REPLICATED (OR INTERPENETRATING) SAMPLES OF UNEQUAL SIZES[1]

By J. C. Koop[2]

North Carolina State University at Raleigh

1. Introduction.

1.1. The technique of replicated (or interpenetrating) samples, introduced by Mahalanobis, is now well known. Some aspects of this technique, and its underlying theory, relative to equal-sized samples, were clarified by Lahiri (1954) and Koop (1960). Deming (1956) has given a simplified version of the technique based on the use of two or more systematic samples drawn from the entire universe of ultimate units or elements.

1.2. In practice equal-sized replicated samples are used. However if considerations of field work, and/or other causes, should dictate a departure from this practice, the outcome would still be favorable. The main purpose of this paper is to point out this surprising result; that is, replicated samples of unequal sizes are more efficient than those with equal sizes.

2. Technique underlying the theory.

2.1. We consider a single universe U containing N first-stage units

$$u_1, u_2, \cdots, u_j, \cdots, u_N.$$

The first-stage units are selected with equal probabilities, and without replacement after each draw. Beyond this stage, the procedure, say D_j, ($j = 1, 2, \cdots, N$) for selecting the units at the second and subsequent stages, with equal or unequal probabilities and with or without replacement, is specified in advance for each u_j. For the sake of generality we do not explicitly define this procedure.

2.2. Let X_j ($j = 1, 2, \cdots, N$) be the total value of a characteristic of interest in the first-stage unit u_j. We note that X_j is equal to the sum of all the variate values in the ultimate units of u_j. It is desired to estimate

$$T = \sum_{j=1}^{N} X_j, \tag{1}$$

on the basis of k independent replicated samples each of size $m_i > 1$ ($i = 1, 2, \cdots, k$). For this purpose, m_1 first-stage units are selected as specified in the previous paragraph. Denote this collection of first-stage units by s_1. Then for every unit u_j, which is a member of s_1, the procedure D_j is applied for selecting the second- and subsequent-stage units and for ascertaining the variate values of the ultimate-stage units. Next these m_1 first-stage units are returned to U. The same procedure is repeated for the selection of sets of $m_2, m_3, \cdots, m_k$ units, and for the selection of the multi-stage units internal to them.

Received 11 October 1965; revised 27 December 1966.

[1] Paper read at the 28th Annual Meeting of the Institute of Mathematical Statistics at Philadelphia, Pennsylvania, on September 10, 1965.

[2] Now at the Dominion Bureau of Statistics, Ottawa.

2.3. We digress to note that all the k replicated samples (of unequal sizes) are mutually statistically independent simply because of the replacement of first-stage units after each set of draws.

2.4. In all there are $\sum_1^k m_i = \bar{m}k$ first-stage units, and the m_i's are chosen so that $\bar{m}$ is an integer. The overall first-stage sample size, $\bar{m}k$, is fixed.

2.5. Similarly we may draw k sets of independent replicated samples each containing an equal number of first-stage units $\bar{m}$. In current practice this is usually done.

3. Theory and solution of problem.

3.1. Now consider the replicated sample s_i based on m_i first-stage units ($i = 1, 2, \cdots, k$). Let $\hat{X}_j$ be an unbiased estimate of X_j, with variance $V(\hat{X}_j)$, for all j. Then T_i ($i = 1, 2, \cdots, k$), an unbiased estimate of T will be given by

$$(2) \qquad T_i = (N/m_i) \textstyle\sum_{u_j \varepsilon s_i} \hat{X}_j ,$$

with variance (Madow (1949))

$$(3) \qquad V(T_i) = N^2[(S^2/m_i)(1 - m_i/N) + (1/m_i)(N^{-1}\textstyle\sum_{j=1}^N V(\hat{X}_j))]$$

where $S^2 = \sum_1^N (X_i - T/N)^2/(N - 1)$. The expression for $V(\hat{X}_j)$ can be obtained by applying Madow's theorem again as soon as the procedure D_j, defined in paragraph 2.1, is specified in detail for all j. Writing $N^{-1}\sum_1^N V(\hat{X}_j) = E(u)$, suggestive of the fact that this value is the mathematical expectation of the variance functions internal to each u_j, $\alpha = S^2/(S^2 + E(u))$ and $f_i = m_i/N$, (3) can be rewritten as

$$(4) \qquad V(T_i) = (1/m_i)N^2(S^2 + E(u))(1 - f_i\alpha).$$

We note that $0 < \alpha \leqq 1$, and in uni-stage sampling, when $E(u)$ is formally zero, α attains its greatest possible value one.

3.2. Now an unbiased estimator of T, with the least variance, based on the k independent estimates defined at (2), will be given by

$$(5) \qquad T_u = \textstyle\sum_{i=1}^k w_i T_i ,$$

where $w_i > 0$ and $\sum_1^k w_i = 1$, when

$$(6) \qquad w_i = (1/V(T_i))/\textstyle\sum_1^k (1/V(T_i)), \qquad (i = 1, 2, \cdots, k).$$

Substituting for the $V(T_i)$ in (6) we find

$$(7) \qquad w_i = (m_i/(1 - f_i\alpha))/\textstyle\sum_1^k (m_i/(1 - f_i\alpha)).$$

The variance of the estimator T_u (with weights specified by (7)) is

$$(8) \qquad V_u = 1/\textstyle\sum_1^k (1/V(T_i)).$$

The results at (6) and (8), under other notations, were well known to surveyors and astronomers in the problem of combining observations. Later in paragraph 4.3 the approximations for the w_i's will be considered. Substituting $V(T_i)$ given by (4) in (8), we find

$$V_u = N^2(S^2 + E(u))/\textstyle\sum_1^k [m_i/(1 - f_i\alpha)]. \tag{9}$$

As a subsidiary problem, we shall consider the estimation of V_u in Section 5.

3.3. Equally, on the basis of results at (4), (6) and (8), an unbiased estimator of T, based on k independent replicated samples each containing $\bar{m}$ first-stage units and having the least variance, will be

$$T_e = \textstyle\sum_{i=1}^k T_i/k, \tag{10}$$

with variance

$$V_e = V(T_i)/k = N^2(S^2 + E(u))(1 - \bar{f}\alpha)/\bar{m}k, \tag{11}$$

where $\bar{f} = \sum_1^k f_i/k$. We observe that $V_e = V_u$ when $m_1 = m_2 = \cdots = m_k = \bar{m}$.

3.4. In what follows it will be demonstrated that $V_e > V_u$ when at least two among k values of the m_i's are unequal. From (11) and (9) we find

$$V_e/V_u = (1/k\bar{m})(1 - \bar{f}\alpha)\textstyle\sum_1^k [m_i/(1 - f_i\alpha)]. \tag{12}$$

We recall that $k\bar{m}$, the overall first-stage sample size, is fixed; $\bar{f} = \bar{m}/N$ is also fixed. Only the expression under the summation sign in (12) can vary, subject to the choice of (positive) values of m_i which sum to the fixed value $k\bar{m}$. In view of this, consider the expression multiplied by $\sum_1^k m_i$. By Cauchy's inequality

$$(\textstyle\sum_{i=1}^k m_i)(\sum_{i=1}^k m_i/(1 - f_i\alpha)) \geqq [\sum_1^k m_i^{\frac{1}{2}}(m_i/(1 - f_i\alpha))^{\frac{1}{2}}]^2. \tag{13}$$

Equality in (13) is attained if and only if

$$(m_i/(1 - f_i\alpha))^{\frac{1}{2}}/m_i^{\frac{1}{2}} = 1/(1 - f_i\alpha)^{\frac{1}{2}} = \text{a constant}, \qquad \text{for all } i. \tag{14}$$

Even if two of the m_i's are unequal (13) will still be true. When condition (14) is satisfied, which implies that

$$1 - f_i\alpha = 1 - (m_i/N)\alpha = \lambda \qquad (i = 1, 2, \cdots, k), \tag{15}$$

where λ is a constant, the lowest bound for the expression on the left-hand side of (13) will be obtained. Noting that the restricting condition $\sum_1^k m_i = k\bar{m}$ holds, and summing over all i in (15) we find

$$k\lambda = k - (\textstyle\sum_1^k m_i)\alpha/N = k - k\bar{m}\alpha/N,$$

so that

$$\lambda = 1 - \bar{f}\alpha = 1 - f_i\alpha \qquad (i = 1, 2, \cdots, k). \tag{16}$$

Further from (16) we have

$$f_i = \bar{f} \quad \text{or} \quad m_i = \bar{m}, \qquad \text{for all } i. \tag{16.1}$$

Hence the lowest value which the expression on the left-hand side of (13) can attain, subject to the restriction that $\sum m_i = k\bar{m}$, is when (16.1) holds. This value is $(\sum_1^k \bar{m}/(1 - \bar{f}\alpha)^{\frac{1}{2}})^2 = (k\bar{m})^2/(1 - \bar{f}\alpha)$. Hence even when two of the

m_i's are unequal,

$$k\bar{m}\sum_1^k [m_i/(1 - f_i\alpha)] > (k\bar{m})^2/(1 - \bar{f}\alpha),$$

so that

$$V_e/V_u = (1/k\bar{m})(1 - \bar{f}\alpha)\sum_1^k [m_i/(1 - f_i\alpha)] > 1. \tag{17}$$

4. Comments.

4.1. We note that (17) is true regardless of the value of $\alpha = S^2/(S^2 + E(u))$, and regardless of the nature of the sample design internal to the first-stage units.

4.2. Numerical calculations will show that for any given k and α, the ratio V_e/V_u increases with increasing dispersion among the m_i's, and it attains its maximum when $m_1 = m_2 = \cdots = m_{k-1} = 2$ and $m_k = \bar{m}k - 2(k - 1)$. We present some of these calculations for a universe of $N = 100$ first-stage units and for overall sample sizes of $\bar{m}k = 10, 20, 30$.

The tables speak for themselves. We only note that when the overall sampling fraction $\bar{m}k/N$ is as high as 3/10, the ratio V_e/V_u can rise as high as 1.177 for uni-stage sampling when dispersion among the sizes of the three replicated samples is highest.

4.3. Regarding the problem of weights for T_u, for uni-stage sampling, since $\alpha = 1$ for all characteristics of interest, we find

$$w_i = (m_i/(1 - f_i))/\sum_1^k (m_i/(1 - f_i)) \qquad (i = 1, 2, \cdots, k)\cdot$$

TABLE (i)
$k = 2$

Sample sizes (m_1, m_2)	(2, 8)			(4, 6)			(2, 18)		
α	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{2}$	$\frac{3}{4}$	1
V_e/V_u	1.009	1.015	1.020	1.0010	1.0016	1.0022	1.036	1.056	1.080
Sample sizes (m_1, m_2)	(4, 16)			(2, 28)			(4, 26)		
α	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{2}$	$\frac{3}{4}$	1
V_e/V_u	1.022	1.032	1.045	1.066	1.109	1.160	1.047	1.077	1.114

TABLE (ii)
$k = 3$

Sample sizes (m_1, m_2, m_3)	(2, 2, 26)			(4, 4, 22)		
α	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{2}$	$\frac{3}{4}$	1
V_e/V_u	1.074	1.121	1.177	1.014	1.067	1.096

However, in multi-stage sampling α (which lies between 0 and 1) will vary from characteristic to characteristic. For calculating the weights for each characteristic, α can be estimated from one of the samples. Fortunately this will not be necessary as the w_i's can be approximated as $m_i/k\bar{m}$ $(i = 1, 2, \cdots, k)$ (since $f_i\alpha$ will be closer to zero than one, particularly if the individual f_i's are less than 1/10, as they usually are), thus leading to a uniform vector of weights for all characteristics, which is a desirable property of the technique.

5. Estimation of variance.

5.1. One of the objectives of the technique of replicated samples is to simplify the estimation of sampling variance. Recently Srikantan (1964) showed that an unbiased estimate of the variance of T_u, with weights w_i different from those given by (7) is

$$\{\sum w_i^2 T_i^2 - (\sum w_i^2) T_u^2\}/(1 - \sum w_i^2).$$

He computed estimates of variance for the two cases when these weights were chosen (i) equal and (ii) unequal. Of course in case (1), the formula reduces to $\sum_1^k (T_i - T_e)^2/k(k-1)$, where T_e is given by (10), which is well known. In case (ii), an estimate can assume negative values.[3] We shall show that positive estimates of variance, which are unbiased, can always be constructed.

5.2. One of them is given by

$$(18) \quad \hat{V}_u = [(m_H/(1 - \alpha f_H))/(\textstyle\sum_1^k m_i/(1 - \alpha f_i))]\cdot \textstyle\sum_1^k (T_i - \bar{T})^2/(k-1)$$

where m_H is the harmonic mean of the m_i's, i.e. $m_H^{-1} = \sum_1^k m_i^{-1}/k$, $f_H = m_H/N$ and $\bar{T} = \sum_1^k T_i/k$. We recall that $\alpha = S^2/(S^2 + E(u))$. Remembering that $E(T_i^2) = V(T_i) + T^2$ for all i, $E(T_iT_{i'}) = T^2$ for $i \neq i'$, and using the formula for $V(T_i)$ given by (4), the essential steps in the proof of (18) are as follows:

$$\begin{aligned}(19) \quad & E[\textstyle\sum_1^k (T_i - \bar{T})^2/(k-1)] \\ &= E[k^{-1}\textstyle\sum T_i^2 - (1/k(k-1))\textstyle\sum_{i\neq i'} T_iT_{i'}] = k^{-1}\textstyle\sum V(T_i) \\ &= N^2(S^2 + E(u))(1 - \alpha f_H)/m_H .\end{aligned}$$

The proof is completed by eliminating $N^2(S^2 + E(u))$ between (19) and (9).

5.3. For any given α, the expression in the square brackets in (18) can be approximated as $m_H/k\bar{m}$, because it can be shown that in the underlying binomial expansion, where second order terms are neglected, the first order term, $\alpha\{f_H - (\sum f_im_i)/(\sum m_i)\}$, is nearly zero, so that

$$(20) \quad \hat{V}_u \text{ (approximately)} = (m_H/\bar{m})\textstyle\sum_1^k (T_i - \bar{T})^2/k(k-1).$$

5.4. Another unbiased estimator of V_u is

$$(21) \quad V'_u = \textstyle\sum_1^k w_i(T_i - T_u)^2/(k-1)$$

[3] Shrikantan had the courage to publish these estimates. The general problem of negative estimates of variance in sampling theory remains open (Koop (1957), (1964)), both in regard to interpretation, and to the nature of the circumstances giving rise to such estimates.

where the w_i's are given by (7). The proof that $E(V'_u) = V_u$ is much simpler. The essential steps are as follows:

$$E[\textstyle\sum_1^k w_i(T_i - T_u)^2/(k-1)] = [\sum_1^k w_i V(T_i) - \sum_1^k w_i^2 V(T_i)]/(k-1).$$

Next by substituting for the w_i's, V_u as given by (8) will be obtained. As argued in paragraph 4.3, w_i in (21) can be approximated by $m_i/k\bar{m}$. This estimator will be more stable if the m_i's, and therefore the w_i's, vary substantially.

5.5. In uni-stage sampling $\alpha = 1$. By substituting this value in (18) and (21), we obtain the exact analogous expressions for the estimates of V_u.

Acknowledgment. The author is grateful to the referee, and an associate editor, for various criticisms which have improved the presentation of the paper and for a suggestion leading to an alternative estimator of variance given in paragraph 5.4.

REFERENCES

DEMING, W. E. (1956). On simplification of sampling design through replication with equal probabilities without stages. *J. Amer. Statist. Assoc.* **51** 24–53.

KOOP, J. C. (1957). Contributions to the general theory of sampling finite populations without replacement and with unequal probabilities. N. Carolina Inst. Statist. Mimeo Series 296.

KOOP, J. C. (1960). On theoretical questions underlying the technique of replicated or interpenetrating samples. *Proc. Soc. Statist. Sec. Amer. Statist. Assoc.* 196–205.

KOOP, J. C. (1964). Some properties of random variables. *Nature, London* **203** 1097–1098.

LAHIRI, D. B. (1954). National Sample Survey No. 5. Technical paper on some aspects of the sample design. *Sankhyā,* **14** 268–316.

MADOW, W. G. (1949). On the theory of systematic sampling II. *Ann. Math. Statist.* **20** 333–354.

SRIKANTAN, K. S. (1964). A note on interpenetrating sub-samples of unequal sizes. *Sankhyā Ser. B* **25** 345–350.