A NOTE ON SAMPLING OVER TWO OCCASIONS

BY J. N. K. RAO AND HISAKO SHIMIZUGAWA
Texas A and M. University and Otaru University of Commerce
(Received in March, 1968)

1. INTRODUCTION

Suppose we have a finite population of N units and a character whose values can be measured on two different occasions for each unit in the population. Let x_j and y_j denote the values of the character for the j^{th} unit on the first and second occasions respectively. Let $\bar{X}$ and $\bar{Y}$ denote the population means, S_x^2 and S_y^2 the population mean squares and ρ the population correlation. Select on the first occasion a simple random sample s of n units and observe their x-values. On the second occasion, select a random subsample s_1 of m units from s and a random sample s_2 of u units from the $(N-n)$ units in the population not included in s, and observe their y-values.

Kulldorff (1963) has shown that the best linear unbiased estimator of $\bar{Y}$ is given by

$$\bar{y}_r' = \frac{n_r'[\bar{y}_1+\beta(\bar{x}-\bar{x}_1)]+u\bar{y}_2}{n_r'+u} \qquad \ldots(1)$$

where the population regression coefficient β is assumed to be known,

$$\bar{y}_1=m^{-1}\sum_{s_1} y_j,\ \bar{x}_1=m^{-1}\sum_{s_1} x_j,\ \bar{y}_2=u^{-1}\sum_{s_2} y_j,\ \bar{x}=n^{-1}\sum_{s} x_j$$

and n_r' is defined by

$$\frac{1}{n_r'}=\frac{\rho^2}{n}+\frac{1-\rho^2}{m} \qquad \ldots(2)$$

and does not contain u. The variance of $\bar{y}_r'$ is

$$V(\bar{y}_r')=\left(\frac{1}{n_r'+u}-\frac{1}{N}\right)S_y^2. \qquad \ldots(3)$$

Kulldorff investigated the optimum allocation of m and u assuming the simple cost function

$$C=c_0+c_1m+c_2u \qquad \ldots(4)$$

and tabulated the optimum values of m/n and u/n for the important case $R=(C-c_0)/c_2=n$ and for selected values of the cost ratio $\delta=c_1/c_2$ and the correlation ρ, where c_0 is the fixed cost, c_1 is the cost per matched unit and c_2 is the cost per unmatched unit. If β is not known (which often is the case in paractice), the sample regression coefficient b computed from the values of the matched units is used in place of β, provided m is sufficiently large.

Des Raj (1965*a*) used the difference estimator

$$\bar{y}_d'=w_d[\bar{y}_1+(\bar{x}-\bar{x}_1)]+(1-w_d)\bar{y}_2 \qquad ...(5)$$

where w_d is a constant weight, $0 \leqslant w_d \leqslant 1$. He investigated the relative efficiency of $\bar{y}_d'$ over $\bar{y}_r'$ for the special case $S_x^2=S_y^2$, $\delta=1$, $m+u=n$ and N infinite. Des Raj (1965*b*) extended the results to unequal probability sampling with replacement. Hansen *et al* (1953) used the composite estimator

$$\bar{y}_c'=w_c(\bar{y}_1+\bar{x}-\bar{x}_1)+(1-w_c)\bar{y} \qquad ...(6)$$

where

$$\bar{y}=(m+u)^{-1}\sum_{s_1+s_2} y_j$$

and w_c is a constant weight, $0 \leqslant w_c \leqslant 1$. The difference and composite estimators, unlike the regression estimator when b is used in place of β, are unbiased and computationally simpler. Moreover, unbiased variance estimators of $\bar{y}_d'$ and $\bar{y}_c'$ are available for all m.

In this note we investigate the relative efficiencies of $\bar{y}_d'$ and $\bar{y}_c'$ for the general case, along the lines of Kulldorff. We also provide a table of optimum values of m/n, u/n and the optimum weights w_d and w_c.

2. Difference Estimator

It is easily seen, by using conditional expectations, that the variance of $\bar{y}_d'$ is given by

$$V(\bar{y}_d')=\left[w_d^2\left\{\frac{1}{m}+\left(\frac{1}{m}-\frac{1}{n}\right)(k^2-2\rho k)\right\}+\frac{(1-w_d)^2}{u}-\frac{1}{N}\right]S_y^2 \qquad ...(7)$$

where $k=S_x/S_y$. Minimization of (7) with respect to w_d leads to

$$w_{d(opt)}=\frac{n_d'}{n_d'+u} \qquad ...(8)$$

and

$$V_{opt}(\bar{y}_d')=\frac{1}{n_d'+u}-\left\{\frac{1}{N}\right\}S_y^2 \qquad ...(9)$$

TABLE 1

Optimum values of m/n, u/n and w for $\bar{y}_d'$ (and in parantheses for $\bar{y}_e'$) when the total cost is fixed to c_0+c_2n (i.e., $R=n$) and $S_x/S_y=k=1$.

ρ	$=\frac{c_m}{c_u}$	0·1	0·2	0·3	0 4	0·5	0·6	0·7	0·8	0·9	1·0	1·5	2 0	2·5	3·0
0·6	m/n	1·000	1·000	1·000	1·000	1·000	1·000	1·000	1·000	0·714	0·472	0·000	0·000	0·000	0·000
		(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(0·573)	(0·390)	(0·000)	(0·000)	(0·000)	(0·000)
	u/n	0·900	0·800	0·700	0·600	0·500	0·400	0·300	0·200	0·357	0·528	1·000	1·000	1·000	1·000
		(0·900)	(0·800)	(0·700)	(0·600)	(0·500)	(0·400)	(0·300)	(0·200)	(0·487)	(0·610)	(1·000)	(1·000)	(1·000)	(1·000)
	w	0·526	0·556	0·588	0·625	0·667	0·714	0·769	0·833	0·679	0·500	0·000	0·000	0 000	0·000
		(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·213)	(0·185)	(0·000)	(0·000)	(0·000)	(0·000)
0·7	m/n	1·000	1·000	1·000	1·000	1·000	1·000	0 815	0·665	0·541	0·436	0·081	0·000	0·000	0·000
		(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(0·610)	(0·450)	(0·350)	(0·090)	(0·000)	(0·000)	(0·000)
	u/n	0·900	0·800	0·700	0·600	0·500	0·400	0·430	0·468	0·513	0·564	0·878	1·000	1·000	1·000
		(0·900)	(0·800)	(0·700)	(0·600)	(0·500)	(0·400)	(0·300)	(0·512)	(0·595)	(0·650)	(0·865)	(1·000)	(1·000)	(1·000)
	w	0·526	0·556	0·588	0·625	0·667	0·714	0·672	0·621	0 564	0·500	0·127	0·000	0·000	0·000
		(0·000)	(0·000)	(0·000)	(0·000)	(0 000)	(0 000)	(0·000)	(0·249)	(0·243)	(0·225)	(0·091)	(0·000)	(0·000)	(0·000)
0·8	m/n	1·000	1·000	1·000	1·000	0·824	0·694	0·593	0·512	0·446	0·387	0·194	0·079	0·000	0;000
		(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(0·430)	(0·360)	(0·310)	(0·150)	(0·070)	(0;010)	(0·000)
	u/n	0·900	0·800	0·700	0·600	0·588	0·584	0·585	0·591	0·600	0·613	0·709	0·843	1·000	1·000
		(0·900)	(0·800)	(0·700)	(0·600)	(0·500)	(0·400)	(0·300)	(0·656)	(0 676)	(0·690)	(0·775)	(0·860)	(0·975)	(1·000)
	w	0·526	0·556	0·588	0·625	0·610	0·593	0·573	0·551	0·526	0·500	0 346	0·173	0·000	0 000
		(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·290)	(0·285)	(0·277)	(0·201)	(0·117)	(0·020)	(0·000)
0·9	m/n	1·000	1·000	0·771	0·634	0·541	0·472	0·418	0·375	0·339	0 309	0·206	0·145	0·104	0·073
		(1·000)	(1·000)	(1·000)	(1·000)	(1·000)	(0·380)	(0·330)	(0·290)	(0·260)	(0·240)	(0·160)	(0·120)	(0·080)	(0·060)
	u/n	0·900	0·800	0·769	0·746	0·730	0·717	0·707	0·700	0·695	0·691	0·690	0·709	0·741	0·782
		(0·900)	(0·800)	(0·700)	(0·600)	(0·500)	(0·772)	(0·764)	(0·768)	(0·766)	(0·760)	(0·760)	(0·760)	(0·800)	(0·820)
	w	0·526	0·556	0·551	0·546	0·539	0·533	0·525	0·517	0·509	0·500	0·450	0·393	0·331	0·262
		(0·000)	(0·000)	(0·000)	(0·000)	(0·000)	(0·339)	(0·344)	(0·345)	(0 343)	(0·342)	(0 315)	(0·288)	(0·233)	(0·195)
0·95	m/n	1·000	0·675	0·530	0·444	0·386	0;342	0·309	0·282	0·259	0·240	0·176	0·137	0·111	0·092
		(1·000)	(1·000)	(1·000)	(1·000)	(0·280)	(0·250)	(0·230)	(0·210)	(0·200)	(0·180)	(0·140)	(0·110)	(0·090)	(0·080)
	u/n	0·900	0·865	0·841	0·822	0·807	0·795	0·784	0·775	0·767	0·760	0·736	0·725	0·722	0·725
		(0·900)	(0·800)	(0·700)	(0·600)	(0·860)	(0 850)	(0·839)	(0·832)	(0·820)	(0·820)	(0·790)	(0·780)	(0·775)	(0·760)
	w	0·526	0·524	0·522	0·519	0·517	0·514	0·510	0·507	0·504	0·500	0·480	0·459	0 435	0·409
		(0·000)	(0·000)	(0·000)	(0·000)	(0·380)	(0·384)	(0·387)	(0·388)	(0·390)	(0·387)	(0·383)	(0·367)	(0·350)	(0·342)

TABLE 2

Percent gain in efficiency of $\bar{y}_r'$ over $\bar{y}_a'$ (and in parantheses over $\bar{y}_c'$) when $S_x/S_y=k=0{\cdot}75$ or $1{\cdot}00$ or $1{\cdot}25$ using the optimum values of m/n, u/n and w from Table 1 (for $k=1$).

ρ	$\delta=\frac{c_m}{c_u}$	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0	1·5	2·0	2·5	3·0
0·6		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	2(2)	1(6)	1(6)	0(0)	0(0)	0(0)	0(0)
0·7		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(2)	0(8)	0(8)	0(8)	1(4)	0(0)	0(0)	0(0)
0·8	k=0·75	0(0)	0(0)	0(0)	0(0)	0(1)	0(3)	0(7)	0(12)	0(12)	0(12)	0(9)	0(5)	0(1)	0(0)
0·9		0(0)	0(0)	0(1)	1(3)	1(6)	1(17)	1(18)	2(18)	2(18)	2(18)	2(17)	2(15)	2(13)	2(11)
0·95		0(0)	1(1)	2(4)	2(7)	3(23)	3(23)	4(24)	4(24)	5(24)	5(25)	6(25)	7(25)	8(24)	8(24)
0·6		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	2(2)	4(4)	5(4)	0(0)	0(0)	0(0)	0(0)
0·7		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	2(2)	2(5)	3(5)	4(5)	2(2)	0(0)	0(0)	0(0)
0·8	k=1·00	0(0)	0(0)	0(0)	0(0)	0(1)	1(3)	1(7)	1(7)	2(6)	2(6)	2(4)	2(2)	0(0)	0(0)
0·9		0(0)	0(0)	0(1)	0(3)	0(6)	0(9)	0(8)	1(8)	1(7)	1(7)	1(5)	1(4)	1(3)	1(2)
0·95		0(0)	0(1)	0(4)	0(7)	0(9)	0(8)	0(8)	0(7)	0(7)	0(7)	0(5)	0(4)	0(3)	0(3)
0·6		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	2(2)	10(2)	13(2)	0(0)	0(0)	0(0)	0(0)
0·7		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	4(2)	8(3)	10(3)	12(2)	7(1)	0(0)	0(0)	0(0)
0·8	k=1·25	0(0)	0(0)	0(0)	0(0)	3(1)	5(3)	6(7)	8(3)	9(3)	10(3)	11(1)	8(0)	0(0)	0(0)
0·9		0(0)	0(0)	2(1)	4(3)	5(6)	6(3)	7(3)	8(2)	9(2)	10(2)	12(1)	13(1)	13(1)	12(1)
0·95		0(0)	2(1)	4(4)	5(7)	6(2)	7(2)	8(2)	9(1)	10(1)	11(1)	14(2)	16(2)	17(3)	18(3)

where

$$\frac{1}{n_d'} = \frac{k(2\rho-k)}{n} + \frac{1-k(2\rho-k)}{m}. \qquad \ldots(10)$$

Comparison of (9) and (10) with (2) and (3) immediately shows that the formulae for the optimum allocation of m and u for $\bar{y}_d'$ are obtained simply by replacing ρ^2 by $k(2\rho-k)$ in Kulldorff's formulae (see his Table 1). We now consider in detail the optimum allocation when the total cost is fixed to c_0+c_2n (*i.e.*, $R=n$). Since a good guessed value of k is normally not available in practice, we give, in Table 1, the optimum values of m/n, u/n and w_d for $k=1$. Using these values in (7) we computed the gain in efficiency of $\bar{y}_r'$ over $\bar{y}_d'$ when $k=0{\cdot}75$ or $1{\cdot}00$ or $1{\cdot}25$ (ignoring the term $-S_y^2/N$) and the results are given in Table 2—the effect of ignoring the term $-S_y^2/N$ is to decrease the gain in efficiency of $\bar{y}_r'$ over $\bar{y}_d'$. It may be noted that normally in practice $0{\cdot}75 \leqslant k \leqslant 1{\cdot}25$.

The following conclusions may be drawn from Table 2 : (1) when $k=0{\cdot}75$ and $\rho \leqslant 0{\cdot}90$, the gain in efficiency of $\bar{y}_r'$ over $\bar{y}_d'$ (G say) is less than or equal to 2% ; for $\rho=0{\cdot}95$ and $\delta > 0{\cdot}6$, G is moderate but not larger that 8%. (2) When $k=1{\cdot}0$, G is less than or equal to 2% except when $\rho=0{\cdot}6$ or $0{\cdot}7$ and $\delta=0{\cdot}9$ or $1{\cdot}0$; however, it is never larger than 5%. (3) When $k=1{\cdot}25$ and $\delta > 0{\cdot}5$, G can be substantial, especially when $\rho \geqslant 0{\cdot}8$ (as high as 18%). For the important case of $\delta=1{\cdot}0$, G is greater than or equal to 10% for all values of ρ considered here.

3. Composite Estimator

It is easily seen that the variance of the composite estimator $\bar{y}_c'$ is given by

$$V(\bar{y}'_c) = \left[w_c^2 \left\{ \frac{1}{m} + \left(\frac{1}{m} - \frac{1}{n} \right) \left(k^2 - 2\rho k \right) \right\} + \frac{(1-w_c)^2}{m+u} \right.$$

$$\left. + 2w_c(1-w_c) \left(\frac{m}{m+u} \right) \left(\frac{\rho k}{n} + \frac{1-\rho k}{m} \right) - \frac{1}{N} \right] S_y^2. \qquad (11)$$

For $k=1$, minimization of (11) with respect to w_c leads to

$$w_{c(opt)} = \frac{m\rho(n-m)}{(m+u)\{(2\rho-1)m+2(1-\rho)n\}+2\rho(n-m)m-nm} \qquad (12)$$

and

$$V_{opt}\bar{y}_c' = \left[\frac{1}{n_c'+u} - \frac{1}{N} \right] S_y^2 \qquad (13)$$

where

$$\frac{1}{n_c'}=\frac{(n-m)\left[(m+u)-2\rho u-\frac{m}{n}\rho^2(n-m)\right]+mu}{m[(n-m)(m+u)-2\rho n+\frac{u}{n}\rho^2(u-m)\}+nu]} \qquad (14)$$

Since n'_c contains u, it is not possible to use kulldorff's method. For the case $R=n$ we can substitute $u=n-\delta_m$ in (13) and minimize the resulting expression with respect to $\lambda=m/n$; however, this would lead to a fourth degree equation in λ (except when $\delta=1$). Therefore, we found the optimum value of λ for each selected combination (ρ, δ) empirically using a high speed computer, noting that $0\leqslant\lambda\leqslant 1/\delta$. The optimum values of w_c and u/n were obtained by substituting these values of λ in (12) and $u/n=1-\delta\lambda$. The results are given in Table 1 in parentheses. For the special case $\delta=1$, it is easily seen that $\lambda_{opt}=\sqrt{1-\rho}/[1+\sqrt{1-\rho}\ ]$.

Using the optimum values of λ, u/n and w_c in (11) we computed the gain in efficiency of $\bar{y}'_r$ over $\bar{y}'_c$ (ignoring $-S_y^2/N$) when $k=0\cdot 75$ or $1\cdot 00$ or $1\cdot 25$ and the results are given in Table 2 in parantheses. It is evident from Table 2 that $\bar{y}'_c$ is less efficient than $\bar{y}'_d$ except when $k=1\cdot 25$ and $\delta>0\cdot 5$. The loss in efficiency of $\bar{y}'_c$ is considerable for $k=0\cdot 75$, specially when $\rho\geqslant 0\cdot 9$ and $\delta\geqslant 0\cdot 5$. On the other hand, $\bar{y}'_c$ is considerably more efficient than $\bar{y}'_d$ when $k=1\cdot 25$ and $\delta>0\cdot 5$, and in these cases the gain in efficiency of $\bar{y}'_r$ over $\bar{y}'_c$ is small (less than 3%).

The results given in this note will be applicable to unequal probability sampling with replacement if ρ is defined as

$$\rho=\sum_1^N p_i\left(\frac{y_i}{p_i}-Y\right)\left(\frac{x_i}{p_i}-X\right)\Big/\left\{\left[\sum_1^N p_i\left(\frac{y_i}{p_i}-Y\right)^2\right]^{1/2}\left[\sum_1^N p_i\left(\frac{x_i}{p_i}-X\right)^2\right]^{1/2}\right\}$$

where X and Y are the population totals and p_i is the probability of selected the i^{th} unit in a draw.

The following recommendations may be made from this study : (1) If S_x^2 is expected to be about equal or smaller than S_y^2, the difference estimator would be satisfactory for all values of δ, (2) If

$\delta \leqslant 0{\cdot}5$, the difference estimator would be satisfactory for all S_x/S_y in the range of 0·25 to 1·25. (3) If S_x^2 is expected to be larger than S_y^2 and $\delta > 0{\cdot}5$, the composite estimator may be preferable over the difference estimator.

REFERENCES

Des Raj (1965a), "On a method of using multi-auxiliary information in sample surveys", *Jour. Amer. Statist. Assoc.*, Vol. 60, pp. 270-277.

Des Raj (1965*b*), "On sampling over two occasions with probability proportional to size", *Ann. Math. Statist.*, Vol. 36, pp. 327-330.

M.H. Hansen, W.H. Hurwitz and W.G. Madow (1953), *Sample Survey Methods and Theory*. Vol. 2, John Wiley and Sons, New York.

G. Kulldorff (1963), "Some problems of optimum allocation for sampling on two occasions", *Rev. International Statist. Inst.*, Vol. 31, pp. 24-56.