USE OF RANDOMISED ROUNDED OFF WEIGHTS IN SAMPLE SURVEYS

By J. ROY and G. KALYANASUNDARAM

Indian Statistical Institute

SUMMARY. Use of randomised rounded off weights in large-scale sample surveys was suggested by Murthy and Sethi (1959, 1961). In this paper, two practical methods of rounding off are proposed, which leave the estimator unbiassed and are amenable to mechanised processing on punched cards. An expression for the addition to the sampling variance due to such rounding off is worked out and empirical results suggest that its magnitude is not likely to be large. Punched card methods are developed for generating such weights and for tabulation of survey data using these weights.

1. INTRODUCTION

To increase the efficiency of estimation and also to ensure a rational distribution of work-load in field operations, ultimate sample units (s.u.) in a large-scale sample survey are very often chosen with unequal probabilities. For such schemes of sampling, an unbiassed estimate of a total for the population is generally obtained as a weighted sum of the characteristics of the s.u.'s : the weight for a particular s.u. being the reciprocal of the probability of its selection. When standard punched card equipment are used in processing the data obtained from such surveys, the first stage of processing consists in passing the data-cards through a calculating punch, multiplying the s.u. characteristic by the weight, and punching the products on the cards. These product cards are next passed through a tabulator to get the required totals. This puts a considerable load on calculating punches. Also, there is the additional source of processing error in undetected wrong products.

One simple way of obviating these difficulties would be the use of self-weighted sampling designs. This is generally impracticable for several reasons : non-availability of a complete, accurate frame at the stage of drawing up the sampling design, non-response in the case of some selected s.u.'s, uneven work-load on field-investigators, inadequate precision of estimates for sub-populations etc. What is called for is clearly a method of adjustment at the stage of processing the data.

Several different methods of adjustment have been suggested by Murthy and Sethi (1959). One method is to introduce an additional stage of sampling before processing : a sample can be drawn from the selected s.u.'s and only the s.u.'s so selected are to be included in the processing; the probabilities of selection at this stage can be so adjusted that in effect the design becomes self-weighted. This would mean throwing away the information collected from s.u.'s which are later excluded from the processing. Though the sampling design can be adapted to minimise this loss, a layman is unlikely to be happy to accept this solution.

An alternative procedure is to include all the selected s.u.'s, but to round off their weights in a suitable way. Murthy and Sethi (1959) considered stochastic methods of rounding off which leave the estimator unbiassed. Later (Murthy and Sethi, 1961) they gave an iterative procedure for determining the optimum randomised rounded off weights in the sense of minimising the number of different rounded off weights

when an upper bound is set for the additional sampling variance due to such rounding off. This procedure, however, is rather cumbersome for adoption in large-scale surveys.

In this paper, we present two alternative practical procedures of randomised rounding off that are amenable to mechanised processing. We round off integers, not the original weights, but the ratios they bear to the smallest weight. The rounding off is done, not in the usual arithmetical way, but in a stochastic way, so as to ensure that the estimates of totals remain unbiassed. We are thus left to work with small integral modified weights. The processing can then be done on the original cards, using the progressive digital total method on a tabulator; and for enumeration surveys, by repeated passes on a counting sorter. Of course, the simplicity of processing is achieved at the cost of some increase in the sampling variance of the estimate, but empirical results suggest that this increase in variance is more than offset by the reduction in processing costs.

In Section 2 we discuss the statistical aspects of the problem and work out an expression for the additional variance. In Section 3 methods are given for computing rounded off weights using punched card equipment. Section 4 gives simplified methods of tabulation using rounded off multipliers. Section 5 gives some empirical results about the magnitude of the additional sampling variance.

2. RANDOMISED ROUNDED OFF WEIGHTS

Let $T = \sum_{i=1}^{n} w_i y_i$ be an unbiassed estimate of a parameter θ, where y_i is a numerical characteristic of the i-th s.u., $i = 1, 2, \ldots, n$ and w_i the positive 'weight' associated with it : the weights being different for different s.u.'s.

Let $w_0 = \min(w_1, w_2, \ldots, w_n)$ and write w_i in the form

$$w_i = w_0(q_i + p_i)$$

where q_i is the largest integer not exceeding w_i/w_0 and $0 \leqslant p_i < 1$. Consider random variables $m_1, m_2, \ldots, m_n$ whose marginal distributions, given the sample, are defined by

$$\begin{aligned} \text{prob}\,(m_i = q_i) &= 1 - p_i \\ \text{prob}\,(m_i = q_i + 1) &= p_i. \end{aligned} \qquad \ldots \quad (2.1)$$

We propose to use the randomised rounded off weight $w_0 m_i$ in place of w_i, so that the statistic

$$t = w_0 \sum_{i=1}^{n} m_i y_i$$

can be used, instead of T, as an estimate of θ. It follows that

$$E^*(t) = T$$

where E^* denotes conditional expectation, given the sample. Consequently t also is an unbiassed estimate of θ. The variance of t is obtained as

$$V(t) = V(T) + E(w_0^2 \sum_{i,j=1}^{n} v_{ij} y_i y_j) \qquad \ldots \quad (2.2)$$

where $((v_{ij}))$ is the conditional dispersion matrix of the random vector $m = (m_1, m_2, \ldots, m_n)$. The second term on the right hand side of (2.2) represents the additional variance due to the use of randomised rounded off weights. Obviously

$$v_{ii} = p_i(1-p_i), \quad i = 1, 2, \ldots, n$$

but the values of v_{ij} for $i \neq j$ will depend on the bivariate marginal distributions of the random variables $m_1, m_2, \ldots, m_n$.

The random vector m can be generated in many different ways. In this paper we shall consider two particular methods, to be called (i) independent and (ii) systematic methods of generation which can be readily mechanised. In method (i), given the sample, the random variables $m_1, m_2, \ldots, m_n$ are generated independently, so that

$$v_{ij} = 0 \quad \text{for} \quad i \neq j.$$

This can be done by taking independent random variables $x_1, x_2, \ldots, x_n$ each distributed uniformly in (0, 1) and defining m_i as the largest integer not exceeding $p_i+q_i+x_i$. Hence, if we denote by t_1 the estimate obtained with this set of randomised rounded off weights, we get

$$V(t_1) = V(T)+E[w_0^2 \sum_{i=1}^{n} y_i^2 p_i(1-p_i)]. \qquad \ldots \quad (2.3)$$

In method (ii), the random vector m is generated in the following way.

Let

$$W_1 = w_1, \; W_2 = w_1+w_2, \ldots, \; W_n = w_1+w_2+\ldots+w_n.$$

We write

$$W_i = w_0(Q_i+P_i)$$

where Q_i is the largest integer not exceeding W_i/w_0 and $0 \leqslant P_i < 1$. Let x be a random variable distributed uniformly in (0, 1) and independently of the sample selected. Let M_i be the largest integer not exceeding $\left(\frac{W_i}{w_0} + x\right)$. We then define

$$\begin{aligned} m_1 &= M_1 \\ m_i &= M_i - M_{i-1}, \quad i = 2, 3, \ldots, n. \end{aligned} \qquad \ldots \quad (2.4)$$

The random vector $m = (m_1, m_2, \ldots, m_n)$ so defined is immediately seen to satisfy the condition (2.1). A little calculation shows that

$$v_{ij} = \Delta_{ij} - \delta_i\delta_j, \quad i \neq j = 1, 2, \ldots, n \qquad \ldots \quad (2.5)$$

where

$$\begin{aligned} \delta_i &= P_i - P_{i-1} \\ \Delta_{ij} &= P_{i,j} - P_{i,j-1} - P_{i-1,j} + P_{i-1,j-1} \end{aligned}$$

P_{ij} being defined as the smaller of P_i and P_j and for the purpose of these definitions $P_0 = 0$. If in particular the groups are so arranged that $P_1 \leqslant P_2 \leqslant \ldots \leqslant P_n$, then $\delta_i = p_i$ and $\Delta_{ij} = 0$, so that $v_{ij} = -p_ip_j$ for $i \neq j = 1, 2, \ldots, n$. If we denote by t_2, the estimate obtained with weights rounded off this way, we get

$$V(t_2) = V(T)+E\Big[w_0^2\Big\{\sum_{i=1}^{n} y_i^2 p_i-\Big(\sum_{i=1}^{n} y_i p_i\Big)^2\Big\}\Big], \qquad \ldots \quad (2.6)$$

Thus from (2.3) and (2.6)

$$V(t_1)-V(t_2) = E\Big[\ w_0^2 \sum_{i\neq j=1}^{n}\Sigma\ y_i y_j p_i p_j\Big]$$

which must be non-negative if $y_i \geqslant 0$ for $i = 1, 2, \ldots, n$. It follows therefore that method (ii) of randomised rounding off is better than method (i) when the groups are arranged in ascending order of the values of their P_i's. In practice, however, it is possible to get such an arrangement only if $\sum_{i=1}^{n} p_i < 1$.

3. PUNCHED CARD METHOD FOR COMPUTING RANDOMISED ROUNDED OFF WEIGHTS

Standard punched card equipment can be used with advantage for computing randomised rounded off weights, whenever the number of sample units is large. One starts with a deck of punched cards (one card per sample unit) giving the indentification number of the 's.u.' and its 'weight' w.

For rounding off the weights with independent binomial probabilities using method (i), one has to take another deck of punched cards on each one of which a random number distributed uniformly in the interval (0, 1) is punched. The two decks are collated in such a way that each 's.u. card' is preceded by a 'random number card.' This merged pack is sent into a calculating punch. From each 's.u.' card the weight 'w_i' is taken to a counter, divided by the minimum weight w_0 supplied by digit emitter and to the fractional part of the quotient formed, the random number x_i from the corresponding random number card is added. The integral part of this total is the required randomised rounded off weight m_i and it is punched on the corresponding 's.u.' card in an assigned field. The counters are cleared for each card.

For computing the randomised rounded off weights using method (ii) the deck of s.u. cards is sorted in the order in which the units were sampled and is fed into the calculating punch. As the first s.u. card passes through the machine, w_1 is taken into the dividend counter. It is divided by w_0 and the quotient formed in a pair of counters coupled together so that the integral part of the quotient will be formed in the least significant positions of the higher order counter and fractional part in the lower order counter. To this fractional part is added the random number $x\ (0 < x < 1)$ from the digit emitter and the integral part is punched as 'm_1'. The higher order quotient counter is now cleared but not the lower order quotient counter nor the dividend counter. From the second 's.u.' card w_2 is added to the remainder left over in the dividend counter and this net figure is divided by w_0. The integral part of the quotient formed is m_2 and is punched in the assigned columns. The process is continued for the rest of the cards, clearing each time only the higher order counter allotted for quotient.

4. Simplified methods of tabulation with randomised rounded off weights

In most cases, the types of estimates required are obtained as the weighted sum of observed values of characteristics of sampled units belonging to each of many different classificatory groups. Randomised rounding off the weights leaves us with weights which are mostly single digit figures and sometimes two-digit figures which can be split up to two single digit numbers adding to that number. Multiplication by these single digit weights of the observed value and summing them over units belonging to each classificatory group becomes quite simple and the need for calculating punches can altogether be eliminated at this stage.

The weighted sums can be obtained on the tabulator itself by applying the progressive digital method.

One starts with a pack of punched cards, one card for each 's.u.' containing the identification number, classificatory codes, the characteristics, and the randomised rounded off weight m_i. We assume that the rounded off weights are all single digit numbers and hence are punched in a single column field. These cards are sorted in descending order of the rounded weights and fed into the tabulator. Every gap in the sequence between 1 and the maximum rounded weight must be filled by inserting blank cards. The machine is set to add the characteristic values 'y' and controlled on the weight. At every break of control the cumulated sum of y's is transferred to another counter without resetting the first counter. The final total accumulated in the second counter after all the 's.u. cards have passed through the machine is the required weighted sum and it is printed out or summary punched as desired. The weighted sums for different classificatory groups are obtained by distributing the y-field among different sets of counters.

When the characteristics are simply indicator variables, that is when each y takes the value 1 or 0 according as the s.u. has a certain attribute or not, the tabulating machine can be dispensed with. A sort counting machine of the IBM type 101-Electronic Statistical Machine can be used with advantage. The Electronic Statistical Machine (ESM) has a feeding speed of 450 cards per minute which is almost thrice that of a tabulator and in addition has about sixty sets of recode selectors and unit counters for purposes of distribution and counting. The problem now reduces to adding up the rounded off weights of the s.u.'s separately for the different classificatory groups. The cards are fed into the machine and the machine is instructed to count the number of cards belonging to different groups. Cards are repeatedly passed through the machine so that any 's.u.' card having weight r passes through the machine r-times and gets counted r-times. This is done by instructing the ESM to sort the cards according to the weights, and at the same time count the number of cards in each group. At the end of j-th run all the 's.u.' cards which fall in the pockets 1, 2, 3, ..., j are removed and the rest of the cards fed into the machine for the $(j+1)$-th run. If the weights

range from 1 to 9, this will involve 9 runs. In case the maximum weight is less than 9 say m only, then the operation terminates after m runs, and the total in the respective counters are printed out. This process has many built-in checks. The ESM itself is equipped to cross-foot check the totals held in the different counters. Also any mis-sorting which is not systematic can be detected by this repetitive removing of some cards and sorting the result again. For instance, if during the third run we get some cards in pocket number 1 or 2 it indicates missorting at some stage and further work on wrong lines can be stopped forthwith. It should also be noted that though there may be 9 runs of cards through the machine not all the cards are run through the machine nine times. At each stage a part of the pack of cards is removed and only the rest run again.

Our experience is that the number of card passages is likely to be of the order of three times the number of cards in most practical applications. The cost of this additional card passage is more than offset by the higher speed and larger counter capacity of the ESM.

5. Numerical examples

From a list of 54 districts of the State of Uttar Pradesh which give., for each district, the area in square miles and the population according to the 1961 census, a sample of 10 districts was selected one by one with replacement, in such a way that the probability of selecting any district, in a particular draw, was proportional to the area of the district.

An unbiassed estimate of θ, the total population of the State is then obtained as $T = \sum_{i=1}^{10} y_i w_i$ where y_i is the population of the i-th selected district and $w_i = (1/10)(A/a_i)$ where $A = 113453$ sq. miles is the total area of the State and a_i is the area in sq. miles of the i-th selected village.

Table 5.1 gives these statistics and also shows the computation of the randomised rounded off multipliers $m_i^{(1)}$ and $m_i^{(2)}$ by the two methods. The modified estimates are then obtained as

$$t_1 = w_0 \sum_{i=1}^{10} y_i m_i^{(1)} \text{ (by 'independent' rounding off),}$$

and,
$$t_2 = w_0 \sum_{i=1}^{10} y_i m_i^{(2)} \text{ (by 'systematic' rounding off),}$$

where $w_0 = \min(w_1, w_2 \ldots, w_{10}) = 2.918030$.

We thus obtain $T = 7.962 \times 10^7$, $t_1 = 8.120 \times 10^7$, $t_2 = 7.752 \times 10^7$ as estimates of the parameter $\theta = 7.375 \times 10^7$.

TABLE 5.1. GENERATION OF RANDOMISED ROUNDED OFF WEIGHTS BY TWO METHODS

population and area				method (i)			method (ii)		
i	y_i (thousands)	a_i (sq. miles)	$w_i=A/10a_i$	w_i/w_0	x_i	$m_i^{(1)}$	W_i/w_0	M_i	$m_i^{(2)}$
1	1728	1543	7.352754	2.5198	0.5261	3	2.5198	3	3
2	1086	3888	2.918030	1.0000	0.3853	1	3.5198	4	1
3	1616	2236	5.073027	1.7388	0.2679	2	5.2586	5	1
4	1860	1816	6.096346	2.0892	0.7840	2	7.3478	7	2
5	1227	1774	6.395321	2.1917	0.3412	2	9.5395	10	3
6	2372	2087	5.436176	1.8630	0.5685	2	11.4025	11	1
7	1498	2620	4.330267	1.4840	0.1264	1	12.8865	13	2
8	1736	1887	6.012347	2.0604	0.3816	2	14.9469	15	2
9	430	1201	9.446544	3.2373	0.7334	3	18.1842	18	3
10	574	2626	4.320373	1.4806	0.9928	2	19.6648	20	2

$A=113453$ sq.miles ; $w_0=2.918030$; $(W_i=\sum_{x=1}^{i} w_x$; $x=0.5261)$.

An estimate of the variance of T is obtained as

$$\text{est } V(T)=\frac{1}{9}\left[10\sum_{i=1}^{10}(y_iw_i)^2-T^2\right]=1.4983\times10^{14}.$$

An estimate of the additional variance in $t^{(1)}$ due to rounding off the multipliers by method (i) is obtained as

$$w_0^2\sum_{i=1}^{10}y_i^2p_i(1-p_i)=0.2789\times10^{14}$$

where p_i is the fractional part of w_i/w_0. Thus an estimate of the variance of $t^{(1)}$ is

$$\text{est } V(t_1)=\text{est } V(T)+w_0^2\sum_{i=1}^{10}y_i^2p_i(1-p_i)=1.7772\times10^{14}$$

which means that rounding off by method (i) causes an increase of only about 9.3% in the standard error of the estimate.

An estimate of the additional variance in t_2 due to rounding off the multipliers by method (ii) is given by

$$w_0^2\left[\sum_{i,j=1}^{10}\Delta_{ij}y_iy_j-\left(\sum_{i=1}^{10}\delta_iy_i\right)^2\right]$$

$$=0.02585\times10^{14}.$$

Thus $\text{est } V(t_2) = 1.5242 \times 10^{14}$

which means that rounding off by method (ii) causes an increase of only 1.2% in the standard error of the estimate. The details of the computation are shown in Tables 5.2 and 5.3.

TABLE 5.2. VALUES OF $P_{i,j}$ AND P_i

$i\backslash j$	1	2	3	4	5	6	7	8	9	10	P_i
1	0.5198	0.5198	0.2586	0.3478	0.5198	0.4025	0.5198	0.5198	0.1842	0.5198	0.5198
2	0.5198	0.5198	0.2586	0.3478	0.5198	0.4025	0.5198	0.5198	0.1842	0.5198	0.5198
3	0.2586	0.2586	0.2586	0.2586	0.2586	0.2586	0.2586	0.2586	0.1842	0.2586	0.2586
4	0.3478	0.3478	0.2586	0.3478	0.3478	0.3478	0.3478	0.3478	0.1842	0.3478	0.3478
5	0.5198	0.5198	0.2586	0.3478	0.5395	0.4025	0.5395	0.5395	0.1842	0.5395	0.5395
6	0.4025	0.4025	0.2586	0.3478	0.4025	0.4025	0.4025	0.4025	0.1842	0.4025	0.4025
7	0.5198	0.5198	0.2586	0.3478	0.5395	0.4025	0.8865	0.8865	0.1842	0.6648	0.8865
8	0.5198	0.5198	0.2586	0.3478	0.5395	0.4025	0.8865	0.9469	0.1842	0.6648	0.9469
9	0.1842	0.1842	0.1842	0.1842	0.1842	0.1842	0.1842	0.1842	0.1842	0.1842	0.1842
10	0.5198	0.5198	0.2586	0.3478	0.5395	0.4025	0.6648	0.6648	0.1842	0.6648	0.6648
P_j	0.5198	0.5198	0.2586	0.3478	0.5395	0.4025	0.8865	0.9469	0.1842	0.6648	—

TABLE 5.3. VALUES OF Δ_{ij} AND δ_i

$i\backslash j$	1	2	3	4	5	6	7	8	9	10	δ_i
1	0.5198	0.	−0.2612	0.0892	0.1720	−0.1174	0.1174	0.	−0.3357	0.3357	0.5198
2		0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3			0.2612	−0.0892	−0.1720	0.1174	−0.1174	0.	0.2612	−0.2612	−0.2612
4				0.0892	0.	0.	0.	0.	−0.0892	0.0892	0.0892
5					0.1916	−0.1370	0.1370	0.	−0.1916	0.1916	0.1916
6						0.1370	−0.1370	0.	0.1370	−0.1370	−0.1370
7							0.4840	0.	−0.4840	0.2623	0.4840
8								0.0604	−0.0604	0.	0.0604
9									0.7627	−0.4806	−0.7627
10										0.4806	0.4806

REFERENCES

MURTHY, M. N. and SETHI, V. K. (1959) : *Self-weighting Designs at Tabulation Stage*, Indian Statistical Institute : National Sample Survey : Working Paper 6.

——— (1961) : Randomised rounded off multipliers in sampling theory. *J. Am. Stat. Ass.*, **56**, 328-339.

Paper received : October, 1962.
Revised : February, 1963.