The Annals of Statistics
1977, Vol. 5, No. 4, 795-802

REPEATED SAMPLING WITH PARTIAL REPLACEMENT OF UNITS

BY E. MANOUSSAKIS

Hellenic Telecom Organization, Research Dept., Athens, Greece

This paper is concerned with the minimum variance estimation of a time-dependent population mean, assuming that one is restricted to the case of linear unbiased estimators.

A number of results are given for a new rotation sampling model (RSM), in which unequal sample sizes are used on each occasion. Also results corresponding to the special case of sampling with a fixed sample size on all the occasions are derived.

Finally the optimum structure of the suggested model is discussed and a comparison of this sampling scheme with Patterson's and Eckler's schemes is made.

1. Introduction. When changes in time-dependent population values are to be examined several sampling alternatives, as listed by Yates (1960), can be used. These involve: (i) A new sample on each occasion; (ii) a fixed sample used on all occasions; (iii) a subsample of the original sample on a second occasion; (iv) a partial replacement of units from occasion to occasion.

The relative merits of these alternative methods will depend on the extent to which any relationship between the values of a particular character observed, on the same unit of the population on two successive occasions, can be used to improve the current estimate of the population mean. This can be done by using information obtained from samples taken on regular previous occasions.

Although in this paper, as in Patterson's (1950), repeated sampling with partial replacement of units from occasion to occasion is used, a different procedure is followed here in order to obtain an estimate of the true population mean on each occasion (after the first). The repeated sampling or rotation sampling model discussed here will be referred to as RSM in the following.

The existence of an infinite population U is assumed along occasions (time points) $t_1, t_2, \cdots, t_\nu$. It is assumed that $t_j - t_{j-1} =$ constant for $j = 2, 3, \cdots, \nu$. Any unit, U_l say, of the population is associated with a sample value z_{lj} corresponding to the t_jth occasion. Since the population is infinite, the sample values z_{lj}, z_{mj} are uncorrelated. The sample values z_{lj} on the t_jth occasion are regarded as values of a random variable Z_j with expected value equal to the true population mean, $\bar{Y}_j$ say, and variance σ^2 independent of time. Also for the random variables Z_j ($j = 1, 2, \cdots$) it is assumed the correlation coefficient $\rho(Z_j, Z_{j-1}) = \rho\ \forall j > 1$ and the partial correlation coefficients $\rho_{jj'.s}$ are zero

Received November 1975; revised December 1976.

AMS 1970 *subject classification.* Primary 62D05.

Key words and phrases. Random sampling, repeated sampling, rotation sampling, partial correlation coefficient, minimum variance linear unbiased estimator.

$\forall s$ between j, j'. These last properties lead to the result:

$$\rho(Z_j, Z_{j'}) = \rho^{|j-j'|}, \quad \forall j, j' .$$

Now the sampling procedure suggested here is as follows: On each occasion t_j, after the first, a random sample s_{1j} of size n_{1j} is selected from the units appearing on the previous occasion t_{j-1}, and the values of a particular character observed on these units on both occasions t_j, t_{j-1} are recorded. The means of these sample values are denoted by $\bar{x}_{1j}$, $\bar{y}_{1j}$, respectively. (The sample values for $\bar{y}_{1j}$ are already available from the previous occasion.) On the same occasion t_j, another random sample s_{2j} of size n_{2j}, consisting of entirely new units (not appearing in any previous sample), is chosen. The sample values of these units are recorded for both occasions t_j, t_{j-1} and their means are denoted by $\bar{x}_{2j}$, $\bar{y}_{2j}$, respectively. (It is usually cheaper to record these sample values z_{lj}, $z_{l,j-1}$, if possible, on the same occasion rather than to record them on two separate occasions.)

The sample sizes n_{1j}, n_{2j} are connected by the relations $n_{1j} + n_{2j} = n_j$, $n_{ij} = u_{ij} n_j$ $\forall j$ and for $i = 1, 2$, where $u_{ij} \in [0, 1]$.

We notice that Patterson (1950) did not involve the sample mean $\bar{y}_{2j}$ in his sampling scheme, which was called one-level sampling by Eckler. Furthermore Eckler (1955) did not use the sample means $\bar{x}_{1j}$, $\bar{y}_{1j}$ in his two-level sampling. The main results of Patterson's and Eckler's papers are summarized in the following section for ready reference.

2. One-level and two-level rotation sampling.

2.1. The minimum variance linear unbiased estimator $\bar{y}_j^{(P)}$ of the true population mean $\bar{Y}_j$, on the t_jth occasion, used by Patterson, is of the form:

$$\bar{y}_j^{(P)} = (1 - \varphi_j)\bar{x}_{1j} + \varphi_j \bar{x}_{2j} + A_j(\bar{y}_{j-1}^{(P)} - \bar{y}_{1j}) , \quad \forall j > 1 .$$

Although this expression differs slightly from that given by Patterson, it leads to exactly the same results as follows:

$$\bar{y}_j^{(P)} = (1 - \varphi_j)\{\bar{x}_{1j} + \rho(\bar{y}_{j-1}^{(P)} - \bar{y}_{1j})\} + \varphi_j \bar{x}_{2j} , \quad \forall j > 1 ,$$

$$(2.1) \qquad V_j^{(P)} \equiv \operatorname{Var}(\bar{y}_j^{(P)}) = (1 - A_j/\rho)(\sigma^2/n_{2j}) , \quad \forall j \geqq 1 ,$$

$$\varphi_j = 1 - A_j/\rho$$

where

$$A_j = \rho n_{1j} n_{2,j-1}/\{n_j n_{2,j-1} - n_{2j}(n_{2,j-1} - n_{1j})\rho^2 - n_{1j} n_{2j} \rho A_{j-1}\}$$

with $A_1 = \rho u_{11}$ (from the initial condition $V_1^{(P)} = \sigma^2/n_1$).

In the case of sampling with fixed sample sizes $n_{ij} = u_i n$, $\forall i, j$, the expression for A_j becomes:

$$(2.2) \qquad A_j = \rho u_1/(1 - \rho^2 + 2\rho^2 u_1 - \rho u_1 A_{j-1}) , \quad \forall j > 1 .$$

The limiting value A of A_j, for $j \to \infty$, is

$$A = [1 - \rho^2 + 2\rho^2 u_1 - \{(1 - \rho^2)(1 - \rho^2 + 4\rho^2 u_1 u_2)\}^{\frac{1}{2}}]/(2\rho u_1) .$$

REMARK. It may often be more convenient to use $\varphi = \lim_{j\to\infty} \varphi_j$ in place of φ_j after any occasion t_i $(i \geqq 1)$. The previous results are then simplified with little loss in efficiency. The true variance of the corresponding estimator of $\bar{Y}_j$ may be denoted by $k_j(\sigma^2/n_2)$, where

$$k_j = \varphi(1 - A^2) + A^2 k_{j-1}, \qquad \forall j > i, \quad (k_i = \varphi_i). \tag{2.3}$$

This is an alternative, simpler expression than that given by Patterson (1950), which can be proved as follows.

Patterson proved that

$$u_1 k_j = (1 - \varphi^2)\{u_2(1 - \rho^2) + u_1\rho^2 k_{j-1}\} + u_1\varphi^2, \quad \forall j > i.$$

Since $A = \rho(1 - \varphi)$ this equation can be written as:

$$k_j = S + A^2 k_{j-1}$$

where

$$S = \varphi^2 + (1 - \rho^2)(1 - \varphi)^2 u_2/u_1.$$

It is obvious now that it is sufficient to prove that $S = \varphi(1 - A^2)$. Actually by eliminating φ from S, after some algebraic manipulation, we have $\rho^2 u_1 S = u_1\rho(\rho - 2A) + [A(1 - u_2\rho^2)]A$.

By taking limits of both sides of equation (2.2) we can obtain the following results: $A(1 - u_2\rho^2) = \rho u_1(A^2 - \rho A + 1)$. Because of this result, $\rho^2 u_1 S$ can be written as $\rho u_1(\rho - A)(1 - A^2)$ which eventually leads to the desired result $S = \varphi(1 - A^2)$, since $\varphi = (\rho - A)/\rho$.

2.2. Eckler (1955) used the following minimum variance linear unbiased estimator $\bar{y}_j^{(E)}$ of the true population mean $\bar{Y}_j$, on the t_jth occasion:

$$\bar{y}_j^{(E)} = \bar{x}_{2j} + A_j'(\bar{y}_{j-1}^{(E)} - \bar{y}_{2j}), \quad \forall j > 1,$$

whence he found:

$$V_j^{(E)} \equiv \operatorname{Var}(\bar{y}_j^{(E)}) = (1 - \rho A_j')(\sigma^2/n_{2j}), \quad \forall j \geqq 1, \tag{2.4}$$

where

$$A_j' = \rho n_{2,j-1}/(n_{2j} + n_{2,j-1} - \rho n_{2j} A_{j-1}'), \quad \forall j > 1,$$

with $A_1' = 0$ (from the initial condition $V_1^{(E)} = \sigma^2/n_{21}$).

The previous results are simplified in the case of sampling with fixed sample size $n_{2j} = n_2$, $\forall j$. If $V^{(E)} = \lim_{j\to\infty} V_j^{(E)}$ then it can easily be proved that $V^{(E)} = (1 - \rho^2)^{\frac{1}{2}}(\sigma^2/n_2)$.

A further simplification may be possible by substituting $A' = \lim_{j\to\infty} A_j'$ for A_j' in the estimator $\bar{y}_j^{(E)}$ after any occasion t_i $(i \geqq 1)$. The true variance of the corresponding estimator of $\bar{Y}_j$ may be denoted by $(1 - \rho k_j')/(\sigma^2/n_2)$ where $k_j' = A'(1 - A'^2) + A'^2 k_{j-1}'$, $\forall j > i$ with the initial condition $k_i' = A_i'$.

3. The new rotation sampling model. We denote by $\bar{y}_j$ the minimum variance linear unbiased estimator of the true population mean $\bar{Y}_j$ on the occasion t_j and define it as follows:

$$\bar{y}_j = (1 - w_j)\bar{x}_{1j} + w_j\bar{x}_{2j} + a_j(\bar{y}_{j-1} - \bar{y}_{2j}) + b_j(\bar{y}_{1j} - \bar{y}_{2j}), \quad \forall j > 1.$$

By minimizing the Var $(\bar{y}_j) \equiv V_j$ this equation becomes

$$\bar{y}_j = u_{1j}\bar{x}_{1j} + u_{2j}\bar{x}_{2j} - \rho u_{1j}(\bar{y}_{1j} - \bar{y}_{2j}) + a_j(\bar{y}_{j-1} - \bar{y}_{2j}) \tag{3.1}$$

where

$$a_j = \rho\sigma^2/(\sigma^2 + n_{2j}V_{j-1}) .$$

The variance V_j can be expressed as:

$$V_j = \{1 - (1-\rho^2)u_{1j} - \rho a_j\}(\sigma^2/n_{2j}) , \quad \forall j \geqq 1 , \tag{3.2}$$

whence we can find the following recurrence formula for a_j:

$$a_j = \rho n_{j-1}n_{2,j-1}/\{(n_{2j} + n_{2,j-1})n_{j-1} - (1-\rho^2)n_{2j}n_{1,j-1} - \rho n_{2j}n_{j-1}a_{j-1}\} , \quad \forall j > 1 , \tag{3.3}$$

with $a_1 = \rho u_{11}$ (because of the initial condition $V_1 = \sigma^2/n_1$).

We notice here that, among the variances $V_j^{(P)}$, $V_j^{(E)}$, V_j of the three sampling schemes described in the present paper, the following inequalities hold:

$$V_j^{(E)}(n_j) \leqq V_j(n_j, u_{1j}) \leqq V_j^{(P)}(n_j, u_{1j}) .$$

These results give a comparison of the three models without considering cost.

4. Sampling with fixed sampling sizes. Here we consider the case of sampling with fixed sample sizes on each occasion. By using the restrictions $n_j = n$, $n_{ij} = u_i n$ for $i = 1, 2$ and $\forall j$ the main results (3.3), (3.2) are simplified, respectively, as follows:

$$a_j = \rho/\{2 - (1-\rho^2)u_1 - \rho a_{j-1}\} , \quad \forall j > 1 . \tag{4.1}$$

$$V_j = \{1 - (1-\rho^2)u_1 - \rho a_j\}\{\sigma^2/(nu_2)\} , \quad \forall j \geqq 1 . \tag{4.2}$$

It can be shown that the sequence $\{a_j\}$ is monotone ($0 \leqq a_j \nearrow < 1$ for $\rho > 0$ and $-1 < a_j \swarrow \leqq 0$ for $\rho \leqq 0$) and it converges to the limiting value α, where $\alpha = [1 + u_2 + \rho a_1 - \{(1 + u_2 + \rho u_1)^2 - 4\rho^2\}^{\frac{1}{2}}]/(2\rho)$.

Now, the quantities V_j, a_j can alternatively be written, in terms of the limiting value α, as follows:

$$V_j = \{\rho(\alpha - a_j) + (\rho - a)/a\}\{\sigma^2/(nu_2)\} , \tag{4.3}$$

$$a_j = \alpha(1 - xa^{2(j-1)})/(1 - xa^{2j}) , \quad \forall j \geqq 1 , \tag{4.4}$$

where

$$x = (\alpha - a_1)/(\alpha - a^2 a_1) .$$

To prove the result (4.3) we notice that from equation (4.1), for $j \to \infty$, we have $\alpha = \rho/\{2 - (1-\rho^2)u_1 - \rho a\}$ whence we find $1 - (1-\rho^2)u_1 = \rho a + (\rho - a)/\alpha$. Using this result in equation (4.2) we obtain (4.3) directly.

Some details are now given to indicate how the result (4.4) was obtained. It is obvious that (4.1) can be written in the form:

$$a_j = 1/(D - Ca_{j-1}) , \quad \forall j > 1 .$$

Now we use the transformation $a_j = (D - v_{j+1}/v_j)/C$, $\forall j \geqq 1$, (where v_1 ($\neq 0$) is arbitrarily chosen) whence we obtain the following homogeneous difference equation:

$$v_{j+1} - Dv_j + Cv_{j-1} = 0 . \tag{4.5}$$

By solving this equation we finally obtain (4.4).

REMARK. It is possible, after any occasion t_i ($i \geqq 1$), to use α in place of a_j in the estimator $\bar{y}_j$. The true variance, say $\hat{V}_j$, of the new estimator of $\bar{Y}_j$ may be denoted by

$$\hat{V}_j = \{\rho(a - k_j) + (\rho - a)/\alpha\}\{\sigma^2/(nu_2)\} , \quad \forall j > i$$

where

$$k_j = a(1 - a^2) + a^2 k_{j-1} , \qquad \text{with} \quad k_i = \alpha_i . \tag{4.6}$$

Of course this replacement simplifies the calculations for the estimate of $\bar{Y}_j$ though it increases the true variance of the corresponding estimator.

The loss in efficiency, due to the incorrect weights, is minimal as we can see by examining the fractional error $B_j = (\hat{V}_j - V_j)/V_j$, $\forall j > i$.

We notice that if $i \geqq 3$ then $\max B_j < 4.4\%$ $\forall\, |\rho| \in [0, 1]$ and $u_1 \in [0.10, 1]$. For $i = 2$, $\max B_j \doteq 17\%$ for $\rho = 0.95$ and $u_1 = 0.10$. The sequence $\{B_j\}$ decreases with ρ and increases with u_1.

5. Optimum number of units on each occasion for RSM. Some consideration will now be given to the question "How many units should be used on a given occasion and how many of these should be new?"

To provide an answer to this question we proceed as follows. Suppose the results up to the occasion t_{j-1} are known and we wish to estimate $\bar{Y}_j$ such that $V_j = \sigma^2/N$, $\forall j \geqq 1$, with a minimum sampling cost on the occasion t_j.

If we assume that it costs c to obtain a single sample value on the occasion t_j, and $c(1 + k)$ to obtain two sample values z_{ij}, $z_{i,j-1}$ on the same occasion t_j, where $0 \leqq k \leqq 1$, then minimization of the cost function $M_j = c\{n_{1j} + (1 + k)n_{2j}\}$ leads to the following results, $\forall j > 1$ (on the first occasion N units are chosen). These results designate the corresponding sampling procedure.

(i) If $0 \leqq k \leqq (1 - (1 - \rho^2)^{\frac{1}{2}})^2/\rho^2$ then $n_j = n_{2j} = N(1 - \rho^2)^{\frac{1}{2}}$ and $a_j = (1 - (1 - \rho^2)^{\frac{1}{2}}/\rho^2$. That is to say, in this case, the RSM degenerates to Eckler's model.

(ii) If

$$(1 - (1 - \rho^2)^{\frac{1}{2}})^2/\rho^2 < k < \min\{1, \rho^2/(1 - \rho^2)\}$$

then

$$n_j = N(1 - \rho^2)^{\frac{1}{2}}\{|\rho|k^{\frac{1}{2}} + (1 - \rho^2)^{\frac{1}{2}}\} ,$$
$$n_{2j} = N(1 - \rho^2)^{\frac{1}{2}}\{|\rho|/k^{\frac{1}{2}} - (1 - \rho^2)^{\frac{1}{2}}\} ,$$
$$a_j = (\rho/|\rho|)k^{\frac{1}{2}}(|\rho|k^{\frac{1}{2}} + (1 - \rho^2)^{\frac{1}{2}})^{-1} .$$

(iii) If $\rho^2/(1 - \rho^2) \leqq k \leqq 1$ (for $\rho^2 \leqq 0.50$), then $n_j = N$, $n_{2j} = 0$, $a_j = \rho$.

In this case the estimator $\bar{y}_j$ coincides with the sample mean $\bar{x}_{1j}$, $\forall j \geqq 1$, as can be seen from equation (3.1). Choosing $\bar{y}_1$ properly (for example $\bar{y}_1 = \bar{Y}_1$, if possible) this coincidence can be avoided. In Table 1 the values of the ratios $P = n_j/N$, $P_2 = n_{2j}/N$ are recorded for various values of ρ, k.

TABLE 1
Percentage optimal values of $P = n_j/N$, $P_2 = n_{2j}/N$ for some values of ρ, k

100ρ	100k: 100	95	90	85	80	75	70	65	60	55	50	45	40	35	30	25	20	15
95	39	39	38	37	36	35	34	34	33	32	$(P = P_2 = (1-\rho^2)^{\frac{1}{2}})$							
	20	21	21	22	23	24	26	27	28	30								
90	58	57	56	55	54	53	52	51	49	48	47	45	44					
	20	21	22	23	25	26	28	30	32	34	36	39	43					
85	72	71	70	69	68	66	65	64	62	61	59	58	56	54				
	17	18	19	21	22	24	26	28	30	33	35	39	43	48				
80	84	83	81	80	79	77	76	75	73	72	70	68	66	64	62			
	12	13	15	16	18	19	21	23	26	29	32	35	40	45	52			
75	93	92	91	89	88	87	85	84	82	80	79	77	75	73	71	68		
	6	7	8	10	12	13	15	18	20	23	26	30	35	40	47	55		
70			98	97	96	94	93	91	90	88	86	84	83	80	78	76	73	
	$\begin{pmatrix} P = 100 \\ P_2 = 0 \end{pmatrix}$		2	3	5	7	9	11	13	16	20	23	28	33	40	49	61	
65							99	97	96	94	93	91	89	87	85	82	80	77
							1	3	6	9	12	16	20	26	32	41	53	70

NOTE. The first entry in each cell (ρ, k) corresponds to P and the second to P_2.

6. Comparison of the three models considering cost. In this section we derive a criterion for deciding when to use one sampling method in preference to the other two. In order to simplify the results we assume that we have patterns of infinite length. As in the previous Section 5, we assume that it costs c to obtain a single sample value for each occasion t_j and $c(1 + k)$ to obtain two sample values z_{lj}, $z_{l,j-1}$ on the same occasion t_j, where $k \in [0, 1]$. If we suppose that we have a fixed amount $M = cN$ to spend for sampling for each occasion t_j then the sample sizes $n = N(1 + ku_2)^{-1}$, $n_P = N$, $n_E = N(1 + k)^{-1}$ can be used, on all the occasions, for RSM, one-level or two-level sampling models respectively. It is also supposed that the same constant fractional replacement rate u_2 is used for both the RSM and one-level sampling on all the occasions.

Now substituting the previous sample sizes in the limiting values V, $V^{(P)}$, $V^{(E)}$ of the variances V_j (equation (4.3)), $V_j^{(P)}$ (equation (2.1)), $V_j^{(E)}$ (equation (2.4)), respectively, we find

$$V = \{(\rho - a)(1 + ku_2)/(au_2)\}(\sigma^2/N) ,$$
$$V^{(P)} = \{(\rho - A)/(\rho u_2)\}(\sigma^2/N) ,$$
$$V^{(E)} = \{(1 + k)(1 - \rho^2)^{\frac{1}{2}}\}(\sigma^2/N) .$$

Comparing the above variances in pairs, for various values of k, we obtain the desired criterion.

6.1. RSM *versus one-level rotation sampling*. Denoting the solution of the equation $V = V^{(P)}$ by k_P we find

$$k_P = \{\alpha(\rho - A) - \rho(\rho - a)\}/\{\rho u_2(\rho - a)\}$$

where k_P is a function of the parameters ρ, u_2. Comparing k with k_P we can decide which model should be used. The required criterion is as follows. For any $k > k_P$ use one-level sampling and for any $k < k_P$ use the RSM. If $k = k_P$ then it is immaterial which scheme is used.

6.2. RSM *versus two-level rotation sampling*. Similarly on requiring $V = V^{(E)}$ we find

$$k_E = \{(\rho - a) - au_2(1 - \rho^2)^{\frac{1}{2}}\}/\{au_2(1 - \rho^2)^{\frac{1}{2}} - u_2(\rho - a)\}$$

where k_E denotes the solution of the previous equation. Now we have the following criterion: if $k > k_E$, use RSM and if $k < k_E$, use two-level sampling. If $k = k_E$ it is immaterial which sampling scheme is used.

We notice that solving equation $V = V^{(E)}$ in terms of u_2 we find the solution

$$u_2^* = \{1 - (1 + k)(1 - \rho^2)^{\frac{1}{2}}\}/\{k^2 + k(1 + k)(1 - \rho^2)^{\frac{1}{2}}\}$$

which leads to the following criterion. For any $u_2 > u_2^*$ use the RSM and for any $u_2 < u_2^*$ use two-level sampling. If $u_2 = u_2^*$ it does not matter which sampling scheme is used.

The previous expression of u_2^* is held provided $(1 - (1 - \rho^2)^{\frac{1}{2}})^2/\rho^2 \leqq k \leqq (1 - (1 - \rho^2)^{\frac{1}{2}})/(1 - \rho^2)^{\frac{1}{2}}$ since $u_2^* \in [0, 1]$. The criterion under consideration is completed as follows. For any $k < (1 - (1 - \rho^2)^{\frac{1}{2}})^2/\rho^2$, use two-level sampling and for any $k > (1 - (1 - \rho^2)^{\frac{1}{2}})/(1 - \rho^2)^{\frac{1}{2}}$, use the RSM, no matter what value is chosen for u_2.

6.3. *One-level versus two-level rotation sampling*. Eckler (1955) compared one-level with two-level rotation sampling for the particular value 0.50 of the parameter u_2 for which $V^{(P)}$ becomes a minimum. In the following a comparison is made between the two sampling schemes for any value of the parameter u_2 $(= 1 - u_1)$. Equating the variances $V^{(P)}$, $V^{(E)}$ and denoting the solution of this equation by k_E' we find

$$k_E' = \{(1 - \rho^2 + 4\rho^2 u_1 u_2)^{\frac{1}{2}} - (1 - \rho^2)^{\frac{1}{2}} - 2\rho^2 u_1 u_2\}/(2\rho^2 u_1 u_2)\,.$$

Now, for any pair of values (ρ, u_2) the desirable criterion is as follows. If $k > k_E'$, use one-level rotation sampling, otherwise use the two-level scheme.

Acknowledgments. I would like to thank Professor Sampford who introduced me to rotation sampling and for his helpful advice and encouragement.

REFERENCES

COCHRAN, W. G. (1963). *Sampling Techniques*, 2nd ed. Wiley, New York.

ECKLER, A. R. (1955). Rotation sampling. *Ann. Math. Statist.* **26** 664-685.

PATTERSON, H. D. (1950). Sampling on successive occasions with replacement of units. *J. Roy. Statist. Soc.* **B12** 241-255.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, 3rd ed. Griffin, London.

38, G. PAPANDREOU AVENUE
ATHENS 622, GREECE