 \cdots j_{u-1} j_u)$ is the inclusion probability of $u$ units which does not contain the unit $(j_m)$ in a sample of size $r$.

Hence,

$$\sum_{\substack{j_m=1\\ \neq j_1\neq j_2\ldots\neq j_{m-1}}}^{N} P(j_1\ldots j_m) = \frac{1}{\binom{N-r}{n-r}} \sum_{u=0}^{m} \binom{N-r-m}{N-n-u}$$

$$\left[\sum_{(m-1;\,u-1)} \sum_{\substack{j=1\\ \neq j_1\neq j_2\ldots\neq j_{u-1}}}^{N} Q_r(j_1\ldots j_{u-1}j_m)\right.$$

$$\left.-\sum_{(m-1;\,u-1)} \sum_{l=u}^{m-1} Q_r(j_1\ldots j_{u-1}j_l)+(N-m+1)\sum_{(m-1;u)} Q_r(j_1\ldots j_{u-1}j_u)\right] \qquad \ldots (21)$$

As a consequence of the generalization of Horvitz and Thompson's (1952) result given in (3), the first term inside the square brackets reduces to

$$(r-u+1)\sum_{(m-1;\,u-1)} Q_r(j_1\ldots j_{u-1}\,j_u) \qquad \ldots (22)$$

A little examination will show that the second term inside the square brackets becomes

$$-u \sum_{(m-1;\,u)} Q_r(j_1\ldots j_{u-1}j_n) \qquad \ldots (23)$$

Substituting from (22) and (23) in (21) and simplifying we get

$$\sum_{\substack{j_m=1\\ \neq j_1\neq j_2\ldots\neq j_{m-1}}}^{N} P(j_1\ldots j_m) = \frac{1}{\binom{N-r}{n-r}} \sum_{u=0}^{m} \binom{N-r-m}{N-n-u}$$

$$\left[(r-u+1)\sum_{(m-1;\,u-1)} Q_r(j_1\ldots j_{u-1})\right.$$

$$\left.+(N-m-u+1) \sum_{(m-1;\,u)} Q_r(j_1\ldots j_{u-1}j_u)\right] \qquad \ldots (24)$$

Adding the second part of $(k+1)$th term and the first part of the $(k+2)$th term of (24) we get

$$(n-m+1)\frac{\binom{N-r-m+1}{N-n-u}}{\binom{N-r}{n-r}} \sum_{(m-1;\,u)} Q_r(j_1\ldots j_{u-1}j_u)$$

This is true for all values of $u$ from $o$ to $m-1$ and when $u=m$ the second part of the last term in (24) will vanish.

Hence (24) reduces to

$$\sum_{\substack{j_m=1 \\ \neq j_1 \neq j_2 \ldots \neq j_{m-1}}}^{N} P(j_1 \ldots j_m) = (n-m+1) \left\{ \frac{1}{\binom{N-r}{n-r}} \sum_{u=0}^{m-1} \binom{N-r-m+1}{N-n-u} \sum_{m-1;\, u} Q_r(j_1 \ldots j_u) \right\}$$

$$= (n-m+1)\, P(j_1 \ldots j_{m-1}) \qquad \ldots(25)$$

from eqn. (11).

Proceeding as above or otherwise it can be shown that

$$\Sigma' \, P(j_1 \ldots j_m) = n(n-1) \ldots (n-m+1) \qquad \ldots(26)$$

where $\Sigma'$ stands for the summation over all the permutations of $m$ units out of $N$ in the population.

The estimator of population total, its variance and the variance estimator can be easily obtained by substituting the values of $P(i)$ and $P(ij)$ (for desired value of $r$), evaluated from eqn. (11), in the corresponding formulae given by Horvitz and Thompson (1952).

## Acknowledgment

The author is very much thankful to Prof. M.V. Jambunathan, Department of Statistics, K.U.D., for his guidance and encouragement in preparing this note.

## References


1. Feller, W. "An Introduction to probability theory and its applications", John Wiley and Sons, New York, 1960.
2. Horvitz, D.G. and Thompson, D.J. "A generalization of sampling without replacement from a finite universe", J. Amer. Statists. Assoc., 1952, *47*, 663-85.
3. Midzuno, H. "An outline of the theory of sampling systems", Ann. Inst. Statist. Math. 1950, *1*, 149-156.
4. Rao, J.N.K. "On the estimation of the variance in unequal probability sampling", Ann. Inst. Statist. Math., 1961, *13*, 57-60.
5. Sen, A.R. "On the selection of $n$ primary sampling units from a structure ($n$=2)", Ann. Math. Statist., 1955, *26*, 744-751.