Sampling Systems for Dependent Data

R. M. WILLIAMS,
Applied Mathematics Laboratory, D.S.I.R., Wellington.

THIS paper gives a general and necessarily incomplete survey of some of the recent advances in the theory of sampling which have some use in research work in animal production.

Any sampling scheme is simply a plan to measure a limited part of some population (e.g. a flock of sheep, the grass in a field or the daily milk yield of a cow) in order to obtain an estimate of some quantity such as the average weight of the sheep in the flock or the average daily yield of the milk by the cow throughout her lactation. One may also measure such things as the correlation between two or more factors, but we shall confine ourselves to the case of the simple average. Such methods have been widely employed, as for instance in herd testing, but the theory (except for the most straight forward cases) has been slow to develop, much of it having been done during and since the war. The advances that have been made do not always suggest very original methods of sampling, but they do help us to understand the efficiency of different sampling schemes under different circumstances and so to choose the one that is most appropriate.

The commonest reasons for using a sampling scheme are because it provides an estimate sufficiently accurate for the purpose and is much more economical than a full enumeration, or because we wish to make measurements halfway through an experiment which, for instance, involve killing those animals on which the measurements are being made and keeping the rest of the flock for later measurements. In either case it is for the experimentalist to decide what accuracy is required, and for the statistician to try to help him find the most economical scheme which provides it. Ideally such a scheme should meet the following requirements:

1. It should be unbiased i.e. if the sampling procedure is repeated sufficiently often the average of the estimates obtained will approach the true value as closely as required.
2. It should provide some estimate of its accuracy.
3. It is desirable that besides knowing how accurate an existing scheme is, we should be able to predict roughly how the accuracy will change if we make changes in, say, the density of sampling or minor changes in the design.
4. It must be efficient in the sense of providing the required accuracy with the least amount of work.
5. It must be simple to carry out.

If the variations of the property to be measured are completely independent of the relative positions in, say, the plot of grass whose production is to be measured then all these conditions are fulfilled by placing the sampling plots systematically over the area. This condition is, however, very rarely fulfilled in biological data. In the overwhelming majority of cases the difference likely to occur between any pair of observations does depend on their relative positions in the field, or the relative times at which the measurements are made e.g. measurements of milk yield made on consecutive days may be expected to differ on the average less than those made a week apart. In this case, if we ensure that the observations are chosen randomly with respect to this factor i.e. position when measuring in a field, time when measuring in lactation, we ensure that the first three conditions are fulfilled, but since a properly randomised design is in general harder to carry out (particularly with unskilled labour) than a systematic

one, we sacrifice much in simplicity and economy of execution, and in the great majority of cases occurring in agricultural work we sacrifice a lot in accuracy.

No doubt it is this fact of the greater accuracy that can often be achieved by systematic sampling schemes or schemes with limited randomisation, that has led to their general use even when their theory has been little understood. The object of theoretical work has been to formulate general conditions covering a wide range of practical cases, under which we can recover at least some information about the accuracy of such designs.

The most obvious advance from purely random schemes is the stratified random scheme under which the region to be sampled is divided into strata and a number of sampling units allocated to each strata. This raises the question of how best to allocate a given number of sampling units; in a number of experimental surveys it was found that the most efficient system was to divide the region into the same number of strata and to allocate only one sample to each; but this fails to provide us with any estimate of accuracy; a doubling of the stratum size and allocating two sampling units to each gives us an estimate of accuracy, but at the cost of reducing the efficiency. (See for example Yates, 1948, where the first scheme was often at least twice as accurate as the second). A compromise is sometimes made by allocating two units to some of the strata, but one unit to most of them. This provides us with some estimate of accuracy, but is only useful if there are a large number of strata. These general considerations only hold if the variances within each stratum are comparable from strata to strata. If these variances differ greatly from one stratum to the next the greatest efficiency is obtained by making the number of sampling units allocated proportional to the variance of each stratum; but this assumes a prior knowledge of these variances, which is not always available.

We conclude then that stratified random sampling systems are usually more efficient than fully random schemes, although this efficiency can only be fully exploited by sacrificing the estimate of its accuracy. Against it, a stratified design is usually no simpler to carry out than a fully randomised one and is often less efficient than a systematic one.

The systematic sample is drawn by taking sampling units at equal intervals in time (in the case of a lactation curve) or along a line (when sampling pasture); if the region is a two-dimensional one such as a field one may either lay it out as a two-dimensional grid of small sampling units or treat it as a one-dimensional case by taking strips right across it in one direction, each strip being treated as an element in a one-dimensional sample taken at right angles to the first direction.

Provided that the starting point is chosen at random such a systematic system is unbiased; it is easy to carry out; and in most cases appreciably more efficient than a stratified random scheme, particularly where two or more sampling units are required in each stratum (see for example Osborne, 1952, where an example is given of systematic cover survey which was four times as accurate as a stratified random scheme with two units in each stratum and thirty-six times as accurate as a fully randomised scheme). The most notable exception is when the data have a regular periodic variation such as daily or seasonal effects, or such fluctuations as are said to occur in fields subject to regular methods of cultivation. In these cases, if the interval between sampling units is a multiple of the period (or very close to one) the sytematic scheme can be grossly misleading and the stratified random scheme is to be preferred. Disregarding this particular case several special cases have been considered which make

it possible to see how the accuracy may be expected to vary with sampling density and to obtain at least some estimate of its magnitude.

Yates considers the errors which will occur when sampling from a distribution which can be regarded as a smooth curve. A very high degree of accuracy (Yates, 1948) can be obtained, particularly if the quantity considered increases and decreases very slowly at the beginning and end of the region to be sampled; if this is not the case, rather less accuracy is obtainable, and end corrections are required; but it is still satisfactory and the manner in which the sampling variance changes can be given a mathematical form. This case is rather artificial and a more practical model is found if we superimpose on the smooth curve a small random fluctuation. Such a model can lead to the situation where the accuracy increases with the number of units sampling much more rapidly than in the random case; and the application of such a model as this might well lead to a better understanding of changes in the sampling variances of lactation curves with change in sampling frequency. (Dick, 1950).

Yates has also developed a method, applicable both to this and to more general cases, for obtaining not a valid estimate, but at least an upper limit to the sampling variance. This method depends on assumptions which can only be expressed rigorously in mathematical form, but which might be expected to be true in many cases when samples taken at close intervals differ on the average, less than samples taken far part. This assumption is a plausible one except in those cases where there are either periodic or competitive effects. This upper limit can sometimes seriously over-estimate the sampling variance and its accuracy can be greatly improved by a limited number of supplementary sampling units, placed between the main sampling points.

Another model, which has its theoretical basis in the study of time series, makes the assumption that the probability relations which exist between pairs of sampling units depends only on the spacing between the sampling units. (Jowett, 1952).

It is clear that all these assumptions are restrictive, and one may fit where another does not; the systematic examination of various types of biological data to see which, if any, is appropriate should lead to considerable advances in sampling technique.

REFERENCES:

Dick, I. D. (1950). N.Z.J. Sci. Tech. **32** : 25.
Jowett, G. H. (1952). App. Stat. **1** : 50.
Osborne, J. G. (1952). J. Amer. stat. Assn. **37** : 256.
Yates, F. (1948). Phil. Trans. roy. Soc. **241A** : 345.

Discussion

Professor RAE: Is there any similar work on the sampling accuracy of the variances or components of variance, rather than the means?

Dr. WILLIAMS: As far as I know, very little. I think that at this stage the problem could best be tackled by considering a specific set of data.