Chapter (non-refereed)

Lakhani, K. H.. 1983 Strategy for successful survey. In: Fuller, R. M., (ed.) Ecological mapping from ground, air and space. NERC/ITE, 8-15. (ITE Symposium, 10).

II STATISTICAL CONSIDERATIONS

STRATEGY FOR SUCCESSFUL SURVEY

K H LAKHANI
Institute of Terrestrial Ecology, Monks Wood Experimental Station, Huntingdon

ABSTRACT

The scientist engaged on an observational programme should seek to obtain the maximum amount of relevant information for minimum cost. The conclusions should be clear and precise, and seen to be based on acceptable scientific method. Thus, the scientist should be familiar not only with the observational material but also with the essentials of sampling techniques, experimental designs and statistical methodology. The literature in these areas is extensive and growing, requiring a close collaboration between the scientist and the statistician. This paper gives an introduction to the main steps underlying a survey.

INTRODUCTION

Our knowledge, opinions and attitudes are based on our limited experience and interactions with other people and events. Events in nature - the very phenomena of birth, of living and experiencing, even death - are the results of complex sampling processes. In everyday life, the many decisions that we make are based on samples of past experience.

What is true in everyday life is equally true in scientific investigations. Quite a lot of the complex interactions occurring in nature might be intrinsically uncertain and beyond our myopic comprehension. Even if the universe was deterministic, we have neither the ability nor the resources to assess, with complete precision, all relevant particular considerations; and we find ourselves forced to describe and explain from a limited amount of information.

In any given study, the extent of our success depends upon the nature of the population being studied, the variable of interest, the sampling design and the size of the sample, the accuracy of the measuring methods, and the appropriateness and the adequacy of the mathematical and analytical methodology used. The scientist, engaged to carry out a survey, is thus placed in an almost impossible position: not only must he have expert biological and ecological knowlege of the nature of the observational material, he must also have a detailed understanding of the subject of the design and analysis of experiments, and also of the sampling methods for censuses and surveys. There are, of course, many text books on these subjects: a selected sample is listed at the end of this paper. However, the text books tend to be highly technical. For most scientists, an optimum policy is, perhaps, not to attempt to master the books, but to gain familiarity with the basic concepts underlying the subjects of sampling methods and experimental designs, and, then, to seek out a friendly statistician who is interested in the scientist's problems, and who is willing to act collaboratively and collectively with the scientist.

ESSENTIAL CONSIDERATIONS FOR A SUCCESSFUL SURVEY

It is obviously naive to suppose that it is possible to provide, in so short a paper as this one, a comprehensive guide to the principles of statistical methods and sampling design. This does not preclude, however, the presentation of a brief account, highlighting the importance of some of the considerations. Thus, a good summary of the basic principles is given in the first 64 pages of Green's (1979) book *Sampling design and statistical methods for environmental biologists*. The statistical checklists by Jeffers (1978 & 1979) on *Design of experiments* and on *Sampling* cover only a few pages, but succeed in getting across to biologists and statisticians alike the important questions which must be taken into account in the application of statistical and sampling methods to practical research. The present paper discusses a number of features which are common to most investigations (see Cochran 1963). In organizing a new survey, the scientist should consider carefully the following features.

Objectives

It is essential that the objectives of the survey are clearly defined and well known to all staff participating in the study. If this is not done, it is very easy, during the later stages of planning or measuring or analysing, to make decisions which do not quite agree with the objectives. The degree to which the investigation succeeds depends, of course, on how one measures success, but an obvious criterion is to assess the degree to which the specified objectives were attained. An additional advantage of clearly set out objectives is that their very existence presents a challenge to the team of workers to direct their collective creative efforts towards the achievement of the aims of the exercise.

Population

The population which is being studied, and to which the conclusions from the survey will be applicable, must be clearly defined. The word *population* refers not to a biological population of the members of a given species, but to the complete collection of all individual entities of interest, eg the population of all books in a given library.

In an ecological mapping survey, the interest may be in the population of all non-overlapping quadrats, say 1 km^2, making up the study area. It is not always straightforward to identify clearly the collection of interest. Apart from boundary problems, there may be features within the study area (eg lakes, towns, railway lines, etc) which might be thought of as not belonging to the population. Even with a detailed definition, borderline cases are bound to arise, and it will be necessary to have clear rules enabling the field worker to decide, without much hesitation, whether or not the borderline case belongs to the population.

Since conclusions drawn from a sample apply only to the *sampled population*, it is necessary for the sample population to coincide with the *target population*, ie the population about which information is required. For example, a survey of the invertebrates of a given locality based on light trap catching methods may be misleading not only because the traps might succeed in catching invertebrates from other nearby localities, but also because the traps might fail to catch in appropriate proportions the different species which are in fact present, but are differentially active during the night. The extent to which conclusions from such a sample population apply to the target population

must then depend upon other supplementary information which may not always be available readily.

Sampling units

Before the sample is selected, the population must be divided into parts called *sampling units*. For a biological population made up of all members of a given species, the sampling unit may be the individual member. Equally, the sampling unit may be a quadrat of a given size, the measurement of interest being, perhaps, the individuals found in a given quadrat. In a large scale mapping programme, the quadrat may be quite large, and the only practical measurement may be the recording of presence or absence of a given attribute.

Data

It is necessary to check that the data to be collected are *sufficient* and *relevant* for the objectives of the survey. A preliminary consideration of how the data will be analysed and presented may draw attention not only to the additional variables which must also be measured, but also to the possibility of discarding some of the variables which may be superfluous to the study.

Degree of accuracy

The results of the survey will be subject to some uncertainty because of the chance variation inherent in the sampling procedure and because of the errors of measurement. It is tempting, therefore, to increase the sample size and to improve the instruments and the procedures of measurement. These actions cost time and money. If the extent of error that can be tolerated is specified, it may be possible to recommend an economic sampling procedure, consistent with drawing reliable conclusions from the survey.

Sampling scheme

For most surveys there will be a number of ways in which the sample may be selected. To illustrate a few simple sampling schemes, we will suppose that the basic sampling unit is the 1 km^2 of the Ordnance Survey National Grid, and that the study area is made up of N such squares, numbered 1, 2, 3, ...,N.

Systematic sampling is obtained by randomly locating a grid of points over the study area. Such a scheme is convenient to use, and ensures that the sample points are well spread over the whole study area, but has a weakness that, if there is any periodicity in the population, and, if the sampling coincides with this periodicity, then the particular sample estimate may be grossly in error, and it will have a misleadingly small standard error.

A simple random sample gives each sampling unit the same chance of being included in the sample, regardless of which other units have already been selected. Given a population of N grid squares, a random sample can be obtained by drawing numbered tickets, which have been well mixed in a large hat, or by using a table of random numbers or by programming a computer to provide a random selection of numbers.

Relative to systematic sampling, simple random sampling requires greater care and work, but the advantage is that, if the data arising out of the mapping survey are used to estimate the population parameters of interest to the ecologist, the resulting sample estimates will be unbiased, and the calculated standard errors will not be misleading.

Stratified random sampling is the application of simple random sampling, not within the entire population of, say, N units, but separately within subpopulations of N_1, N_2, N_3, . . . etc, units. These subpopulations are non-overlapping and together they comprise the whole of the population, so that $N_1+N_2+N_3$ ...=N. These subpopulations are called *strata*, and it is important to bear in mind that, to obtain the full benefit from stratification, the subpopulation sizes N_1, N_2, N_3 . . . etc, must be known. There are a number of reasons why we may use *stratified random sampling*:

1. We may simply wish to also obtain separate estimates for the different subdivisions of the population. This makes it necessary to treat the subdivisions as if they were separate populations. Thus, a national survey of Great Britain may be subdivided into England, Wales and Scotland, or into other geographical or environmental regions.

2. It may be convenient administratively to subdivide the population of interest. Thus, a national survey may be easier to conduct if the population is subdivided by regions or counties or other areas, with the regional staff carrying out the sampling locally. Care must be taken to provide, and enforce, adequate training for the active workers in the different regions. The training should be centralised, and the field workers' performance should be assessed to verify that their observational ability is of uniform and comparable standard. If this is not done, it will be difficult to distinguish genuine differences between regions from spurious differences brought about as a result of regional differences in the expertise of the field workers from different regions.

3. Stratification may improve the precision of the estimates. It may be possible to divide a heterogeneous population into strata (subpopulations), such that each stratum is reasonably homogeneous, in the sense that the measurements vary little from one unit to another within a stratum. Each stratum mean can then be estimated precisely from only a small sample in that stratum. The estimates for the different strata can be combined to obtain a precise estimate for the whole population. For example, in surveying the vegetation of a large area, it may be reasonable to suppose that, within each square, the values of the variables of interest would depend on the site characteristics such as rainfall, altitude, soil types and slope. If so, before sampling begins, the population of grid squares should be stratified (ie classified) by these factors.

If stratification is based on, say, k *factors*, and if the i^{th} factor has n_i levels, then the total number of strata will be the product of all the levels, $n_1 \times n_2 \times n_3 \times \ldots \times n_k$.

The theory of stratified sampling deals with the properties of the estimates from a stratified sample, and can be used to design the allocation of the sampling effort to different strata, so that the precision of the estimates will be maximized. Thus, the sampling effort to be allocated to a given stratum will be determined by the size of the stratum, the variability within the stratum and also on the cost of sampling, in terms of money or effort, within the stratum. The stratum variability and the cost of sampling in a stratum may not be known before sampling begins; but sampling can be controlled to take place sequentially, so that, for example, an initial allocation of a part of the total sampling effort may be made proportional to the stratum size, and the information on the variability and the cost of sampling within the strata, gained from the initial sampling effort, can be used to determine the allocation of the remaining sampling effort.

Apart from the 3 sampling schemes (systematic, simple random and stratified random), there are also other sampling designs (eg multi-stage sampling and

cluster sampling) which are also frequently used. The sampling theory for surveys is mainly concerned about efficiently estimating the population parameters from the sample data.

Not all surveys are *analytical*, however, and, if the main objective of a survey is to produce a map or maps, then it can be argued that most of the sampling methodology in statistical literature is immaterial for achieving the goals of such a wholly descriptive survey. For ecological surveys, such an argument is short-sighted and unacceptable. Initially, the aim might well be to just describe. However, as soon as this phase is anywhere near completion, the ecologist would want to make comparisons between different subgroups of the population, with a view to discovering whether differences exist among them that may enable him to form or verify hypotheses about the processes at work in nature.

The recent advances in automatic recording techniques, and storage of almost unlimited amount of information on modern computer systems, together with our ability to tackle elaborate computational problems, mean that a purely descriptive ecological survey is now very much a thing of the past.

Methods of measurement

The sampling design defines the exact procedure to be followed for the selection of the sample units. Having identified a given sampling unit, it is obviously necessary to be clear as to what attributes or variables are to be measured or counted in each sample unit, and also how these measurements are to be made. A countable variable such as 'the number of plants in a quadrat' may present no difficulty if the individual plants are readily identifiable. On the other hand, the variable 'the number of earthworms in a quadrat' will generally mean not only the earthworms on the surface of the ground but also those underground. The numbers actually observed would then depend upon the extraction procedure used. The efficiency of the extraction procedure may itself depend on the earthworm activity which in turn may depend on variables such as soil temperature and soil moisture content.

Similarly, our ability to see and count other fauna will depend also on the size and colour of the fauna, and upon their mobility, abundance and behaviour. Data arising from pitfall traps, light-traps or suction sampling devices may not be easy to interpret. The green carpet of prevailing vegetation may produce ever-changing patterns through the seasons, and the insect species' variety and abundance may follow regular rhythms requiring great care in making decisions about the timing of field measurements.

Having made sure that the measurements are meaningful, relevant and adequate for the objectives of the survey, they need to be systematically recorded in a well-designed record form, which should be designed to also encourage the recording of unexpected occurrences in the field.

A pilot study

Unexpected occurrences always occur, and are, in a sense, expected. It is a good policy, therefore, to try out a preliminary field study on a small scale. This nearly always results in improvements in the record forms, may lead to modifications in the measuring methods, and may reveal other difficulties, such as a revised higher estimate of the total cost. Preliminary information, about the cost of sampling and relative variability in the different parts of the population, may provide a valuable basis for modifying the design of the sampling scheme, or the sampling allocation within the existing scheme.

Analysis of data

The type of analysis of the data will depend upon the procedure used for data collection, ie the sampling design and the method of measurement, the aim of the analysis being to achieve the objectives of the study. Preliminary arrangements for data handling on a suitable computer, including the development of routine computer programs, could be made before the sampling begins.

However, before rushing into any large scale 'number-crunching' exercise, it will be an advantage to discuss the analysis with an interested statistician, who should also be kept informed of the progress of the analysis. Statistical methodology is developing rapidly and continuously, and it is just possible that improved or alternative methodological techniques might have been publicized in the statistical literature since the initiation of the study. It is often the case that the proposed analysis of the observations requires that the data satisfy certain assumptions about the form of the distribution of the variables. The statistician, being aware of the assumptions, is often better placed to carry out initial checks of these assumptions, and, if necessary, to suggest appropriate transformations of the observations.

A preliminary examination may also indicate certain aspects of the data which may suggest modifications of the earlier decisions about the analysis. For example, it may be decided, after the data have been collected from a simple random sampling scheme, to carry out a stratification of the data. For example, in studying the rates of hedge removal from the interpretations of aerial photographs, Hooper (1968) classified his randomly selected study areas into arable, mixed and dairying farm types. Such a post-stratification amounts to correcting, to some extent, the earlier omission of not having foreseen the desirability of stratifying.

At times post-stratification may be even more efficient than stratification before sampling begins, because, after the sampling has been carried out, the stratification factors can be chosen in different ways, for different sets of variables, in order to maximize the gains in precision. This technique is particularly useful in multipurpose surveys where stratification factors selected before sampling may be poorly correlated with large numbers of secondary variables. Situations also arise where identifiable subgroups of the population do exist, but individual members can only be classified after sample selection. Thus, quadrats already observed may be classified by their soil pH measurements, but in practice this can only be done after the sample selection.

Presentation of results

The whole purpose of the statistical and numerical methodology is to isolate and describe the main trends and variations in the data, with a view to allowing and enabling the data to tell their own story. The results obtained, and also the data or their summary, need to be so presented that they leap to the eye. This condensation of detailed information into tables, graphs, charts and relationships is also part of statistical methods, and so, the statistician should be able to play a prominent part in the presentation of results.

Information gained for future surveys

A difficulty underlying the design of a sampling scheme for any population is that the choice of the scheme depends on the initial information we have about the population. As noted earlier, in stratified random sampling, the allocation of the sampling effort to different strata can be improved upon by an element of

sequential sampling in the scheme. In the same way, it is obvious that any completed study is potentially a good guide for a similar future study. Thus, information about various estimates and their standard errors, and the cost of sampling in the earlier study, can be used to make judgements about sampling design, sample size and costs in a later study. In addition, a knowledge of the type of unexpected occurrences, and mistakes made in the earlier study are also of obvious value to the future worker. It is important, therefore, to assemble, record and make available all such information for future surveys.

GENERAL COMMENTS

1. Unlike experiments, in which treatments are allocated to randomly selected plots and other variables are controlled as far as possible, the data arising from a survey are purely observational in nature, and require care and caution in analysis and interpretation. Thus, in an experiment, the observed differences may readily be attributed to the differences in treatments. On the other hand, in data arising from a survey, the observed differences in sampling units from different levels of a factor may well not be due to that particular factor, but to some other unknown but correlated factor.

2. Care must also be exercised when the survey, perhaps because of its sheer size, is spread over a number of years. If different geographic regions are observed over different years, the apparent regional differences may well be due to the differential effects of the different calendar years on the variables of interest. In other words, using statistical language, the regional (spatial) effects may be *confounded* with the yearly (temporal) effects. A proper design would, of course, include a number of control units, to be observed each year. Such additional observations would be valuable to disentangle the regional effects from the yearly effects.

3. Finally, it must be borne in mind that a survey, being a programme of observations at one time point, is like a single snapshot of nature. It will describe the status of the variables as measured at a given time point, and it will describe the prevailing relationships between different variables at the time of the survey; but, the survey will not, by itself, provide information on the rates of change in the values of the variables, in time.

CONCLUDING REMARKS

To summarize, ecological populations are complex, and their study requires great care and ability on the part of the scientist. Usually, it is neither possible nor wise to attempt to study the entire population. If the programme of observations is properly planned and conducted, the inductive approach of the statistical theory makes it possible to make probabilistic statements about the whole population, having observed only a part of it, a suitable sample of it.

To put it another way, nature does not yield her secrets readily - a price has to be paid in terms of manpower and resources. We would like to see and describe all of nature, in all her glorious and detailed contours and colours. This, alas, we cannot do. We can, however, have a peep at her, through the keyhole as it were, and, having seen a bit of her, create a complete image of her.

REFERENCES

COCHRAN, W.G. 1963. *Sampling techniques*. New York, London: Wiley.

GREEN, R.H. 1979. *Sampling design and statistical methods for environmental biologists*. New York: Wiley.

HOOPER, M.D. 1968. The rates of hedgerow removal. In: *Hedges and hedgerow trees*, edited by M.D. Hooper and M.W. Holdgate, 9-11. (Monks Wood Symposium no. 4). Abbots Ripton: Monks Wood Experimental Station.

JEFFERS, J.N.R. 1978. *Design of experiments*. (Statistical checklist 1). Cambridge: Institute of Terrestrial Ecology.

JEFFERS, J.N.R. 1979. *Sampling*. (Statistical checklist 2). Cambridge: Institute of Terrestrial Ecology.

SELECTED TEXT BOOKS

COCHRAN, W.G. & COX, G.M. 1950. *Experimental designs*. New York: Wiley.

COX, D.R. 1958. *Planning of experiments*. New York: Wiley.

FISHER, R.A. 1935. *The design of experiments*. Edinburgh: Oliver & Boyd.

SAMPFORD, M.R. 1962. *Introduction to sampling theory*. Edinburgh: Oliver & Boyd.

SCHEFFÉ, H. 1959. *The analysis of variance*. New York: Wiley.

SUKHATME, P.V. 1954. *Sampling theory of surveys with applications*. Iowa: Iowa State College Press.

YATES, F. 1960. *Sampling methods for censuses and surveys*. 3rd ed. London: Griffin.