AREA SAMPLING FOR SOIL CONSERVATION NEEDS

by

J. E. Dowd

PRESENTED to the CONSERVATION NEEDS WORKSHOP

at

BOSTON and PHILADELPHIA

JULY 17 and JULY 20, 1956

Sampling may be defined as that method of gathering information about a specified population or universe wherebÿ the population is divided into a set of mutually exclusive and exhaustive units and a subset of these units are selected to represent the whole population. The significance of the word "area" in the title of this paper is that the units into which the population has been divided are areas of land of a certain pre-determined size.

From the selected units information is then gathered in some detail and expansion of this information is made up to the population level with the hope that the information collected from the sampled areas is representative of conditions occurring in the population. Since in most populations the material about which information is required displays dissimilarities from one unit to the next, it is not the case that conditions occurring in the population can be estimated with perfect accuracy by the expansions made from the sampling units. If there were no dissimilarities among the units (e.g. a county contained only one mapping unit), or if the material in our population were so thoroughly intermixed so that every unit contained the same amount of each type of material found in the county then there would be no problem of sample selection since one unit selected in any manner whatsoever, would supply complete information about the whole population. The one unit would be completely representative of soil and land characteristics in the population.

As everyone here is aware, the material dealt with in the Conservation Needs Inventory is neither similar throughout the population (in our case a county) nor is it thoroughly intermixed. No matter what size of unit the population is divided into, there will exist dissimilarities amongst the units so that nothing less than a 100% sample of the population will be able to estimate all soil breakdowns with complete accuracy. A sample can provide a relatively inexpensive and quick method of providing estimates which will be sufficient for estimating conservation needs although certainly not of sufficient accuracy to meet the needs which a complete soil survey is expected to fulfill.

The definition of sampling stated in part is that "....a subset of these units are selected to represent the whole population". How this subset of units is to be selected is a crucial determinant of the precision which can be achieved by our sample estimates and much of the mathematical theory of sampling is concerned with this question. Certain methods of selection which at first appear to fill the bill, such as deliberate selection of those units which in the opinion of an expert appear to be representative, or the casual or haphazard selection of units with no particular scheme or model in mind, can be shown (and has been shown) to give samples having serious biases in them. Any sample where an individual's judgment is allowed to determine which units shall be selected will consequently reflect the biases of that individual's judgment. This is not to say, of course, that expert opinion and judgment is of no value in selecting a sample, It is just that this judgment and experience should enter the sample survey at other points than the final selection of units to be included in the sample.

The simplest and only universal way of avoiding biases in the selection process is the use of a method whereby every unit in the population has a known and independent probability of entering the sample and each unit is drawn into the sample with this probability. In our case we can assume that all units in the population will be assigned an equal probability of being chose, although in general this need not be the case. When a sample is selected in accordance to the above scheme, the sample is said to be random. It is possible however, to put restrictions upon the randomness of the sample without affecting the validity of the selection. Such a restricted procedure will be discussed shortly.

As an example of a probability sample, suppose we have a county of 400,000 acres and we wish to choose a 4% sample (16,000 acres), and that we have decided to use a sampling unit of 100 acres)giving a sample of 160 units). The procedure would be to divide our county (by means of a grid, for instance) into 4,000 units of 100-acres each, and assign each unit a probability of 1/4,000 of entering the sample. Our next step would then be to choose 160 units from the 4,000, any particular unit having 1 chance in 4,000 of being chosen before the selection begins. We could accomplish this selection by first numbering all the units in any arbitrary fashion, then for each numbered unit placing a ball containing the same number into an urn and mixing the balls thoroughly, Next, 160 of these balls are selected from the urn, and the units corresponding to the balls selected constitute the sample.

Fortunately, it is not necessary to go through this process of drawing balls from the urn, since tables of random numbers generated by electronic machines are available and it is merely necessary to select 160 numbers between 1 and 4,000 from the tables to achieve a random sample.

Whether or not a sample will give results which are sufficiently representative of the population depends upon whether or not errors introduced by the sampling procedure are sufficiently small not to invalidate the results for the purposes for which they are desired. Since the population is composed of dissimilar units the sample cannot be completely representative of the whole population. The errors occurring are termed random sampling errors. Their magnitude will depend upon the size of the sample, the variability of the material sampled, the sampling procedure adopted, and the methods used in calculating the results. Although variability of the material cannot be directly controlled it can be counteracted by increased sample size, by use of certain sampling procedures, and by certain methods of calculating the results which utilize outside information.

One great advantage of using a sampling procedure that is random is the fact that estimates of the sampling error involved (and hence of the precision of the estimate) can be derived from the sample itself. Thus, we are in the position of not only having estimates of the totals for the material in the population, but of also having some idea as to the precision of these estimates. We can use these estimates of precision to make probability statements about the true value we were estimating. As an example of this procedure, suppose our estimate of soil separation x from a 4% sample for a particular county was 10,000 acres, and suppose the sampling error (in terms of the standard diviation of the total) was 2,000 acres, then the chances are roughly 2 to 1 that the true value lies between 8,000 and 12,000 acres, roughly 19 to 1 that the true value lies between 6,000 and 14,000 acres and roughly 99 to 1 that it lies between 4,000 and 16,000 acres.

It must also be stated that the width of the interval within which we expect the true population value to lie is also a function of the degree to which the results have to be broken down and results of the same precision can be achieved by a far smaller sample when one is dealing with a large population then would be necessary if detailed results for different parts of the population are required. Thus, while a two percent sample might give adequate estimates of county totals, it will not provide estimates of the same precision for, say, a watershed within the county. If estimates of the same fixed precision are desired for a smaller subsection of the total population, the sample within that subsection will have to be supplemented.

It has been shown both by means of mathematical statistics and by empirical sampling studies that the sampling error of an estimate varies inversely as the square root of the sample size. Thus in the example cited above, if we desired to reduce the sampling error from 2,000 acres to 1,000 acres, it would be necessary to quadruple the sampling size. If we had have had a 4% sample of 160 units giving us a sampling error of 2,000 acres it would take 640 sampling units to give us a sampling error of 1,000 acres. Obviously, we quickly reach a point where time and resources available limit the amount of reduction in sampling error that can be achieved by increasing the size of the sample.

One other method of decreasing sampling error (a much less expensive way) is to use the judgment of an expert to stratify the population into homogeneous sub-populations (called strata) from which independent random samples will be drawn. This entrance of expert judgment into the sample selection does not cause any bias in the estimate, and, while restricting the randomization to the subpopulations, can achieve some reduction in sampling error. If the same proportion is sampled from each stratum, differences between strata will be eliminated from the sampling error. Large reductions in sampling error can be achieved only where the population can be divided in such a way that large differences in the quantities we are estimating occur between at least some of the strata. It is planned to use strata composed of small compact geographical areas of (in most cases) 4900 acres with the proviso that these strata boundaries be made to coincide with land resource area boundaries wherever the boundaries are deemed to have real meaning as far as differences in conservation needs are concerned.

Besides tending to reduce sampling error, stratification also insures that the sample will be evenly distributed throughout the whole county and that it is possible to get separate estimates for land resource areas falling within the county. This can be achieved simply by combining estimates for all those strata falling within the land resource area.

It should be noted that it is necessary to select at least two sampling units from each stratum in order to get a valid estimate of sampling error, but, good approximations for sampling error can be achieved even if only one sampling unit is located in each stratum.

One other method of reducing sampling error is the use of outside information in calculation of the estimates. Thus, in those counties that have been partially surveyed (either in block mapping, or in individual farm mapping) this information will be used in compiling estimates to help reduce sampling error.

Something should also be said about the choice of the sampling unit to be used in the Northeast. Both were chosen on the basis of information derived from a pilot study carried out on three counties; Livingston and Tioga in New York, and Frederick in Maryland; in all of which there had previously been a completed soil survey. The size of sampling unit chosed was that size which gave the highest precision for a fixed budget. This optimum size of unit depends both upon the distribution of the particular soil separation being estimated and the way in which various costs arise in the sampling, mapping, and measuring processes. Thus it was necessary to determine a sampling unit that would, making specific assumptions about costs, yield an optimum on the average.

While the choice of the sampling rate was an administrative decision in which desired precision had to be balanced against the manpower and funds available to do the job, estimates of the average precision obtainable with different sampling rates, along with some idea of the relative cost of achieving these different rates were obtained from the pilot study.

It was decided in the Northeast to use a 100-acre sampling unit and to draw a 4% sample from each county. In case it is not possible to complete the mapping of a 4% sample within the time set out for completion of the field work and with the manpower available, the sample will be drawn in the form of two independent 2% samples. Thus, if time and manpower make it possible to complete only a 2% sample, valid estimates of population quantities along with estimates of the precision can be obtained from this sample.

In order to examine the accuracy obtainable with a relatively small sample, we shall look at the results obtained by using a 2 1/2% sample of 40-acre units taken from Frederick and Tioga counties. The stratification used here was much coarser than that being used in the survey itself, strata being 16,000 acres rather than 4,900 acres. There was also no attempt made to use land resource areas as a guide to setting up strata.