USING SAMPLING TO ASSESS LIBRARY COLLECTIONS

by Jim Self
Director, Management Information Services
University of Virginia Library
December 2001
self@virginia.edu

This document is designed to demonstrate the use of sampling to obtain accurate information about a library collection. Two research projects are explicitly described; the sampling methodology as described can be adapted for use in other situations and projects.

In a research library with millions of volumes, a complete census examining every volume is not a very practical possibility. To learn about a sizable collection, we need to sample it.

Sampling is a means of obtaining information about a population without examining every single member of the population. Sampling saves time and money at a cost of precision and accuracy.

We can use sampling to estimate:

1. the percentage of filled and available shelving
2. the number of volumes in a collection
3. the incidence of a particular characteristic, e.g., what percentage of the collection is damaged or brittle?
4. the growth rate of a collection, i.e., how fast are the stacks filling up?

The size of a sample necessary to produce a good estimate is a matter of judgment. A larger sample will of course bring greater precision and certainty than a smaller sample, but larger samples cost more than smaller samples. After considering the consequences of error and the costs of sampling, the researcher must decide how much time and money to invest in a sampling project.

This brief guide will not specify sample size, rather it will inform the researcher just how much precision and certainty can be obtained from a given sample size. In addition, it will offer practical advice and suggestions as to how to conduct a sampling project in a library collection.

SELECTING THE SAMPLE

For a statistical procedure to be valid, the sample must be unbiased. Every member of the population should have an equal chance of being chosen. Extracting an unbiased sample from a large library collection is a challenging task. A typical research library contains a number of stack floors of varying size and arrangement. Some floors may be tightly packed with books while others contain many empty shelves. No one can make a good estimate of the collection by walking through it and looking at the shelves.

For research libraries, with their numerous and separated spaces of irregular layout, systematic sampling is generally the best procedure. The researcher counts all the shelves in the collection and selects every nth shelf for the sample. The same number, henceforth called the sampling index value, is used for the entire collection under study. Thus a sample might include every 17th shelf. Another researcher on another project might decide to sample every 37th shelf.

The sampling index value should be selected to yield a sample of appropriate size - not too large, not too small. It should be a number that does not fit a regular pattern of shelves per section in the library. If there are 7 shelves in most sections of a collection, a sampling index value of 7 would not work well, but 10 or 17 or 23 would be fine.

SAMPLE SIZE

The sample size is determined by the need for certainty and precision. The level of confidence indicates the certainty of an estimate. The confidence interval indicates the precision.

As an example we might be estimating the "fullness" of the shelves in a library. The project might find that the shelves are 74% full with a confidence interval of (plus or minus) 3.5% at a confidence level of 95%.

This means we are 95 percent sure that the true percentage of "fullness" is somewhere between 70.5% (74-3.5) and 77.5% (74+3.5). It also means that in five percent of the cases the true percentage would be less than 70.5 or above 77.5.

In most library work a 95% level of confidence is sufficient. In other fields (e.g., biomedical research or airplane manufacture) greater certainty is required.

ILLUSTRATIONS OF LIBRARY RESEARCH PROJECTS

Described below is the sampling methodology used in two research projects. The same methodology may be used or adapted for other library projects.

Project I: Estimating the amount of used and unused shelf space in a collection

In this scenario a researcher determines the percentage of the shelf space that is actually in use. As a practical matter, if an active collection has 85% of its shelf space filled, the collection is "full." Shelving books is very inefficient and time consuming when space is so limited.

1. Carrying out this project involves several steps:

   Determine the level of confidence and confidence interval. The answers to the questions listed below will determine the size of the sample needed.
   a. How much certainty is needed? What is the consequence of error? Is a 95% level of confidence sufficient for this estimate? If not, then the 99% level should be used.
   b. How precise should the estimate be? This depends on the purpose of the estimate. In some cases, a 5% (plus or minus) interval is fine; sometimes it may need to be plus or minus 1%.

2. Make a rough estimate of the number of shelves in a collection.
   a. If the collection is reasonably small, someone can simply walk through the collection, and count the shelves.
   b. If counting all the shelves is too time consuming, an estimate must be made, based on surveying one or two floors, and multiplying by the number of floors. (This technique produces a very rough estimate indeed, but it is acceptable for this stage of the project.)

3. Estimate the needed sample size
   a. One may use Table 1 below to get a rough idea of the number needed in the sample. If one is willing to accept a 10% confidence interval at a 95% level of confidence, a rather small sample (98) will suffice. If a 1% confidence interval is needed, a much larger sample (9800) is required. To increase the precision by a factor of 10, the sample size must be increased by a factor of 100 (or 10 squared). (The confidence interval is inversely proportional to sample size squared.)

Confidence Interval	**Approximate Sample Size for a 95% Level of Confidence**	**Approximate Sample Size for a 99% Level of Confidence**
1%	9800	12,800
2%	2450	3200
3%	1089	1422
4%	613	800
5%	392	512
6%	272	356
7%	200	261
8%	153	200
9%	121	158
10%	98	128

Table 1. Estimating the Appropriate Sample Size

b. For most library applications, a confidence interval between 1% and 5% is appropriate. And for most library applications a 95% level of confidence is adequate.

4. Choose the sampling index value

   a. For purposes of illustration, let us assume we have selected a 4% confidence interval at a 95% level of confidence. From the chart above we find this combination calls for a sample of 613. Let us further assume the rough estimate of number of shelves is 11,000.

   b. We divide the number of shelves (11,000) by the sample size (613), we get a result of 17.9. Rounding down, we have 17 as our sampling index value. That is, we will measure and tally every 17th shelf. (Rounding down increases the sample size slightly.)

   c. When a large collection is being sampled, the sampling index value should ideally be greater than the number of shelves in a typical section. As noted above, the most important consideration is to avoid a regular pattern; the sample shelves should not occur at the same place in every section.

5. Measure the shelves

   a. The researcher goes through the stacks, starting at the beginning, measuring every 17th shelf. The occupied and unoccupied portions of the shelf are measured in inches or centimeters, and the two numbers are recorded.

   b. It is recommended that measurements be kept separately for each floor or major section, so that subtotals may be calculated, if desired.

6. Calculate the findings

   a. Once the measuring is finished, the totals of occupied linear space and unoccupied linear space are tallied. The two totals can then be added together, and the percentages of occupied space and unoccupied space can be calculated, e.g., %Occupied Space=(Occupied Space/Total Space)*100.

   b. Once the sampling and tallying are completed, the actual confidence interval should be determined. It will probably be slightly more precise than the originally chosen confidence interval.

   c. Table 2 below is an embedded Excel chart. It contains four rows of sample data and space for additional data. Double clicking the chart will activate it. Once the chart is activated, data may be directly entered, and the chart will perform calculations and display results.

   d. After Table 2 is activated, enter the data found in the measurements above. Enter the sampling index value in the first column of the table; in the second column enter the number of shelves measured in the sample; in the third column enter the percentage of shelf space occupied, as found in the project. The table will calculate and display the confidence interval, as well as the minimum and maximum values, at a 95% level of confidence. We can be 95% certain the actual percentage of occupied shelf space will be no less than the minimum, and no more than the maximum.

Calculating Confidence Intervals
Determining the Fullness of Library Shelves

Sampling Index Value	Number of Shelves in Sample	Sample Results: % of Shelves Occupied	Plus or Minus Sample Error	Minimum Value in Confidence Interval	Maximum Value in Confidence Interval
11	105	11.0%	5.7%	5.3%	16.7%
8	235	48.4%	6.0%	42.4%	54.4%
13	400	50.0%	4.7%	45.3%	54.7%
27	350	64.0%	4.9%	59.1%	68.9%

Table 2. Calculating confidence intervals at a 95% level of confidence.

Project II: Estimating the number of volumes in a collection

This project requires the researcher to find two statistical values: the mean of the sample, and the standard deviation of the sample.

The mean is a measure of central tendency; it is calculated by taking the sum of all values in the sample, and dividing by the number of observations in the sample. The term "average" usually refers to the mean. The standard deviation is a measure of dispersion; it is the square root of the sum of the squared differences of the mean and the values. Fortunately, it is easy to use Excel to calculate these statistical values; the procedure is outlined below.

To estimate the number of volumes in the collection, start with steps 1-4 as described in Project I, then proceed as follows:

1. Count the volumes
   a. Count and record the number of volumes on each nth shelf (the sampling index value).
   b. As noted above, it is advisable to keep separate tallies for each floor or major section.

2. Use Excel to determine the mean and standard deviation of the sample.
   a. Enter the volume count from each shelf in the sample. The entries may be in rows or columns, or in a combination of rows and columns.
   b. Click a blank cell where the mean (average) should appear.
   c. Pull down the Insert menu from the toolbar, and choose Function.
   d. Choose the function AVERAGE on the right side of the screen.
   e. Choose OK.
   f. Identify the data by clicking and dragging, or by entering the first and last cell names.

g. Choose OK.
h. The mean of the sample will appear in the selected cell.
i. Select another blank cell to display the standard deviation of the sample.
j. Repeat the above process using the function STDEV
k. The standard deviation of the sample will appear in the selected cell.

3. Calculate the estimated number of volumes in the collection.
a. Table 3 is also an embedded Excel chart; it contains two rows of sample data. Double click the chart to activate it.
b. After activating the chart, enter the sampling index value in the first column; in the second column enter the number of shelves measured in the sample; in the third and fourth columns enter the mean (average) and standard deviation of the sample, as calculated in the Excel worksheet.
c. The table will calculate and display the estimated number of volumes in the collection, as well as the minimum and maximum number of volumes, at a 95% level of confidence. We can be 95% certain the collection will be no smaller than the minimum and no larger than the maximum.

Estimating Collection Size From a Sample

Sample Index Value	Number of Shelves in Sample	Mean Volumes per Shelf	Sample Standard Deviation	Estimated Number of Volumes In Collection	Minimum Number of Volumes	Maximum Number of Volumes
17	800	27.30	10.60	371,280	361,588	380,972
2	670	22.63	9.72	30,324	29,627	31,022

Table 3. Estimating Collection Size.

The results in Table 3 are calculated as follows:

Multiplying the "sample index variable" and the "number of shelves sampled" gives the total number of shelves in the collection (not displayed in the chart). The total number of shelves is multiplied by the "mean volumes per shelf" to produce the "estimated number of volumes in collection."

Four variables are used to calculate a mean error of the sample: "number of shelves sampled," total number of shelves in the collection (not displayed), "mean volumes per shelf," and "sample standard deviation." The mean error of the sample is used to calculate a confidence interval for the "mean volumes per shelf" variable, as well as the "minimum number of volumes" and the "maximum number of volumes."

All calculated variables may be viewed if the table is activated, and the columns are "unhidden." Also, the calculation formulas are visible if one clicks on the appropriate cell. The formulas assume a sample size of at least 100 shelves; if fewer than 100 shelves are sampled, the estimates will be somewhat less reliable.

Other Projects

As noted above, we can use sampling to estimate the incidence of a particular characteristic in a collection (e.g., damaged books). Such an undertaking would be a variant of Project I. Table 2 would calculate the estimates; one simply changes the header from "Percentage of Shelves Occupied" to an appropriate description such as "Percentage of Books found to be Damaged."

To find the number of linear feet (or meters) of filled and available shelf space, we can use a modification of Project II. Instead of counting volumes, we measure (in inches or centimeters) the available and filled shelf space, and adapt the procedures for Project II.

To determine a collection growth rate, we can carry out sampling projects at regular intervals and compare the results over time. Alternatively, if we know the number of volumes added per year, we can adjust the results from Project II to estimate the number of years before a facility is completely filled.

If it is done properly, sampling gives us the opportunity to obtain reliable information at a comparatively low cost. But to insure reliability, we must select the sample in an unbiased fashion, and we must choose a sample of appropriate size.

Jim Self
Director, Management Information Services
University of Virginia Library
December 2001
self@virginia.edu