M. CARL DROTT

Random Sampling: a Tool for Library Research

Questions about the accuracy of library records, the behavior or attitudes of patrons, or the conditions of the books in the collection can often be answered by a random sampling study. Use of this time and money saving technique requires no special mathematical ability or statistical background. The concept of accuracy is discussed and a table is provided to simplify the determination of an appropriate sample size. A method of selecting a sample using random numbers is shown. Three examples illustrate the application of the technique to library problems.

LIBRARIANS ARE continually called upon to make decisions based on imperfect data. How many books in the collection are in need of repair? What percentage of our patrons use the card catalog? Which categories of books can be stored without greatly inconveniencing our patrons? The cost of keeping accurate records to answer this type of question is great. On the other hand the librarian should have something better than an informed guess. One way of providing this information is to study a small sample of the collection or user population, and to draw conclusions based on this sample.

Sampling is, of course, a compromise measure. If unlimited amounts of time and money were available there would be no need for using methods of approximation. But if one is faced with both a need for information and limited resources, it is an important management tool. In order to make the most effective use of this tool statisticians have developed scientific sampling methods. These methods, based on mathematical precepts, assure maximum usefulness and validity of sampling data.

ACCURACY AND SAMPLE SIZE

One of the first things to decide upon is the size of a sample. Intuitively one recognizes that larger samples give rise to more accurate data. To quantify this accuracy he must distinguish between two types of errors possible in a sampling study.

The first type of error is one which might be called tolerance. Most commonly, results are reported as percentages, for example, "25 percent of the books in our collection have circulated in the last two years." Because they are based upon a sample these figures are not exact. Thus we say something like, "Between 23 and 27 percent of our collection has circulated in the last two years." This tolerance is commonly written 25% ± 2 and is read "twenty-five percent plus or minus two." This tolerance is a measure of the accuracy of our result.

Mr. Drott is with the Community Systems Foundation, Ann Arbor, Michigan.

The second type of error measure is called confidence. It is a measure of how certain one is that the true answer lies within the limits stated in his tolerance. For instance a confidence of 90 per cent means that there is one chance in ten (10 per cent) that the true value of the number he is predicting lies outside of the tolerance he has set. A statement that 10% ± 4 of the patrons entering the library go directly to the card catalog reported with a 95 per cent confidence means that there is only one chance in twenty (5 per cent) that the actual percentage of patrons going directly to the catalog is either greater than 14 per cent or less than 6 per cent. Confidence can also be interpreted in terms of the results expected if a sampling study were repeated. Thus a 90 per cent confidence means that if a sampling study were repeated ten times (using the same sample size and tolerance but each time using a different sample) the results would be correct within the specified tolerance for nine of the replications.

Once a tolerance and a confidence have been decided upon one can use Table 1 to find the appropriate sample size. There are several important points in the use of the sample size table. First, the sample size needed for a given tolerance and confidence is dependent on the relative percentages which are observed. The correction can be made by applying a very simple formula. First estimate what per cent of the sample will be in the most important category you are dealing with. For example, if we were sampling the number of patrons who go directly to the card catalog, we might estimate on the basis of preliminary observation that the number is no greater than 20 per cent. We write this as decimal fraction .20. Next we subtract the fraction from one (1.00) and multiply our two fractions together. Thus:

$$1.00 - .20 = .80$$
$$.20 \times .80 = .16$$

This result multiplied by four (4) gives our correction factor.

$$.16 \times 4 = .64$$

This factor is to be multiplied by the sample size in Table 1 to give a revised sample size. If in the example above we had decided on a confidence of 95 per cent and a tolerance of 2 per cent Table 1 gives a sample size of 2,401. Multiplying by our correction factor gives a revised sample size of 1,737. If we cannot predict our sample percentage in advance we can use the sample sizes directly from the table since these represent the most conservative size estimates.

Two further points should be observed in using the sample size table. The first is that the approach used in preparing the table is valid only for sample sizes which are greater than thirty but less than 10 per cent of the total population. A second point is that the sample size must be calculated before performing the survey. The table is not appropriate for calculating the confidence and tolerance of a sample already collected.

Selecting the Sample

The entire validity of this sampling technique is based upon the use of an unbiased sample. That is to say the sample must be as representative as possible of the entire population. To this end one should use the mathematical concept of a random sample. Randomness in this sense means that for each selection (collecting one datum) every member of the population has an equal chance of being drawn. For example if we wanted to select several cards randomly from a new (ordered) deck of playing cards we might throw the cards into a hat, stir them around, and draw the sample with our eyes closed. On the other hand suppose we were to close our eyes, remove a small stack of cards from the top of the deck, draw the next card for our first datum, remove another small stack, draw our next datum, and so on. This would not be a random sample since by removing the stack of cards we gave them no chance of being selected in the following draws. This kind of non-randomness most often appears in so-called fractional sampling. An example of this would be sampling a card catalog by taking every twenty-fifth card (from a random starting point). The effects of this type of violation of mathematical randomness are often difficult to determine. Sometimes the results of a study may be invalidated, other times the violation may have no effect. The critical factor is the order in which the population is arranged.

Let us consider two examples from public opinion surveys. In the first situation respondents were selected by calling at every twenty-fifth house. The interviewers proceeded from block to block in an orderly manner. As they went around each block they stopped at every twenty-fifth dwelling. When the results were compared to known data they were found to be unbiased. The order in which peoples' houses are arranged in a neighborhood seems to be independent of their opinions. Thus the fractional technique did not introduce any bias. In another survey respondents were selected by contacting every twenty-fifth person in the telephone book. The results of this survey were found to be incorrect. It was later recognized that the relationship between names and ethnic groups introduced opinion bias when the names were in an alphabetical list.

To summarize, sampling techniques which are non-random can produce serious and often undetectable errors. Techniques do exist for using certain types of structured samples but these designs require careful statistical analysis and should only be employed after careful consideration.

We have stressed the importance of choosing a random sample, but how does one assure randomness? One method makes use of random number tables. Many books on sampling or statistics include tables of random numbers (see bibliography). For example in a random number table we may find a column of numbers like this:

174393
533251
081831
987384
381849

To use these numbers in sampling we must develop rules for each sampling situation. Suppose we wish to draw a sample from a shelflist which consists of 9 drawers, each drawer having no more than 1,600 cards (about 16 inches). First we select a drawer. For this we can use a very simple rule. Namely, let the first digit of the random number equal the drawer number. We will delete any numbers which begin with zero, since there is no drawer zero. Next, to select a card within a drawer we could use the next four digits of each random number and count that number of cards into each drawer. This, however, would make the data collection extremely tedious. We may decide that measuring a distance into each drawer would be sufficiently unbiased for our purposes. Thus we will wish to choose a number of inches between zero and fifteen and a number of sixteenths of an inch between zero and fifteen. In combination this will allow us to have measurements of from zero to almost sixteen inches. First let us devise a rule for converting the second and third digits of our random number to a number of inches between zero and fifteen. The two random digits form one hundred combinations from 00 to 99. Since we want sixteen numbers (counting zero and fifteen) each group will have six numbers per group. This is because sixteen goes into one hundred a little over six times. Our rule will be:

If the random digits are:	Convert them to inches:
00 to 05	0
06 to 11	1
12 to 17	2
18 to 23	3
24 to 29	4
30 to 35	5
36 to 41	6
42 to 47	7
48 to 53	8
54 to 59	9
60 to 65	10
66 to 71	11
72 to 77	12
78 to 83	13
84 to 89	14
90 to 95	15
96 to 99	Delete

To get a number of sixteenths of an inch, we want to convert the fourth and fifth random digits to numbers between zero and fifteen. We can use exactly the same sixteen division rule developed above. Thus to convert a random number to a card location we convert the first random digit to a drawer number, the second and third digits to a number of inches, and the fourth and fifth to sixteenths of an inch. For example if our random number is 17439 we would draw a card from drawer number one at a distance of $12^{6}/_{16}$ inches from the front.

Note that we have not permitted a sixteen since this could have given us $^{16}/_{16}$. This illustrates a very important point. Suppose for example we had allowed $8^{16}/_{16}$ to equal nine (9). Then there would be two ways to get a nine ($9^{0}/_{16}$ or $8^{16}/_{16}$) but only one way to get a number like $9^{3}/_{16}$. This means that whole numbers (like 9) would be more likely to occur than fractional numbers (like $9^{3}/_{16}$). But our definition of randomness required that all numbers be equally likely. This is the reason that sixteen has been excluded.

Now let us consider some applications of this sampling technique to specific problems. These examples involve the three most commonly sampled items in the library; card files, patrons, and the collection itself.

Example 1

A large research library is concerned about recent discoveries of inaccuracies in their holding records. An inventory would be extremely expensive and would consume a great deal of professional time; thus the librarian wishes to conduct a sample study to determine if an inventory is actually necessary.

In order to set a tolerance and confidence the librarian considers what he will do with the results of his study. The librarian has decided that if less than 2 or 3 per cent of the collection is missing he will take no action. If more than 6 per cent of the collection is missing he is certain that he will conduct an inventory. He is not sure what action he will take if the percentage missing is between 3 and 6 per cent. We can see that the tolerance must be less than 3 per cent; in making a decision it will be important to distinguish between 3 and 6 per cent. The librarian believes that a tolerance of 1 per cent will make this information most useful to him. He is not certain exactly what confidence he desires, but because of the costs involved in being wrong (*e.g.*, performing an unnecessary inventory) he has tentatively set a confidence of 99 per cent. The table indicates a sample size of 16,590 for these values. The librarian is certain that no more than 10 per cent of the collection is missing. Therefore we can calculate a correction factor for the sample size as follows:

$$.10 \times .90 \times 4 = .36$$

We multiply our sample size by this correction factor to get a revised sample size of 5,973. Because of the importance of this measurement the librarian is willing to take a sample of the required size. Thus readjustment of the tolerance and confidence is not necessary.

The sample will be drawn from the shelflist. Since this source is biased against serials and other series, only monographic entries will be considered. The shelflist consists of 1,200 drawers (all numbered consecutively) each containing up to 14 inches of cards. The sample will be drawn by measuring the cards. The random number table used by the library has the digits arranged in columns thus:

47	68	96	90
38	14	42	64
18	11	30	98
55	60	53	30
97	83	71	30

Drawer numbers must be numbers from 0000 to 1199. To make drawer numbers the first two random digits must be converted to numbers between 00 and 11, while the next two digits must become numbers between 00 and 99. To convert the first two digits the following rule is developed:

If random digits are:	Convert them to:
00 to 07	0
08 to 15	1
16 to 23	2
24 to 31	3
32 to 39	4
40 to 47	5
48 to 55	6
56 to 63	7
64 to 71	8
72 to 79	9
80 to 87	10
88 to 95	11
96 to 99	Delete

Each group of random numbers includes eight numbers because twelve (the number of numbers in the range 0 to 11) goes into 100 (number of numbers in the range 00 to 99) eight times plus a fraction. The second part of each drawer number can come directly from the random list. To pick a number of inches between 0 and 13 a similar rule is used except this time each division contains seven numbers thus the rule will be:

00 to 06	0
07 to 13	1
91 to 97	13
98 to 99	Delete

To get sixteenths of an inch we use a rule with six numbers per division:

00 to 05	0
06 to 11	1
90 to 95	15
96 to 99	Delete

This is the same rule developed earlier.

We can now use all of our rules to pick a sample. For example, using the random numbers we would have:

Random Number				Drawer	Inches	16's
47	68	96	90	568	13	15/16
38	14	42	64	414	6	10/16
18	11	30	98	211	4	Delete
55	60	53	30	660	7	5/16
97	83	71	30	Delete	—	—

To make the actual task of taking the sample easier the selection can be ordered by drawer number before collecting the data. If in collecting the data we should find too few cards in a drawer to take the required measurement the data point should be deleted. This is important in order to preserve randomness.

Example 2

A library is taking a survey of user's opinions about library services. The data will be collected by handing out questionnaires to a random sample of patrons as they enter the library. This survey will be only one of many things which the librarian will use in deciding on changes in user service. Thus a tolerance of 5 per cent and a confidence of 90 per cent seem adequate. The librarian has no idea what percentages of the users will hold various opinions thus the sample size of 271 will be used directly from the table. The librarian has decided that the survey should cover a period of two weeks to assure a representative sample of users. The sample will be drawn by converting random numbers to times. The library is open Monday through Friday from 9:00 A.M. to 9:00 P.M. and Saturdays from 9:00 A.M. to 6:00 P.M. We will need rules to convert random numbers to twelve days, twelve hours, and sixty minutes. Our rule for converting to days will use the first two random digits.

If random digits are:	Convert them to:
00 to 07	1
08 to 15	2
16 to 23	3
88 to 95	12
95 to 99	Delete

The same rule can be used on the next two random digits to give us time. In this case one will be equivalent to 9:00 A.M., two to 10:00 A.M. and so on, with twelve being 8:00 P.M. Next we need to convert to minutes. A rule with sixty steps would be tedious to construct and to use. We may decide that no bias would be introduced by using time in five-minute intervals. Since there are twelve five-minute intervals in an hour, we need a rule with twelve divisions. We can use the same rule that we developed above. We can convert by setting one equal to five minutes after the hour, two equal to ten minutes after, and so on with twelve being equal to sixty minutes after which is the next whole hour. Part of our sample would look like this:

Random Number		Day	Hour	Minute
0301	1594	1	9 A.M.	10
8460	8881	11	4 P.M.	60 (5 P.M.)
8393	6703	11	8 P.M.	45
6694	4640	9	8 P.M.	30
9632	0065	Delete	—	—

Note that even though we have used the same rule we used different random digits. The last two random digits in each line were not used. In taking the survey a questionnaire will be given to the first person (old enough to understand it) to enter the library after each sampling time.

Example 3

A librarian wishes to determine whether significant shelf space can be obtained by removing little-used books from the collection. The criterion has been established that a little-used book is one which has not circulated in the last five years. The sample will be drawn by examining the date due slips and book cards in the back of randomly selected books. If 15 per cent or more of the collection can be removed, the librarian will take action. A confidence of 95 per cent and a tolerance of 3 per cent are desired. But budgetary restrictions limit the sample size to 500. The librarian believes the confidence to be more important and thus adjusts the tolerance to 5 per cent. The sample size from Table 1 is 384. There is little doubt that the number of books satisfying the criterion will be less than 25 per cent of the collection. Thus the correction factor is:

$$.25 \times .75 \times 4 = .75$$

This gives a final sample size of 288.

TABLE 1[*]

CONFIDENCE AND TOLERANCE DETERMINE SAMPLE SIZE

CONF.	TOL.	SIZE	CONF.	TOL.	SIZE
99%	± .5%	66,358	90%	± .5%	27,060
	1.0	16,590		1.0	6,765
	2	4,147		2	1,691
	3	1,843		3	752
	5	664		5	271
	7	339		7	138
	10	166		10	68
95%	± .5%	38,416	.80	± .5%	16,435
	1.0	9,604		1.0	4,109
	2	2,401		2	1,027
	3	1,067		3	457
	5	384		5	164
	7	196		7	84
	10	96		10	41

[*] Values in this table are based upon formulae derived in Report No. MG-ML-100, Community Systems Foundation, Ann Arbor, Michigan.

The collection consists of about 19,000 volumes arranged on 234 sections of shelving. Each section has six shelves and there are 25 books or less on each shelf. To pick a section we will want rules to convert the first random digit to a number between zero and two. This calls for a rule with three numbers per division.

0 to 2	is converted to	0
3 to 5		1
6 to 8		2
9		Delete

We can use the second and third random digits directly as the second and third digits of the section number. It is important that we recognize that the second and third digits must range from 00 to 99 since we need to be able to obtain section numbers such as 095 and 173.

To select a shelf within a section we need a rule with six divisions (there is no shelf zero). We will use the fourth and fifth random digits. Each division will have sixteen numbers in it.

00 to 15	is converted to	1
16 to 31		2
32 to 47		3
48 to 63		4
64 to 79		5
80 to 95		6
96 to 99		Delete

To pick a book from the chosen shelf we need a number between 00 and 25. We can use the rule developed for the section number to convert the sixth random digit to a number between zero and two. The seventh random digit can be taken directly to be the second digit of the book number. Combining all of our rules we can draw our sample.

Random number		Section	Shelf	Book
17340	44906	073	3	14
37589	96988	175	6	Delete
70322	75172	203	2	25
63492	26401	234	6	06

Again in drawing the sample it may be convenient to convert all of our random numbers first and order them by section and by shelf before drawing the sample.

Final Remarks

Random sampling is not necessarily an easy operation. Much thought must go into selecting a confidence and tolerance and developing rules for converting random numbers. Furthermore the tasks of actually converting random numbers and drawing the sample may be tedious. On the other hand the entire job of running a library is becoming more complex. To make decisions which are more technical and involve larger amounts of money librarians need both data and an understanding of how accurate it is. The material presented in this article should make it possible for librarians to perform many sampling studies by themselves. For more complex studies there are specially developed statistical techniques. Among these are methods of analyzing data in order to obtain more information from them, methods for more efficient sampling, and for recognizing and avoiding biases. Finally computers may be used for generating the sample and for analyzing the results. These techniques, however, are in the domain of the specialized researcher rather than that of the librarian. ■■