CONTROLLED SIMPLE RANDOM SAMPLING

By M. S. Avadhani and B. V. Sukhatme
Institute of Agricultural Research Statistics, New Delhi

1. Introduction and Summary

When units are drawn one after another with equal probabilities and without replacement from a finite population it is well known that all possible samples of a given size are equally likely to materialize. As such the sampling procedure may result in the selection of a sample which is not quite desirable. For example, it may happen that the sample contains quite a few units which are not important from the point of view of the character under study. Or, it may happen that the sampling units are spread out very much into the interior thereby not only increasing considerably the expenditure on travel but also affecting adversely the supervision and organization of fieldwork. All these factors would seriously affect the quality of the data collected and consequently the precision of the estimate of the parameter in question would be reduced. Such samples which are not desirable will hereafter be referred to as 'Non-preferred' samples (Goodman and Kish, 1950). Hence arises the need of developing a suitable sampling methodology which reduces the risk of getting a non-preferred sample from the population to the minimum possible extent and yet conforming to the fundamental principles of random sampling procedures.

The authors have attempted this problem with varying probabilities (Sukhatme and Avadhani, 1965) and have given an explicit solution for drawing a sample of size 2 from a given population together with appropriate procedures of estimation. The procedure is perfectly general but becomes complicated and tedious with increasing size. Further, the precision of the estimate of the parameter in question increases as the probability, say α, of drawing a non-preferred sample increases and decreases as α decreases. This constitutes a severe drawback of the procedure in question. Thus, arises the question whether or not there exists a random sampling scheme by means of which α can be reduced to the minimum possible extent without affecting the precision of the estimate of the parameter under consideration.

Assuming that no ancillary information on the units of the population is available, the authors have shown in this paper that there exists a random sampling method, alternative to the usual simple random sampling scheme, which reduces α to the minimum possible extent and yet an unbiased estimate of the population total $N\bar{y}_N$ and its variance are respectively given by:

$$N\bar{y}_n \tag{1.1}$$

and

$$\frac{N(N-n)}{n} S_y^2 \tag{1.2}$$

where $\bar{y}_n$ is the mean of the observations obtained from the n units of the sample and

$$S_y^2 = \frac{1}{N-1} \sum_{r=1}^{N} (y_r - \bar{y}_N)^2$$

is the mean square of the population.

2. Notation and Definitions

Let

$$\pi: \{u_r\}, \qquad r = 1, 2, \ldots, N$$

be the population in question and let n be any integer greater than or equal to 3 but less than N. Consider the class S of all possible sets s of n distinct units, u_r of π, *viz.*,

$$S \ni s: (u_{i_1}, u_{i_2} \ldots, u_{i_n}) \qquad 1 \leqslant i_j \leqslant N,$$

so that S consists of only $\binom{N}{n}$ distinct sets 's'.

Definition (2.1).—The class of sets $\{S \ni s\}$ defined on π will be referred to as the *sample space* of size $\binom{N}{n}$.

Definition (2.2).—The sample space S together with an associated P-measure, $\{P \ni P_s, s \epsilon S\}$ where $P_s \geqslant 0$ and $\sum_{s \epsilon S} P_s = 1$ is defined as a *random sampling design* D_n for drawing a sample of size n from π.

Definition (2.3).—The class of all non-preferred sets, $\{C(S) \ni s, s \epsilon S\}$ is said to constitute the '*control space*' of the design D_n.

Hence it is clear that

$$\sum_{s \epsilon C(S)} P_s = \alpha$$

it being assumed that $0 \leqslant \alpha < 1$.

Definition (2.4).—A design D_n is said to be *sufficient* when its P-measure satisfies the system of equations:

$$\sum_{s \epsilon S_r} P_s = \frac{n}{N} \qquad r = 1, \ 2, \ \ldots, N \qquad (2\,A)$$

where S_r is the class of all sets $s \epsilon S$ containing U_r.

The set $\boldsymbol{P}$ of all P-measures satisfying (2 A) is said to constitute the *class of sufficient designs*, $\boldsymbol{D}_n$.

Definition (2.5).—A class of designs $\boldsymbol{D}_n'$ is said to be *admissible* when and only when it is sufficient and for any design $D_n' \epsilon \boldsymbol{D}_n'$, the estimate of the population total and its variance associated with D_n' are independent of α and the variance is never more than in the case of simple random sampling design.

Put in this terminology the problem under consideration reduces to showing that there exists an admissible class of designs $\boldsymbol{D}_n'$ and that the estimate of the population total and its variance associated with any design $D_n' \epsilon \boldsymbol{D}_n'$ are respectively given by (1.1) and (1.2).

3. On the Existence of Admissible Class of Designs

In this section, we shall first prove the existence of admissible class of designs and then show that any member of this class is always more efficient than the usual random sampling design where the units are selected with equal probabilities and without replacement.

Theorem (3.1).—The class of all designs whose P-measures satisfy the system of equations, *viz.*,

$$\sum_{s \epsilon S_r \cap S_{r'}} P_s = \frac{n(n-1)}{N(N-1)} \qquad r \neq r' = 1, \ 2, \ \ldots, N \qquad (3\,A)$$

where $S_r \cap S_{r'}$ is the class of all sets $s \epsilon S$ which contain U_r and $U_{r'}$, is admissible.

Proof.—To prove this it is enough to show that any design D_n satisfying (3 *A*) is a sufficient design and that the estimate of the population total and its variance are independent of α.

Since the system (3 *A*) consists of $\binom{N}{2}$ equations in $\binom{N}{n}$ unknowns and $n \geqslant 3$, it is clear that there exists infinitely many *P*-measures satisfying (3 *A*).

Let D_n be a design whose *P*-measure satisfies (3 *A*). Since

$$\bigcup_{r'(\neq r)=1}^{N} (S_r \cap S_{r'}) = S_r$$

and each P_s, $s \in S_r$ occurs in $(n-1)$ distinct pairs containing U_r, it is clear that

$$\sum_{r'(\neq r)=1}^{N} \sum_{s \in S_r \cap S_{r'}} P_s = \sum_{s \in \bigcup_{r'(\neq r)=1}^{N} (S_r \cap S_{r'})} P_s = (n-1) \sum_{s \in S_r} P_s \qquad (3.1)$$

Now adding the system (3 *A*) over all pairs containing U_r it follows evidently that:

$$\sum_{r'(\neq r)=1}^{N} \sum_{s \in S_r \cap S_{r'}} P_s = (N-1) \cdot \frac{n(n-1)}{N(N-1)} = \frac{n(n-1)}{N} \qquad (3.2)$$

Hence from (3.1) and (3.2) it is clear that:

$$\sum_{s \in S_r} P_s = \frac{n}{N}$$

showing thereby that D_n is a sufficient design. But this is true for any D_n satisfying (3 *A*). Hence the class of all designs D_n satisfying (3 *A*) is a class of sufficient designs.

Further, since the system (3 *A*) preserves pair-wise and individual inclusion probabilities it is clear that the Horvitz-Thompson estimate of the population total and its variance are readily given by:

$$N\bar{y}_\eta$$

and

$$\frac{N(N-n)}{n} \cdot S_y^2$$

which are independent of α and remain unaltered for any design belonging to the class $\boldsymbol{D}_n$.

Hence the theorem.

THEOREM 3.2.—Any member of the admissible class of designs is always more efficient than simple random sampling without replacement design provided:

$$\alpha < 1 - \frac{\binom{N}{2}}{\binom{N}{n}}.$$

Proof.—Let c_1 be the cost of collecting data from a preferred sample and c_2 be the cost of collecting data from a non-preferred sample, where $c_2 > c_1$.

Now the expected cost for simple random sampling without replacement (s.r.s.) design is given by:

$$C_{s.r.s.} = c_1 \cdot \frac{\binom{N}{2}}{\binom{N}{n}} + c_2 \cdot \frac{\binom{N}{n} - \binom{N}{2}}{\binom{N}{n}}$$

it being assumed that there are only $\binom{N}{n} - \binom{N}{2}$ non-preferred samples in the sample space S. Further, the expected cost for any member of the admissible class of designs corresponding to α is given by

$$C_{adm.} = c_1 \sum_{s \in \bar{C}(S)} P_s + c_2 \sum_{s \in C(S)} P_s = c_1(1-\alpha) + c_2\alpha$$

where $\bar{C}(s)$ is the complement of $C(s)$ w.r.t.s. Now in virtue of Theorem (3.1) it is evident that any member of the admissible class is more efficient than the s.r.s. design if and only if $C_{s.r.s.} > C_{adm.}$. But

$$C_{s.r.s.} - C_{adm.}$$

$$= c_1\left[\alpha - \left(1 - \frac{\binom{N}{2}}{\binom{N}{n}}\right)\right] + c_2\left[1 - \frac{\binom{N}{2}}{\binom{N}{n}} - \alpha\right]$$

$$= \left(1 - \frac{\binom{N}{2}}{\binom{N}{n}} - \alpha\right)(c_2 - c_1) > 0$$

since

$$\alpha < 1 - \frac{\binom{N}{2}}{\binom{N}{n}} \quad \text{and} \quad c_2 > c_1.$$

Hence the result.

In view of what has been said above, the procedure for derivation of the admissible class of designs for a given α may be described as follows.

Choose $\binom{N}{n} - \binom{N}{2}$ non-preferred sets s which constitute the class $C(S)$, from S such that when the P_s, $s \epsilon C(S)$, are omitted from the 1.h.s. of $(3A)$, the resulting linear system turns out to be consistent. Then distribute α among the P_s, $s \epsilon C(S)$, and after substituting these assigned values in $(3A)$ solve the resulting system for a non-negative solution which, together with the predetermined P-measure of $C(S)$, determines uniquely a member of the admissible class in question whose members correspond in a (1.1) fashion to each such distribution of α among the components of the P-measure of $C(S)$.

It follows that when $\alpha = 0$ the only value which the P_s, $s \epsilon C(S)$, can take is zero. Hence the best among the members of the admissible class with

$$0 \leqslant \alpha < 1 - \frac{\binom{N}{2}}{\binom{N}{n}}$$

is unique. Hence we have:

THEOREM (3.3).—The admissible class of designs with

$$0 \leqslant \alpha < 1 - \frac{\binom{N}{2}}{\binom{N}{n}}$$

contains one and only one member with $\alpha = 0$.

It may be noted that in order that the admissible class of designs with $\alpha > 0$ exists it is necessary that $\alpha \leqslant n(n-1)/N(N-1)$.

Further, when $\alpha = 0$ there may not exist a random sampling design for *any* choice of $C(S)$. Hence proper care must be exercised in fixing the members of $C(S)$ so that the resulting linear system possesses a *non-negative* solution for the P_s, $s \epsilon C(S)$.

An illustrative example is worked out in the next section to clarify these points.

4. ILLUSTRATION

Consider a population consisting of 6 units U_r, $r = 1, \ldots, 6$ and let a sample of size 3 be desired to be drawn from the best of the admissible class of designs constructed with reference to the population in question.

The sample space S may be defined as the single-rowed matrix, *viz.*, S: $(s_1, s_2, \ldots, s_{20})$ where the s_i are given by:

s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
u_1	u_1	u_1	u_1	u_1	u_1	u_1	u_1	u_1	u_1	u_2	u_2	u_2	u_2	u_2	u_2	u_3	u_3	u_3	u_4
u_2	u_2	u_2	u_2	u_3	u_3	u_3	u_4	u_4	u_5	u_3	u_3	u_3	u_4	u_4	u_5	u_4	u_4	u_5	u_5
u_3	u_4	u_5	u_6	u_4	u_5	u_6	u_5	u_6	u_6	u_4	u_5	u_6	u_5	u_6	u_6	u_5	u_6	u_6	u_6

and let the P-measure of S be denoted by P: $(P_1, P_2, \ldots, P_{20})$, where $P_i \geqslant 0$ and $\sum_{i=1}^{20} P_i = 1$.

Now the admissible class of designs are given by the system (3 *A*), *viz.*,

$$\left.\begin{array}{ll} P_1+P_2+P_3+P_4=\cdot 2 & P_4+P_{13}+P_{15}+P_{16}=\cdot 2 \\ P_1+P_5+P_6+P_7=\cdot 2 & P_5+P_{11}+P_{17}+P_{18}=\cdot 2 \\ P_2+P_5+P_8+P_9=\cdot 2 & P_6+P_{12}+P_{17}+P_{19}=\cdot 2 \\ P_3+P_6+P_8+P_{10}=\cdot 2 & P_7+P_{13}+P_{18}+P_{19}=\cdot 2 \\ P_4+P_7+P_9+P_{10}=\cdot 2 & P_8+P_{14}+P_{17}+P_{20}=\cdot 2 \\ P_1+P_{11}+P_{12}+P_{13}=\cdot 2 & P_9+P_{15}+{}_1P_8+P_{20}=\cdot 2 \\ P_2+P_{11}+P_{14}+P_{15}=\cdot 2 & P_{10}+P_{16}+P_{19}+P_{20}=\cdot 2 \\ P_3+P_{12}+P_{14}+P_{16}=\cdot 2 & \end{array}\right\} \quad (4.1)$$

It may be verified from (4.1) that the prob. that the unit u_r is included in any sample selected from a design satisfying (4.1) is given by $n/N = 3/6 = \cdot 5$ as is the case in virtue of Theorem (3.1).

To solve the system (4.1) the $C(S)$ must consist of only $\binom{N}{n} - \binom{N}{2}$ $= 20-15 = 5$ sets. Further, to get a design which is always more efficient than the s.r.s. design, α must satisfy the inequality $0 \leqslant \alpha < \cdot 2$ as the minimum of

$$\frac{n(n-1)}{N(N-1)} \quad \text{and} \quad 1 - \frac{\binom{N}{2}}{\binom{N}{n}} \text{ is } \frac{n(n-1)}{N(N-1)} = \cdot 2$$

in this case.

(i) Let $\alpha = 0$.

If $s_1, s_2, s_{16}, s_{17}, s_{18}$ are the members of $C(s)$, then setting $P_1 = P_2 = P_{16} = P_{17} = P_{18} = 0$ in (4.1) and solving for the remaining P's we get:

$$P_3 = P_4 = P_5 = 0 \cdot 10.$$
$$P_6 = P_7 = P_8 = P_9 = 0 \cdot 05, \quad P_{10} = 0,$$
$$P_{11} = 0 \cdot 10, \; P_{12} = P_{13} = P_{14} = P_{15} = 0 \cdot 05, \; P_{19} = P_{20} = 0 \cdot 10,$$

which give rise to a unique random sampling design which is the best of the admissible class of designs.

But on the other hand if we take $(s_1, s_6, s_7, s_{17}, s_{18})$ as the $C(S)$, then setting $P_1 = P_6 = P_7 = P_{17} = P_{18} = 0$ in (4.1) and solving for the remaining constants it is seen that:

$$P_2 = P_8 = P_9 = P_{11} = 0, \quad P_3 = P_4 = P_{10} = P_{12} = P_{13} = P_{14} = P_{15} = P_{19} = P_{20} = 0 \cdot 10, \quad P_5 = \cdot 20 \text{ and } \quad P_{16} = -0 \cdot 1,$$

which cannot give rise to a random sampling design since $P_{16} < 0$.

(ii) Let $\alpha = \cdot 20$ and suppose $s_6, s_7, s_8, s_9, s_{10}$ are the non-preferred samples of which s_{10} is the most undesired. Now setting $P_{10} = 0$; $P_6 = P_7 = P_8 = P_9 = 0 \cdot 05$, in (4.1) and solving for the remaining constants we get

$$P_1 = P_2 = 0; \; P_3 = P_4 = P_5 = 0 \cdot 10; \; P_{11} = 0 \cdot 10; \; P_{12} = P_{13} = P_{14} = P_{15} = 0 \cdot 05; \; P_{16} = P_{17} = P_{18} = 0; \; P_{19} = P_{20} = 0 \cdot 10,$$

which determines uniquely for a given $C(S)$ and its measure a member of the admissible class in question. It may be remarked here that in this design while the prob. of getting a non-preferred sample is maintained at the stipulated level of 20% it could not be helped for not getting at all the samples $s_1, s_2, s_{16}, s_{17}, s_{18}$ as the probabilities of these are uniquely determined once the P-measure of $C(S)$ is fixed.

5. Acknowledgement

The authors are grateful to Dr. V. G. Panse, Statistical Adviser, I.C.A.R., for his constant encouragement and keen interest in the preparation of this paper.

6. References

1. Goodman, R. and Kish, L. "Controlled selection—A technique in probability sampling," *Jour. Amer. Stat. Assoc.*, 1950, **45**, 350–72.

2. Sukhatme, B. V. and Avadhani, M. S. "Controlled selection—A technique in random sampling," *Ann. Inst. Stat. Math., Japan*, 1965.