USE OF 'ORDER-STATISTIC' IN SAMPLING WITHOUT REPLACEMENT

By P. K. PATHAK

Indian Statistical Institute

SUMMARY. In sampling without replacement from a finite population, the order in which the units are selected, is immaterial for the purpose of estimation. This point was noted by Basu (1958) and Murthy (1957). Basu showed that the 'order-statistic' (sample units arranged in ascending order of their unit indices) forms a sufficient statistic, and, therefore, any estimator which is not a function of the order-statistic, can be uniformly improved by the use of Rao-Blackwell theorem. In this paper, certain results obtained by Murthy are shown to be immediate consequences of the above observation.

It is shown that sampling with different probabilities with replacement, until we get a specified number of distinct units, is equivalent in some sense to sampling with different probabilities without replacement. Some other related problems are also considered here.

1. Introduction

Let $Y_1, Y_2, \ldots, Y_N$ be the Y-characteristic of the N population units under study. Let P_j be the probability associated with the j-th population unit $(j = 1, \ldots, N)$. For simplicity, we shall always refer population units by capital letters and sample units by small letters, e.g., y_i and p_i denote the variate value and the probability of selection associated with the i-th sample unit respectively. In this paper, we shall throughout follow the notations used by Basu (1958).

2. Sampling without replacement

In sampling without replacement from the above population a particular sample may be recorded as

$$s = (x_1, x_2, \ldots, x_n),$$

where $x_i = (y_i, p_i, u_i)$ $(i = 1, 2, \ldots, n)$, and n is the sample size.

The probability of drawing such a particular sample is given by

$$P(s) = p_1 \cdot \frac{p_2}{(1-p_1)} \cdot \frac{p_3}{(1-p_1-p_2)} \cdots \frac{p_n}{(1-p_1-\cdots-p_{n-1})} \cdot \ldots \quad (2.1)$$

If we record the 'order-statistic' by

$$T = (x_{(1)}, \ldots, x_{(n)}),$$

where $x_{(i)} = (y_{(i)}, p_{(i)}, u_{(i)})$ is the i-th order-statistic ($i = 1, \ldots, n$), we have

$$P(T) = \sum_{s \to T} \frac{p_1 p_2 \cdots p_n}{(1-p_1)(1-p_1-p_2)\cdots(1-p_1-p_2-\cdots-p_{n-1})}, \quad \ldots \quad (2.2)$$

where the summation is taken over all possible samples giving rise to the 'order-statistic' T.

It has been shown by Basu (1958) that T is a sufficient statistic. Thus, if $g(s)$ is some estimator depending only on s, by Rao-Blackwell theorem, a uniformly better estimator than $g(s)$ is given by $E(g(s)|T)$. For any convex loss function, the risk associated with $E(g(s)|T)$ is smaller than the risk associated with $g(s)$.

3. SAMPLING WITH REPLACEMENT : NUMBER OF DISTINCT UNITS FIXED IN ADVANCE

In this case, units are drawn with unequal probabilities and with replacement until we get a specified number 'n' of distinct units. If r denotes the number of draws in a particular case, the sample s may be recorded as

$$s = (x_1, \ldots, x_r).$$

If we define the 'order-statistic' T by

$$T = [x_{(1)}, x_{(2)}, \ldots, x_{(n)}],$$

where $x_{(i)}$ is the i-th 'order-statistic' ($i = 1, 2, \ldots, n$), it is not difficult to show that

$$P(T) = \sum_{r=n}^{\infty} \Big[\sum_{i=1}^{n} p_{(i)}\{\Sigma^{(i)}(p_{(1)}+\ldots+p_{(n)})^{r-1} - \Sigma^{(i)}(p_{(1)}+\ldots+p_{(n-1)})^{r-1} + \ldots(-)^{n-2}\Sigma^{(i)}p_{(1)}{}^{r-1}\}\Big], \quad \ldots \quad (3.1)$$

where $\Sigma^{(i)}$ denotes the summation over all possible combinations out of $p_{(1)}, \ldots, p_{(i-1)}, p_{(i+1)}, \ldots, p_{(n)}$, and the term inside the square brackets denotes the probability of getting T in r draws. Assuming without any loss of generality that $p_{(1)}+\ldots+p_{(n)}<1$, we get on summing (3.1) over r

$$P(T) = \sum_{i=1}^{n} p_{(i)}\Big[\Sigma^{(i)}\frac{p_{(1)}+\ldots+p_{(n)}}{1-p_{(1)}-\ldots-p_{(n)}} - \Sigma^{(i)}\frac{p_{(1)}+\ldots+p_{(n-1)}}{1-p_{(1)}-\ldots-p_{(n-1)}} + (-)^{n-1}\Sigma^{(i)}\frac{p_{(1)}}{1-p_{(1)}}\Big], \quad \ldots \quad (3.2)$$

It can be proved by induction over n that (3.2) and (2.2) are equal.

Thus, if we rely only on the 'order-statistic' T, the two methods of sampling are essentially the same.

4. Improving Des Raj's estimators

Des Raj (1956), in sampling without replacement, gave the following set of uncorrelated estimators of $Y = \sum_{i=1}^{N} Y_i$.

$$\begin{aligned}
t_1(s) &= y_1/p_1; \\
t_2(s) &= y_1 + \frac{y_2}{p_2}(1-p_1); \\
&\cdots \qquad \cdots \qquad \cdots \qquad \cdots \quad (4.1) \\
t_i(s) &= y_1 + y_2 + \cdots + y_{i-1} + \frac{y_i}{p_i}(1 - p_1 - p_2 - \cdots - p_{i-1}); \\
&\cdots \qquad \cdots \qquad \cdots \\
t_n(s) &= y_1 + y_2 + \cdots + y_{n-1} + \frac{y_n}{p_n}(1 - p_1 - p_2 - \cdots - p_{n-1}).
\end{aligned}$$

Theorem 1 : *For any convex loss function a uniformly better estimator than*

$$t(s) = \sum_{i=1}^{n} c_i t_i(s) \quad (\sum_{i=1}^{n} c_i = 1)$$

is given by

$$E[t(s) \mid T] = \sum_{i=1}^{n} y_{(i)} \frac{P(T/(i))}{P(T)}, \qquad \cdots \quad (4.2)$$

where $P(T/(i))$ is the conditional probability of getting the 'order-statistic' T given that i-th order unit was drawn first.

Proof : Since

$$\frac{P[x_i = x_{(i)}, x_{i+1} = x_{(i)} \mid x_1, x_2, \ldots, x_{i-1}]}{P[x_i = x_{(i)}, x_{i+1} = x_{(i)} \mid x_1, x_2, \ldots, x_{i-1}]} = \frac{1 - p_1 - p_2 - \cdots - p_{i-1} - p_{(i)}}{1 - p_1 - p_2 - \cdots - p_{i-1} - p_{(i)}}, \cdots \qquad \cdots \quad (4.3)$$

it follows that

$$E[t_{i+1}(s) - t_i(s) \mid T] = E\left[\frac{y_{i+1}}{p_{i+1}}(1-p_1-p_2-\cdots-p_i) - \frac{y_i}{p_i}(1-p_1-p_2-\cdots-p_i) \mid T\right] = 0.$$

Therefore,

$$E\left[\sum_{i=1}^{n} c_i t_i(s) \mid T\right] = E[t_1(s) \mid T] = \sum_{i=1}^{n} y_{(i)} \frac{P(T \mid (i))}{P(T)}.$$

Corollary 1 : When $n = 2$, we have

$$E\left[\sum_{i=1}^{2} c_i t_i(s) \mid T\right] = \frac{1}{(2-p_1-p_2)} \left[(1-p_2) \frac{y_1}{p_1} + (1-p_1) \frac{y_2}{p_2}\right].$$

Corollary 2 : In simple random sampling without replacement

$$t_i(s) = y_1 + \ldots + y_{i-1} + (N-i+1) y_i, \quad i = 1, \ldots n;$$

and

$$E[t_i(s) \mid T] = E[t_1(s) \mid T] = \frac{N}{n} \sum_{i=1}^{n} y_i.$$

Theorem 2 : *A uniformly better estimator than*

$$g(s) = \sum_{i \neq j=1}^{n} c_{ij} t_j(s) t_j(s)$$

$\left(\sum_{i \neq j=1}^{n} c_{ij} = 1\right)$ *of* Y^2, *is given by*

$$g(T) = E[g(s) \mid T] = \frac{\sum_{i=1}^{n} y^2_{(i)} P(T \mid (i)) + \sum_{i \neq i'=1}^{n} y_{(i)} y_{(i')} P(T \mid (i), (i'))}{P(T)}. \quad \ldots \quad (4.4)^*$$

Proof : Using (4.3), it can be seen that

$$E[t_1(s) t_{j+1}(s) - t_1(s) t_j(s) \mid T] = 0, \qquad (j = 2, \ldots, n)$$

and

$$E[t_{i+1}(s) t_j(s) - t_i(s) t_j(s) \mid T] = 0, \qquad (i = 1, 2, \ldots, j-2)$$

and hence

$$E[g(s) \mid T] = E[t_1(s)\ t_2(s) \mid T]$$

$$= \frac{\sum_{i=1}^{n} y^2_{(i)} P(T \mid (i)) + \sum_{i \neq i'=1}^{n} y_{(i)} y_{(i')} P(T \mid (i), (i'))}{P(T)}$$

---

* *Remark :* The estimators (4.2) and (4.4) can also be obtained by improving the usual estimators of Y and Y^2 under the sampling scheme discussed in Section 3.

Corollary 3 : When $n = 2$, we have

$$E[t_1(s)\,t_2(s)|T] = \frac{1}{(2-p_{(1)}-p_{(2)})}\left[(1-p_{(2)})\,\frac{y^2_{(1)}}{p_{(1)}} + (1-p_{(1)})\,\frac{y^2_{(2)}}{p_{(2)}} + 2(1-p_{(1)})(1-p_{(2)})\,\frac{y_{(1)}}{p_{(1)}}\,\frac{y_{(2)}}{p_{(2)}}\right].$$

Corollary 4 : In simple random sampling (without replacement)

$$E[t_i(s)\,t_j(s)|T] = \frac{N}{n}\sum_{i=1}^{n} y^2_{(i)} + \frac{N(N-1)}{n(n-1)}\sum_{i\neq j=1}^{n} y_{(i)}\,y_{(j)}.$$

The estimator (4.4) is used to derive unbiased variance estimator of

$$\sum_{i=1}^{n} y_{(i)}\,\frac{P(T|(i))}{P(T)}.$$

5. Improving Das' estimators

The set of estimators of Y given by Das (1951) is as follows :

$$u_1(s) = \frac{y_1}{p_1}\ ;$$

$$u_2(s) = \frac{y_2}{p_2}\cdot\frac{(1-p_1)}{p_1}\cdot\frac{1}{(N-1)}\ ;$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$u_r(s) = \frac{y_r}{p_r}\cdot\frac{(1-p_1-p_2\cdots-p_{r-1})}{p_{r-1}}\cdot\cdots\cdot\frac{(1-p_1-p_2)}{p_2}\cdot\frac{1-p_1}{p_1}\cdot\frac{1}{(N-1)(N-2)\cdots(N-r+1)}\ ;$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$u_n(s) = \frac{y_n}{p_n}\cdot\frac{(1-p_1-p_2\cdots-p_{n-1})}{p_{n-1}}\cdot\cdots\cdot\frac{(1-p_1-p_2)}{p_2}\cdot\frac{(1-p_1)}{p_1}\cdot\frac{1}{(N-1)(N-2)\cdots(N-n+1)}\cdot\cdots \quad (5.1)$$

A uniformly better estimator than $u_r(s)$ is, therefore, given by

$$u_r(T) = E[u_r(s)|T]$$

$$= \sum_{i=1}^{n} y_{(i)}\,\frac{\Sigma'\dfrac{1}{(N-1)(N-2)\cdots(N-r+1)}\,P(T|x_1, x_2, \ldots, x_{r-1}, x_r = x_{(i)})}{P(T)} \qquad r = 1, \ldots n \quad \ldots \quad (5.2)$$

where the summation Σ' is taken over all possible $x_1, \ldots, x_{r-1}$.

It is easy to see that the estimators $u_r(T)$ $(r = 1, \ldots, n)$ are identical if and only if the sample is drawn by simple random sampling (without replacement). In this case (5.2) is same as (4.2).

This shows that in simple random sampling (without replacement) the estimator based on the sample mean is more efficient than Das' as well as Des Raj's estimators.

An unbiased estimate of Y^2 based on $u_r(s)$ and $y_k (k < r)$, is given by

$$v_{rk}(s) = u_r(s) y_r + (N-1)\, u_r(s) y_k \quad (k < r = 1, 2, \ldots, n). \qquad \ldots \quad (5.3)$$

A uniformly better estimator than this is given by

$$E[v_{rk}(s)|T] = \sum_{i=1}^{n} y_{(i)}^2 \frac{\left[\Sigma' \frac{1}{(N-1)} \cdots \frac{1}{(N-r+1)} P(T|x_1, \ldots x_{r-1}, x_r = x_{(i)})\right]}{P(T)}$$

$$+ (N-1) \sum_{i \neq j=1}^{n} y_{(i)} y_{(j)} \frac{\left[\Sigma' \frac{1}{(N-1)} \cdots \frac{1}{(N-r+1)} \left\{P(T|x_1, \ldots, x_k = x_{(i)}, \ldots,\right.\right.$$

$$\left.\left. x_{r-1}, x_r = x_{(j)}) + P(T|x_1, \ldots, x_k = x_{(j)}, \ldots, x_{r-1}, x_r = x_{(i)}\right\}\right]. \qquad \ldots \quad (5.4)$$

This expression will also be identical for all r and k if and only if the sample is drawn by simple random sampling (without replacement).

Further, it may be seen on similar lines that in a more general (without replacement) sampling scheme which has been considered by Des Raj (1956), the estimators of Y (or of Y^2) obtained by improving Das' estimators will be identical if and only if the first unit in the sample is selected with pre-assigned probabilities and the remaining units are selected by simple random sampling (without replacement). For further reference about this, one may refer to Des Raj (1956) and Murthy (1957).

ACKNOWLEDGEMENT

I am grateful to Dr. D. Basu for his kind help and guidance in the preparation of the paper.

REFERENCES

BASU, D. (1958) : On sampling with and without replacement. *Sankhyā*, 20, 287—294.

DAS, A. C. (1951) : Two-phase sampling and sampling with varying probabilities. *Bull. Int. Stat. Inst.*, 33, 105-112.

DES RAJ (1956) : Some estimators in sampling with varying probabilities without replacement. *J. Amer. Stat. Ass.*, 51, 274, 269-284.

MURTHY, M. N. (1957) : Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Paper received : December, 1960.

Revised : April, 1961.