MISCELLANEOUS

AN EXISTENCE THEOREM IN SAMPLING THEORY*

By T. V. HANUMANTHA RAO

Indian Statistical Institute

SUMMARY. A (1,1) correspondence is established here, between sampling designs and sampling schemes, for sampling from a finite population. This result enables us to search for optimum sampling procedures in any particular case, through a unified general set-up.

1. Introductions

Let a finite population, of N units be given by

$$U_1, U_2, ..., U_i, ..., U_N. \quad ... \quad (1.1)$$

We give the following definitions.

Definition 1: A *sample* 's' from the above population is an ordered sequence

$$U_{i_1}, U_{i_2}, ..., U_{i_{n_s}}; \quad 1 \leqslant i_t \leqslant N, \quad \text{for} \quad 1 \leqslant t \leqslant n_s, \quad ... \quad (1.2)$$

where the i's need not necessarily be distinct and n_s is called the size of the sample 's'.

Definition 2 : A *sampling scheme* is a process of selecting units one by one from the population (1.1) with pre-determined sets of probabilities of selection for individual units at each of the draws.

Definition 3 : A *sample design* D is an arbitrary collection 'S' of samples 's' with an arbitrary probability measure P defined on it, according to which the samples should be drawn. We can write explicitly

$$D = D(S, P), \quad ... \quad (1.3)$$

where

$$\sum_{s \epsilon S} P_s = 1.$$

It can be seen that this is the most general definition of a sample design.

It is known that any sampling scheme results in a unique sampling design, which is fully determined by it. That the converse also holds good, is shown in Section 2. We restrict our proof to the cases of practical importance when all the samples in the given design are of sizes $\leqslant m$, a fixed positive integer. However, the extension to the cases where this condition is not satisfied, such as sequential procedures, should not offer much difficulty.

A sample design is said to be completely specified if all possible samples including their permutations together with their respective probabilities of selection are specified. Examples of partially specified designs are (i) specification of all possible unordered samples

* Read before the Fortyseventh session of the Indian Science Congress, January 1960, Bombay.

with their respective probabilities of selection and (ii) specification of the probabilities of inclusion in the sample of each of the units. Corresponding to any partially specified sample design, there may be many possibly different completely specified sampling designs.

2. MAIN RESULTS

Theorem : *There is a one to one correspondence between completely specified sample designs and sampling schemes.*

Proof: We introduce a new unit U_0 called the null unit, into our population. The occurrence of U_0 in a draw means that none of the units belonging to (1.1) is selected in that particular draw. We replace any sample

$$s = (U_{i_1}, U_{i_2}, ..., U_{i_{n_s}}) \quad \text{where} \quad n_s < m,$$

by

$$s' = (U_{i_1}, U_{i_2}, ..., U_{i_{n_s}}, U_0, U_0, ..., U_0),$$

such that the size of s' is equal to m. We attach the same probability to 's'' as the corresponding 's'. Let S' be the set of all s' and P' be the probability measure on S', as constructed above.

For any sampling scheme, let $\{p^{(n)}_{i_n} | (i_1, i_2, ..., i_{n-1})\}$, denote the probability of selecting U_{i_n} in the n-th draw, given that the first $(n-1)$ draws resulted in the selections of $U_{i_1}, U_{i_2}, ..., U_{i_{n-1}}$ successively. We shall now find p's such that the resulting sample design is the given design D. It would be sufficient to consider instead, the design $D' = (S', P')$. Then, let S'_{i_1} be the subset of S' consisting of all samples s' for which the first unit is U_{i_1}. Similarly, let $S'_{i_1 i_2}$ be the set of all samples which have U_{i_1} as their first unit and U_{i_2} as their second unit. The definitions of $S'_{i_1 i_2 i_3}$ etc. are now similar. In all the above definitions, the indices $i_1, i_2, ...,$ can be any integers not necessarily distinct, of the index set $0, 1, ..., N$. Let us define $p^{(1)}_{i_1}$ by

$$p^{(1)}_{i_1} = \sum_{s' \epsilon S'_{i_1}} P_{s'}, \qquad 1 \leqslant i_1 \leqslant N. \qquad \dots \quad (2.1)$$

Clearly, $$\sum_{i_1 = 1}^{N} p^{(1)}_{i_1} = 1, \text{ since } \bigcup_{i_1 = 1}^{N} S'_{i_1} = S' \text{ and } \sum_{s' \epsilon S'} P_{s'} = 1.$$

This defines the probabilities of selection of all units in the first draw. Let

$$\left\{p^{(2)}_{i_2} | (i_1)\right\} = \begin{cases} \dfrac{\sum\limits_{s' \epsilon S'_{i_1 i_2}} P_{s'}}{p^{(1)}_{i_1}}, & \text{if} \quad p^{(1)}_{i_1} \neq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \dots \quad (2.2)$$

for $1 \leqslant i_1 \leqslant N$, and $0 \leqslant i_2 \leqslant N$.

Since $$S'_{i_1} = \bigcup_{i_2=0}^{N} S'_{i_1 i_2},$$

it is clear that $$\sum_{i_2=0}^{N} \left\{ p'^{(2)}_{i_2} \mid (i_1) \right\} = 1 \qquad \text{for } 1 \leqslant i_1 \leqslant N,$$

so that (2.2) completely defines the probabilities for selection in the second draw for all possible outcomes of the first draw. We observe here that the null unit cannot be the first unit of a sample s' since s' is the augmentation of a sample 's' of the design D.

Similarly, we define the probabilities of selection in the third draw by

$$\left\{ p'^{(3)}_{i_3} \mid (i_1 i_2) \right\} = \begin{cases} \dfrac{\sum_{s' \in S'_{i_1 i_2 i_3}} P_{s'}}{\sum_{s' \in S'_{i_1 i_2}} P_{s'}} & \text{if } \sum_{s' \in S'_{i_1 i_2}} P_{s'} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where the notation is clear. As before we have

$$\sum_{i_3=0}^{N} \left\{ p'^{(3)}_{i_3} \mid (i_1, i_2) \right\} = 1 \qquad \text{for } \; 1 \leqslant i_1 \leqslant N$$
$$0 \leqslant i_2, i_3 \leqslant N.$$

This process can be continued until the m-th stage where it stops finally. This gives a well-defined sampling scheme giving rise to the design D' which is equivalent to the design D. That there is just one scheme giving rise to D' is clear because at each stage all the conditional probabilities $p'^{(n)}_{i_n} \mid (i_1, i_2, \ldots, i_{n-1})$ should agree for both the schemes. This completes the proof of our assertion.

Remark 1: The introduction of null unit in the population ensures that at each draw, we deal with a probability measure on the population

$$U_0, U_1, \ldots, U_N.$$

This removes undesirable ambiguity in some cases where we come across a draw in which there is a positive probability of no unit getting selected. For example, consider the population of 3 units

$$U_1, U_2 \text{ and } U_3.$$

Let S be the following set of samples

(U_1); (U_2, U_1); (U_1, U_2); (U_3, U_2) and (U_1, U_2, U_3), with probabilities 1/5 attached to each sample. Then

$$\sum_{i=1}^{3} \{ p_i^{(2)} \mid (i) \} = \frac{2}{5} \Big/ \frac{3}{5} = \frac{2}{3}.$$

and with a probability $\frac{2}{5}$ we do not get any unit into the sample, which is an ambiguous situation. However, introducing the null unit U_0, the sample (U_1, U_0, U_0) carries a probability $\frac{1}{5}$ so that

$$\sum_{i=0}^{3} \{p_i^{(2)} | (1)\} = \frac{3}{5} \bigg/ \frac{3}{5} = 1,$$

as it should be.

Remark 2 : We have seen above that for any general design there is a unique scheme which results in that design. We may call this scheme, "*the generating scheme of D*" and D the "*generated design*" of the scheme. However complicated a sampling design may be, we can always consider, conceptually at least, a scheme of drawing units one by one, which gives rise to the given design. Thus any sampling method which does not satisfy the definition of sampling scheme as given in Section 1, can be treated as equivalent to a suitably chosen sampling scheme.

Acknowledgements

The author wishes to thank Dr. V. P. Godambe for suggesting this problem, Dr. J. Roy for his helpful criticism and the referee for his help in the presentation of the paper.

Paper received : October, 1960.
Revised : March, 1962.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. D. Basu	Dr. C. R. Rao
Dr. R. G. Laha	Dr. J. Roy
Shri M. N. Murthy	Dr. D. K. Roychoudhury
Shri P. K. Pathak	Dr. J. Sethuraman

Editor,
Sankhyā : The Indian Journal of Statistics.