Economic and Social Review, Vol. 10, No. 2, January, 1979, pp. 169-174.

Notes and Comments

RANSAM: A Random Sample Design for Ireland

BRENDAN J. WHELAN
The Economic and Social Research Institute, Dublin

I INTRODUCTION

The purpose of this paper is to describe RANSAM, the computer-based system for drawing random samples from the Electoral Register, which has been developed over the past few years at ESRI.[1] This system, which is updated each year by reference to the most recent Electoral Register, permits us to offer a sample selection service to those contemplating carrying out a survey on a local (e.g., regional) or national level. The computer costs in selecting samples are quite low, and the names and addresses can by typed at reasonable cost. Such samples have the advantage that we can ensure that the selected respondents have not cropped up in any of the surveys conducted by the ESRI in the current year.

This paper first discusses the use of the Electoral Register as a sampling frame. It then goes on to describe the main features of RANSAM and the reasons why these features have been included. It should be noted that throughout the paper we are concerned with random or probability samples (i.e., samples in which the probability of inclusion of each element is calcul-

1. In designing the system, I have benefited from the ideas and experiences of several research workers who have drawn samples at the ESRI in the past ten years. The contributions by J. O'Meara, C. A. Ó Muircheartaigh and R. Wiggins have been especially helpful.

able). Non-random samples (e.g., quota samples) can also be based on the present selection system. For instance, we sometimes use RANSAM to select a random sample of names and addresses as starting points for the interviewers and instruct them to call on the selected respondent and on, say, every twentieth house thereafter.

II THE ELECTORAL REGISTER AS A SAMPLING FRAME

The Electoral Register, which comes into force on 15 April of each year, lists the names and addresses of all those eligible to vote in Dail elections (resident citizens) and those eligible to vote in local government elections (all residents). Persons eligible to vote in local government elections only are indicated by the letter L. The Register is compiled by the franchise departments of the county councils and county borough councils.

The Register is published in the form of "books", each relating to "polling districts", in which the electors are numbered sequentially from 1 to n, where n is the total number of people in the book. The county, constituency, and District Electoral Divisions (DEDs) in which the polling district is located are shown. In general, the polling districts are not sub-divisions of the DEDs – one polling district may contain parts of several DEDs. This is not the case in the urban areas of Cork and Dublin where there is a one-to-one correspondence between DEDs (or wards) and the polling districts.

Kish (1965, p. 53) describes a perfect sampling frame as one in which "every element appears on the list separately once, only once, and nothing else appears on the list". Measured against this yardstick, the Irish Electoral Register has certain deficiencies.

Age Restriction: Under current law, the Register excludes those under 18 years of age. Thus, a sample which includes persons under 18 cannot be obtained directly from the Register.

Missing Elements: Individuals over 18 who are resident in the country may fail to appear on the Register. Most of these are people who have only just reached their 18th birthday or who have only recently moved into the district.

Superfluous Elements: The Register includes the names of some deceased electors and of some who have left the district without cancelling their previous registration.

Despite these difficulties, the Electoral Register is still the most appropriate list for use as a sampling frame to select representative samples at reasonable cost.

Each year in May, the Survey Unit of ESRI acquires the complete Electoral Register. A computer card is punched for each polling district (or book), giving the following data:

(a) county,
(b) constituency,
(c) polling district name and identification code,
(d) District Electoral Division(s) into which the polling district falls,
(e) map reference of one of the DEDs into which polling district falls,
(f) number of electors in polling district,
(g) (for Dublin): a ranking of the ward (DED) by percentage of males in unskilled occupations (1971) Census),
(h) (for Dublin): a ranking of the ward (DED) by percentage of households owning cars (1971 Census),
(i) (for Dublin): a ranking of the ward (DED) by an overall socioeconomic status index (based on 1971 Census).

These cards are sorted by alphabetical order of constituency. Within each constituency, the order of the cards is determined in the manner described below under the heading "Stratification". By using these cards as an input into the RANSAM program a national random sample is obtained in the form of a set of polling districts (books) and the registration numbers of the selected electors within these books.

III MAIN FEATURES OF RANSAM

It would, of course, be possible to select a simple random sample (srs) from the two million names on the Register. However, this has many disadvantages such as high cost, impractical scattering of respondents and higher sampling variances than could be achieved by other methods of sampling. Samples selected by RANSAM incorporate three main modifications which distinguish them from simple random samples: (i) multi-stage sampling, (ii) selection with probability proportional to size and (iii) stratification. We now consider each of these in turn.

Multi-stage Sampling

Instead of taking a sample of individuals from the Register, RANSAM first selects a set of primary sampling units (PSUs) and then selects a random sample of individuals within each of the PSUs. The PSUs used are the polling districts into which the Electoral Register is divided. In general, this type of sample will be expected to be less efficient than a srs, but by concentrating the selected individuals into certain geographical areas, one can reduce costs to such an extent as to more than justify the decrease in efficiency.

To illustrate the procedure, consider the sample for the October 1978 EEC Consumer Survey. It was desired to obtain 5,000 completed questionnaires. In order to allow for non-response of various sorts, it was decided to select 6,500 names from the Electoral Register. Thus, 162 polling districts

throughout the country were selected and 40 electors selected from each district. Each interviewer in the survey was allocated one or two of these districts.[2]

The polling districts vary enormously in size – from a few dozen electors to several thousand. A problem may, therefore, arise in small polling districts in that two or more people from the same household may crop up in the sample. This is undesirable, both from the interviewer's point of view and because the intra-household correlation of response is likely to be high. (See Cochran, 1963, p. 210, on this problem.) RANSAM avoids this difficulty by allowing the user to specify a minimum size for the PSUs. Any polling district below this size is amalgamated with a contiguous one to form a combined PSU greater than the required size. A typical value of the minimum size would be about 500.

Probability Proportional to Size

An important feature of the RANSAM design is that it is "epsem" (i.e., each element (elector) in the population has an equal probability of selection). Epsem procedures have some important advantages, notably the fact that the estimates derived from such a sample are self-weighting (see Kish, 1965, pp. 20-22).

In order to show why the sample selected is epsem, let us examine how the sampling procedure operates. To use the procedure one specifies

n = the (constant) number of electors to be chosen from each PSU,
k = the number of PSUs to be selected,
m = the minimum desired PSU size.

The program will then decide the number of PSUs to be selected from each constituency by forming a cumulative list of the populations of the constituencies and sampling systematically from a random start. Within each constituency, it then amalgamates polling districts below the desired minimum size with contiguous polling districts. Next, it selects the polling districts to be included in the sample by forming a cumulative list of the populations of the polling districts and sampling systematically from a random start. Thus, the probability of a given PSU being selected is proportional to its size. For each polling district selected, a random start and an interval are then calculated and the registration numbers of the n selected electors listed. Finally, the names and addresses of the selected electors are found and typed.

2. This means that interviewer effects are confounded with the effects attributable to the polling districts, but cost considerations usually rule out other strategies such as randomisation of interviewers across polling districts or the allocation of several interviewers to one district.

Thus, the probability of selecting the ith element of the jth PSU is

$$\Pr(E_{ij}) = \Pr(PSU_j) \cdot \Pr(E_{ij} | \text{ PSU}_j \text{ selected})$$

$$= \frac{N_j}{\Sigma_j N_j} \cdot \frac{n}{N_j} = \frac{n}{\Sigma_j N_j} = \text{constant}$$

where N_j is the size of the jth PSU, and n the constant number of elements selected from each PSU. Hence, the probability of selection of each element in the population is equal.

Stratification

This means that the population is divided into sub-groups or strata and a srs is selected from each stratum. This procedure will almost always result in an improvement in precision (Cochran, 1963, p. 99). Stratification is incorporated in RANSAM only at the first stage, all PSUs being stratified by geographical location.

In order to stratify the PSUs geographically, one would ideally like to have maps of the polling districts. Unfortunately, no such maps are available, so one must utilise the maps of the DEDs available from the Ordnance Survey. Given the lack of one-to-one correspondence between polling districts and DEDs, these maps only give the approximate location of the polling districts. Furthermore, when one polling district contains parts of several DEDs, it is necessary to identify each polling district uniquely with one of the DEDs in some way. The rather arbitrary rule which we adopted to do this was to identify the polling district with the DED which comes alphabetically first. Though far from perfect, this rule allows one to locate the polling districts approximately, and this is really all that is required for the purposes of stratification.

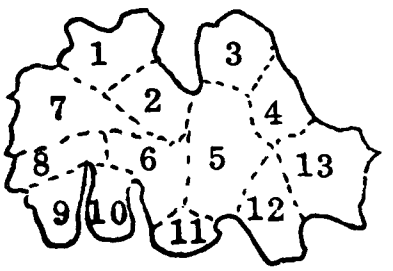
The order of the DEDs within each constituency is determined as follows. Each DED is numbered from 1 to t, where t is the total number of DEDs in the constituency. The numbers are allocated by starting in the extreme north of the constituency, moving west to east, then east to west and so on (see Figure 1). Ordering the polling districts by this numbering system and selecting a systematic random sample means that those selected are stratified geographically across the constituency.

IV FUTURE DEVELOPMENT OF RANSAM

To date, RANSAM has been used to select samples for some twenty surveys and seems to produce sample estimates which correspond well to known data on such variables as age, sex, occupational status, etc. On the whole, the researchers who have used these samples seem satisfied with them.

Figure 1

Constituency boundary _____
DED boundary _ _ _ _ _ _ _ _



It is hoped to develop the system in several ways over the next few years. Work is already at an advanced stage to link the Electoral Register with the Small Area data from the Census of Population. This would permit one to stratify the PSUs by a number of variables (such as age structure, socio-economic status, etc.) in addition to the simple geographical stratification now incorporated in the system. A slightly more long-term objective is the calculation of typical standard errors and design effects for samples drawn by RANSAM. Several surveys already exist at ESRI on which such calculations could be based and more are planned for the future.

REFERENCES

COCHRAN, W. G., 1963. *Sampling Techniques*, New York: Wiley.
KISH, L., 1965. *Survey Sampling*, New York: Wiley.