EDWARD T. O'NEILL

Sampling University Library Collections

Two promising sampling techniques have been developed for large library collections. The first is based on the locations which the books occupy. All possible locations which a volume can occupy are numbered. The resulting sample can then easily be taken from the number locations. This technique has been found applicable to many library sampling problems. In the second method the sample is drawn from the shelf list. For this method the shelflist cards are assigned numbers, and it is the cards that are sampled. This technique is useful only under special circumstances. Both techniques have been used to sample the Purdue University libraries, and the results have been encouraging.

THE ACTIONS AND DECISIONS which are made daily in library operations, as elsewhere, are based on the information available to the decision maker. The better the information on which the decision is based the better the decision will be. Sampling techniques provide a powerful means which the librarian can utilize to improve the information which is available to him.

While sampling techniques can be used to obtain a great variety of information, the methods which will be discussed here will concern only printed library materials. In the design of the sampling plan, several factors must be considered. First, it is necessary to determine exactly the purpose or purposes of the sample. It is also necessary to define precisely the population from which the sample is to be drawn. There are two basic frameworks from which the sample can be drawn: from the locations which the material occupies and from the shelf list. Sampling the locations has been found to be generally more efficient than sampling from the shelf list. The remainder of this paper will deal mainly with simple random location sampling.

Random sampling is, in general, a method of selecting n items out of a total of N items so that each possible sample of size n has an equal chance of being the sample selected. In practice the sample is usually drawn one at a time without replacement. In sampling very large populations, however, the sampling fraction is so small that sampling with replacement can be used with very little loss of precision.

The concept of simple random sampling is easy to understand, and the analysis of the sample is straightforward. The actual sampling, however, can be difficult. Most random sampling procedures require that each of the items in the population of N items be numbered from 1 to N. This requirement makes the selection of the sample relatively easy, but it may be quite difficult to apply such a numbering system to the population. The only way such a system could be implemented for library sampling would be to number individually every item in the population.

Mr. O'Neill is in the School of Industrial Engineering at Purdue University. This paper was read at the 1965 annual meeting of the American Society of Engineering Education in Chicago.

The task of numbering a collection of several hundred thousand volumes would be immense, if not practically impossible.

In library sampling it is desirable to relax the requirement that the items be numbered from 1 to N. All that is required is that each item have a unique number. If a location is defined as any site which is, or could be, occupied by an item belonging to the population being sampled and a unique number is assigned to each location, then each item in the population will have a unique number associated with it. The numbers associated with the locations will be referred to as location numbers. The problem of numbering all N items in the population has been changed to that of assigning a unique number to each of the locations.

While there are several ways to develop the set of location numbers, a nested system seems to be quite satisfactory. Break down the locations as follows: areas, units within areas, sections within units, shelves within sections, and positions within shelves. An area can be any group of shelving units such as one floor of a library or a portion of a floor.[1] A shelving unit as used here refers to a single-faced unit having up to ten sections of shelves, a shelf is just a simple shelf, and a position is an ordered location on a shelf.

To keep the set of possible location numbers to a minimum, it is necessary to establish upper limits on the number of units per area, sections per unit, shelves per section, and positions per shelf. An important criterion in choosing the limits is that they should be an integer power of ten. As will be evident later, this requirement greatly simplifies the numbering system. It has been found that a limit of a hundred units per area seems to be practical. The design of the shelving units usually dictates the limits on the others. It is rare that a single-faced shelving unit would have over ten sections of shelves; therefore ten is a useful limit to set on this value.[2] As for the number of positions per shelf, it is difficult to set an absolute maximum, but for practical purposes a hundred positions per shelf seems to be satisfactory.

The task of relating the set of location numbers to the physical locations is relatively simple. Let the indices "*i, j, k, l, m*" identify the location defined by the *m*th position on the *l*th shelf in the *k*th section of the *j*th unit in the *i*th area. All the locations can then be uniquely described by a set of indices. If A is the location number, and;

$$A = 1{,}000{,}000(i-1) + 10{,}000(j-1) + 1{,}000(k-1) + 100(l-1) + (m-1). \quad (1)$$

The set of indexes are then completely determined from the location number, such that the indexes are the largest integers which satisfy the following conditions:

$$\begin{aligned} i &\leqq [A/1{,}000{,}000] + 1 \\ j &\leqq [(A-1{,}000{,}000i)/10{,}000] + 1 \\ k &\leqq [(A-1{,}000{,}000\,i - 10{,}000\,j)/1{,}000] + 1 \\ l &\leqq [(A-1{,}000{,}000\,i - 10{,}000\,j - 1{,}000\,k)/100] + 1 \\ m &\leqq A - 1{,}000{,}000\,i - 10{,}000\,j - 1{,}000\,k - 100\,l + 1 \quad (2) \end{aligned}$$

Since the location number and the set of location indexes are just different means of representing the same location, they can be used interchangeably.

If ϵ is the number of areas in the collection, the set of location numbers contains all the non-negative integers which are less than $\epsilon \times 10^6$. Since this set of location numbers is much larger than the set of actual location, many of the

[1] Although there is no theoretical limit on the number of areas within the collection, if the number of areas becomes too large it may be practical to define a region which would include a group of areas.

[2] Limiting a shelving unit to no more than ten sections would force a larger unit to be arbitrarily divided into two or more units for sampling purposes.

location numbers will not have any physical location associated with them.

In assigning the location numbers to the locations, the only numbers which must be directly assigned are the area numbers and the unit numbers. This job is relatively easy because the number of areas is usually small, and the units are usually arranged in an orderly fashion within the areas. Some problems are encountered in areas where wall shelving is used since, in this type of shelving the unit is often poorly defined. Once the units are numbered, all the locations within the units can be numbered very simply. An easy way to do this is to number the sections from left to right, the shelves from top to bottom, and the position from left to right.

A location is said to be valid when the location is occupied by an item belonging to the population being sampled. If the location is unoccupied, or if it is occupied by material which is not part of the population being sampled, the location is said to be invalid. A location number is valid if, and only if, the location number represents a valid location. There are two classes of invalid location numbers; (1) the location numbers which have no physical locations associated with them, and (2) the location numbers which represent invalid locations.

What is wanted in random location sampling is a random sample of the valid locations. The invalid locations are of no interest. To select randomly a valid location, location numbers are randomly chosen until a location number is found which represents a valid location. Repeating this process *n* times will yield a random sample of *n* locations.

The numbering system described above includes almost all the possible locations in the library system. To make the numbering system meaningful it was necessary that many invalid location numbers be included; thus a high percentage of the location numbers are invalid. In the Purdue University libraries, for example, it was found that only about 3 per cent of the location numbers were valid.

There is a way to check the location numbers to eliminate a high percentage of the invalid location numbers before the sample is collected. If some of the characteristics of the storage system are known, these can be used to divide the location numbers into two groups; those that are known to be invalid and those that may be valid. The more that is known about the system, the easier it is to classify a location number as invalid. If

α_{ijkl} = the number of positions on the *l*th shelf in the *k*th section of the *j*th unit in the *i*th area,

β_{ijk} = the number of shelves in the *k*th section of the *j*th unit in the *i*th area,

γ_{ij} = the number of sections in the *j*th unit of the *i*th area,

δ_i = the number of units in the *i*th area,

ϵ = the number of areas.

Then for example, if all the α_{ijkl} were known, it would be possible to separate out completely all the location numbers which were associated with unoccupied locations. For the set of location indexes "*i, j, k, l, m,*" if the positions index *m* is greater than α_{ijkl}, then the location number is invalid. The check can be made at several levels as shown in Table 1.

The effectiveness of the prechecking technique decreases as the level increases. This increase in effectiveness is offset,

TABLE 1

Level	Quantities Required	Location Is Invalid If–
I	α_{ijkl}	$m > \alpha_{ijkl}$
II	β_{ijk}	$l > \beta_{ijk}$
III	γ_{ij}	$k > \gamma_{ij}$
IV	δ_i	$j > {}_{i}$
V	ϵ	$i > \epsilon$

however, by the large increase in the amount of information required at the lower level checks. The choice of the level depends on both the size of the sample to be taken and on the physical arrangement of the library collection. Usually the choice is limited to levels II, III, or IV, with level III generally felt to be the most satisfactory.

If the check is made at a level higher than level I, some control is lost since all the variables are not controlled. There is a fairly simple way in which partial control can be gained over the variables not included in the prechecking without requiring too much information. Let

$$\zeta_i = \max_j \left[\max_k \left(\max_l \alpha_{ijkl}\right)\right]$$

$$\eta_i = \max_j \left(\max_k \beta_{ijk}\right)$$

$$\lambda_i = \max_j \gamma_{ij}$$

This yields three more conditions that are required for the location number associated with the location indexes "*i, j, k, l, m*" to be valid:

$$m \leqq \zeta_i$$
$$l \leqq \eta_i$$
$$k \leqq \lambda_i$$

If any of the conditions are not satisfied, then the location is not valid. It is obvious that all or some of these requirements will be redundant depending on what level the check is made.

The technique of checking appears to be quite effective and is easily adaptable to high speed computers. In a large sample taken in the Purdue University libraries, a presort at level III increased the percentage of valid from an estimated 3 per cent before sorting to 42 per cent after the sort. Even if the ζ_i, η_i, and λ_i are only estimates of the true maximums, the checking technique is still highly effective.

Previously only the sampling of shelved materials has been considered. It must be remembered, however, that library materials can be in one of three places; (1) on the shelves of the library, (2) checked out of the library, and (3) in the library but not on the shelves. This third class would include both the material which was actually in use in the library and the material which was waiting to be reshelved.

The techniques developed for sampling shelved materials can be extended to include the materials which are checked out. This requires that each item which is checked out of the library have a location number associated with it. The idea is that every book which is checked out is represented by some form of transaction record. In most libraries, each volume of material which is checked out generates a corresponding transaction record. These transaction records are commonly stored in one or more files of some type. By assigning a location number to each transaction record, the transaction record can be treated as the other locations with one exception. From the transaction record it is only possible to find out what material "occupies" that location. The actual material may still have to be located and physically examined.

The locations have a special meaning when they are part of the checkout file. Generally the checked-out items can be treated as an area, although several areas can be used if required. It is usually convenient to define a file unit as a single drawer. Then the section, the shelf, and the position are effectively just positions within the file drawer. Other interpretations can also be placed on the location number within the file, and the choice of the system to be used is largely a matter of convenience.

The third group of material, that which is in the library but not on the shelf, may present very difficult problems in the assigning of location numbers in open stack libraries. In closed stack libraries this group can be grouped with and treated similarly to the checked-out material. In open stack libraries, there are so many possible locations that it is virtually impossible to assign location numbers to them. A convenient solution is just to exclude this material from the sample. While excluding this material introduces some bias, it may not be significant; and this bias can be reduced by taking the sample at a time when the number of books in this group is smallest, such as during vacation periods or at night.

As an alternate to location sampling, it is also possible to draw the sample from the shelf list. The method of sampling the shelf list is straightforward. It can, however, be very time consuming if a large sample is to be taken unless the required information is recorded in the shelf list. Assume the shelf list consists of *D* drawers with a maximum of *H* cards per drawer. A random sample of the cards can be obtained by randomly selecting a drawer and randomly selecting a position, between *1* and *H* within the drawer. Since *H* is the maximum number of cards per drawer, there will be many cases when the position selected does not exist. When this occurs, a new drawer and position must be selected. If another position within the same drawer is selected, then a bias would be introduced into the sample.

The above procedure will yield a random sample of the cards in the shelf list. Unfortunately, this is not a random sample of all materials covered by the shelf list unless there is a separate card for each volume. This is particularly true of periodicals. Since several volumes are often represented by a single card, a random sample of the card is actually a cluster sample of the volumes. The question often arises as to the necessity of including all the volumes of the cluster in the sample. The way in which the cluster is treated depends largely on both the purpose of the sample and the policy which was used in the construction of the shelf list. It is generally advisable to consult a statistician for advice on this phase of shelflist sampling.

Both the shelf list and the locations sampling have been used to sample at the Purdue University libraries. Each technique has some advantages which under special conditions make it superior to the other technique. However, the shelflist sampling was found to be much more time consuming than location sampling. The use of the shelflist sampling would be recommended only under special circumstances. ■■