Sampling in Archives

An Essay Illustrating the Value of Mathematically Based Sampling and the Usage of Techniques Other Than Simple Random Sampling within an Archives; or, Coping with 10,000 Feet of Invoices before Retirement

FRANK BOLES

THE AMERICAN HISTORICAL ASSOCIATION, in 1948, formed the Ad Hoc Committee on Manuscripts to consider problems of "recent large collections." After years of deliberation, the committee submitted a startling report. Most significantly, it legitimatized the wastebasket as an archival tool. It was not that the committee considered modern documentation worthless. "Practically any paper may conceivably be of some use, to somebody, at some time," wrote the committee. Nevertheless, destruction of marginally useful documents was essential if archivists were to cope with, and make usable to scholars, the immense collections of the twentieth century. The committee concluded, "the archivist must be wise enough, and bold enough, to take a calculated risk, and the historian and the biographer must recognize the difficulties, assist with conference and advice whenever possible, and finally, accept the situation."[1]

In the more than thirty years since the writing of those words, archivists responsible for contemporary records have shown great reluctance in implementing them. Virtually every archival institution seriously collecting twentieth-century material has great quantities of marginal paper littering its stack areas.

One exception to this generally bleak situation has been the application of sampling techniques to these records. While the actual number of archivists who have implemented a systematic sampling procedure is limited, interest in the idea of sampling has been high; because, when properly conceived and implemented, sampling allows the archivist to eliminate quantities of paper, with only a slight possibility of altering their overall historical value. Skeptics of sampling abound, however, and proper conceptualization and implementation of sampling has been more the exception than the rule.

The purposes of this short essay are three. The first is to discuss what sampling is, and in particular to refute the claim that archival sampling and mathematical sampling are distinct and separate methods. The second is to address the problem of implementation, in particular to suggest an alternate technique to simple random sampling. The final purpose is to demonstrate the legitimacy of sampling by reporting the results of it carried out on a typical archival record group, of known values.

The most cogent effort to distinguish between archival and statistical sampling was written by Paul Lewinson. Lewinson

[1] "Report of Ad Hoc Committee on Manuscripts Set Up by the American Historical Association in December 1948" *American Archivist* 14 (July 1951): 229, 232.

argued that archival sampling was broader than mathematically based statistical samples. Lewinson stated that a statistical sample must have mathematically measurable reliability, while an archival sample need only represent or illustrate. Archival samples required no measurable reliability. This was the critical distinction between the two.[2]

There is no dispute over the fact that sampling, in its broadest sense, does include both mathematical and non-mathematical approaches. However, the fundamental limitation of non-mathematical sampling can be presented in a single question: if archival samples need not be of mathematically measurable reliability, what are the criteria by which archivists may all agree that a given sample is either representative or illustrative of the record group from which it is drawn? The only answer is that the judgment of the processing archivist must be trusted. The thought pattern used by the archivist in reaching this judgment follows, logically, certain self-evident steps that turn out to be precursors of mathematical measurement.

The first thing an archivist contemplating sampling a specific record group does is to examine it closely to make certain that the documents are homogeneous. If they are more or less the same, the archivist then decides, by some impressionistic means, how much of the material to retain in order to represent or illustrate the entire record group. When decided by someone thoroughly familiar with the documents and their uses, the results may be successful; but the process is such that it is impossible to document that success or even to discuss it clearly.

A statistical sample begins also with an examination of the record group to determine the homogeneity of its documents. Rather than just a conclusion that the documents are more or less the same, a statistical sample requires that an attempt be made to specify homogeneity in a mathematical measure of variance. Then the exact degree of accuracy desired is specified, as well as the degree of certainty that this degree of accuracy will be obtained. Armed with these three numbers: variance, the specified degree of accuracy, and the level of confidence that the specified degree of accuracy will be achieved, algebraic formulas designed to yield specific sample size may be used to calculate how much of the material to retain.[3]

The archivist's thought processes mimic the mathematics of the statistician. The statistician's approach, however, offers significant advantages. At a minimum, it makes explicit and easily discussed assumptions that were heretofore implicit and in a genuine sense inaccessible. Beyond the minimum, it gives subsequent users of the sample a clearer idea of what they are using, and how much faith to put in it. This increased knowledge about the sample is critical for subsequent, statistically sophisticated users of it. The traditional archives user tended to turn to large sets of documents only to illustrate a point already documented through other material. Large sets of documents supplied color or interest, but rarely evidence. The statistically sophisticated user, in contrast, may use these peripheral documents as fundamental evidence, not simple illustration. To do this reliably, the user will want mathematical information about the sam-

[2] Paul Lewinson, "Archival Sampling," *American Archivist* 20 (October 1957): 291–312. The specific argument regarding the distinction between archival and mathematical sampling is found on 291–92. With the exception of the point with which I take issue, this article is generally sound.

[3] This comparison of mathematical and non-mathematical sampling is an adaptation of an argument made in Leslie Kish, *Survey Sampling* (New York: John Wiley & Sons, 1965), p. 19. For the sake of brevity, I have omitted any coherent discussion of the mathematics of sampling. Those wishing detailed information about either the mathematics or the techniques of sampling will be well served by the Kish volume. Kish, however, does assume a certain level of statistical sophistication on the part of his reader. Hubert M. Blalock, Jr., *Social Statistics: Second Edition* (New York: McGraw-Hill Book Co., 1972) is a good source for finding more about items treated in a cursory manner by Kish.

ples presented. Mathematically undocumented samples, while usable in sophisticated statistical procedures, are unreliable evidence.

Certainly, there are difficulties with mathematical approaches to sampling. Estimating variance can be extremely troublesome. Unfamiliarity with the terminology and concepts of mathematical sampling heighten the archivist's problem. But occasional difficulty and passing unfamiliarity are not reasons to rely upon a more elemental process that is implicit, undebatable, and increasingly unhelpful to the subsequent user.

In archival literature's discussion of mathematical sampling, simple random sampling has been the technique usually discussed.[4] Emphasis has been placed upon simple random sampling because it is the surest way to connect sampling theory with operational procedures. Unfortunately, it is procedurally cumbersome; and for that reason it is not commonly used in field work conducted by survey statisticians.[5]

The most vexing procedural difficulties of simple random sampling involve the use of random number tables. To use the tables at all, each element of the population to be sampled must bear a unique identification number. Unfortunately, many record groups an archivist might wish to sample do not bear identification numbers, or bear duplicate numbers, so that in practice, before the sample can be conducted, each item must either be labeled or checked for duplication of identifying number. Even assuming the best possible circumstance, that the records to be sampled come to the archivist already bearing identification numbers known to be unique, a random number table is awkward to use. The reason is almost self-evident: the numbers have no pattern, making it impossible to work systematically through the records being sampled. Rather, the processor must select, for example, element 100973, followed by 375420, 084226, and 990190.[6]

Because of the practical problems involved in carrying out a simple random sample, survey statisticians have developed an alternative technique: systematic sampling. It differs from simple random sampling in the way elements of the total population are chosen for the sample. Instead of random number tables for choosing elements, in systematic sampling sample elements are picked by their location within the total population. In a typical systematic sample, every hundredth element of the population would be selected for the sample. This procedure is straightforward. It is simpler and quicker. Because it is simpler, it is less prone to human error. Because it is quicker, it is usually cheaper.[7]

Survey statisticians assume that the results of a systematic sample approximate those of a simple random sample. The justification for this assumption rests on the realization that there are two ways by which selection can be randomized. The first is by application of random number tables during the process of selection. The second is by the randomization of the population elements prior to selection. A common example of the latter is found in any game in which playing cards are dealt systematically. Assuming the dealer is experienced, and honest, the result of shuffling is that each player is dealt a truly random set of cards.[8]

In many cases, the arrangement of records tends to randomize the placement of documents. Alphabetical or chronological arrangements, as well as arrangement by social security number, all randomize placement. This being the case, a systematic sampling of the records results in a

[4] The most recent example of this is Eleanor McKay, "Random Sampling Techniques: A Method of Reducing Large, Homogeneous Series in Congressional Papers," *American Archivist* 41 (July 1978): 281–89.

[5] Kish, 113–15.

[6] Random numbers from Blalock, Table B, 554.

[7] Kish, 113–15.

[8] Ibid., 118.

truly random sample just as a systematic dealing of the playing cards results in a truly random hand.[9]

All arrangements of documents are not random. Files arranged by subject are the most obvious example. Such files cannot be systematically sampled. While obviously systematic filing systems are easily recognized, the danger exists that systematic trends which destroy the assumption of random document distribution may exist within a record group that appears to be randomly distributed. The most important of these hidden regularizers are monotonic and periodic trends. A monotonic trend is one in which a persistent pattern occurs. An example would be a constant rate of inflation. Periodic trends are ones in which some sort of regular pattern occurs. For example, those free to choose their own telephone number might regularly prefer certain combinations, such as TY8-7100, while regularly avoiding certain other combinations, such as 313-1313. Periodic trends are often found in time series data, or in lists or ledgers, which often have peculiar regularities regarding the first and last entry on a page.[10]

It is never possible to be absolutely certain that a particular group of records is free from hidden regularities. Therefore, some statisticians, usually those involved in purely theoretical work, reject systematic sampling. The more common position held by statisticans engaged in actual surveying is that careful study of a population, and certain mechanical techniques, such as varying the pattern of selection at intervals, make the risks involved in systematic sampling minimal and acceptable.[11]

The best demonstration that the assumptions made by survey statisticians are accurate, and, for that matter, that the entire theoretical structure erected to justify sampling is valid, is to apply the knowledge to a typical set of archival records of known values. One such set of documents are the sales records of the Stearns Salt and Lumber Company, of Ludington, Michigan. Among the corporate records the Stearns firm has deposited at the Bentley Historical Library, Michigan Historical Collections, are approximately two cubic feet of what appear to be wholesale sales records for the years 1911 through 1914. An individual record was maintained for each firm or person who purchased material from the Stearns Co. during this time. Each account includes the name of the customer, the customer's location, an invoice number for each shipment of goods to the customer, and a standard balance sheet listing the account's debits and credits and the dates of posting. The accounts are arranged alphabetically, by name of customer, with a separate alphabet begun each year to accommodate new customers who placed their first order that year.[12]

Altogether there are 1,207 accounts. The mean value of each account, over the four year period, was $3,745.70. Individual accounts ranged in value from a minimum of $1 to a maximum of $733,850. The record group's variance, as measured by standard deviation, is 24,081. These numbers reveal a record group posing significant challenges for a sampler. The basic problem is that the variance is very high while the total size of the record group is quite small. Because of this, formulas designed to calculate sample size often yield a result larger than the total number of elements within the population.

The problem could be solved in several ways. One would be to accept the lower level of accuracy. A second solution would

[9] Ibid., 115.
[10] Ibid., 120–22.
[11] Ibid., 121–22.
[12] Stearns Salt & Lumber Company, of Ludington, Michigan. The papers are in the Michigan Historical Collections, Bentley Historical Library, the University of Michigan, Ann Arbor. All accounts were converted into machine readable form by myself. All figures were rounded to the nearest dollar. All subsequent calculations and manipulations of the data were accomplished through *Midas,* a package of statistical programs developed at the University of Michigan and available through the Michigan Terminal System (MTS).

be to accept a lower level of confidence that the desired degree of accuracy would be obtained. A third, and preferred, solution is to lower the variance of the total record group.

Close examination of the records reveals that most of the accounts were relatively small and homogeneous. A few, however, were very large. Nine accounts were greater than $100,000, while another sixteen were each valued at between $30,000 and $100,000. The presence of these twenty-five accounts resulted in the very high standard deviation for the total population. If they were segregated from the rest of the accounts, the standard deviation of the remaining 1,182 accounts dropped dramatically. The standard deviation in the remaining accounts became a manageable 3,851.2. Removal of the twenty-five largest accounts also reduced the mean account value. The mean value of the 1,182 accounts was $1,724.30.

The much lower standard deviation figure of the 1,182 remaining accounts made it possible to carry out a sample of these. Establishing plus-or-minus $310.50 of the mean account as an acceptable level of precision, and setting confidence at 95 percent, a level commonly specified in the social sciences, the predicted sample size needed to meet these criteria for a population with a standard deviation of 3851.2 is 591 elements.

First, a sample of 591 accounts was chosen at random. The mean account of this sample was $1,632.90, $81.30 less than the actual mean account of the total population and well within the acceptable deviation of plus-or-minus $310.50. A systematic sample of 591 accounts yielded a sample mean account of $1,855.60, $141.40 above the population mean, but again well within the defined limit.

When the samples were analyzed in greater detail, subsample means drawn from parts of both the random and systematic samples fell within the defined acceptable accuracy of plus-or-minus $310.50 of their comparable total population subsample mean. For example, the true mean account for sales within the state of Michigan was $2,188.70. The Michigan mean account for the random sample was $2,063.50; for the systematic sample it was $2,486.20. The mean account for the states adjoining Michigan: Ohio, Indiana, Illinois, and Wisconsin, and the Canadian province of Ontario, is $1,463.50. The random sample yielded a mean account for these contiguous states and Ontario of $1,371.30. The systematic sample's mean account for the same area was $1,371.10.

Not only do both sampling techniques select sample populations that fall within the previously defined level of precision for mean account, they provide good estimates of other variables. One example is the geographic distribution of accounts. Of the 1,182 accounts, 36.4 percent (430) were within the state of Michigan, 39.3 percent (465) were located in the four contiguous states or Ontario, and the remainder, 24.3 percent (287) were located elsewhere. In the random sample, the distribution of accounts was 36.6 percent Michigan, 41.8 percent contiguous states or Ontario, 21.6 percent elsewhere. The distribution of accounts within the systematic sample were 36.9 percent, 41.3 percent, and 21.8 percent respectively. Overall, the largest difference between the real distribution of accounts and the sample distribution was 2.7 percent.

A calculation of sample size for one variable within a data set will not always supply acceptable samples for subpopulations of that variable, or for other variables. Should a subpopulation or other variable display more variance than the variable studied, the subpopulation or other variable will not be well represented in the sample. Thus, in dealing with populations which contain more than one interesting variable, or interesting subpopulations of a particular variable, each variable or subpopulation should be considered, so that it will be adequately represented in the sample.

This important qualification aside, the results of a controlled sampling experiment on typical archival material is most reassuring. The statistical theories worked as they were supposed to work. Systematic sampling was demonstrably as effective in choosing a representative sample as random sampling was. If these results would

not startle a survey statistician, they are comforting to the less mathematically oriented archival community. The techniques of sampling so effective in the case of the Stearns Company wholesale records would work as effectively when applied to similar archival material.

In point of fact, the mathematical approach used to sample the Stearns papers would be more effective with larger groups of papers. This is so for two reasons. The first is that the larger a homogeneous group of papers is, the lower the variance will be, because the huge mass of similar paper grouped around the mean tend to obscure the few unusual items. The greater the mass, the more obscure the unusual items. Secondly, the formulas used to calculate sample size are not affected by the size of the original population. Given three record groups with the same variance and the same defined precision and accuracy, the needed sample for each will be the same, despite the fact that there are 1,000 items in the first data set, 100,000 in the second, and 1,000,000 in the third. Given these two facts, the bigger the original population, the more advantageous sampling becomes. The bigger the record group, the more paper the archivist may safely dispose of.

While sampling can be carried out in any number of ways, archival sampling that can be rationally discussed within the profession and that will be of maximum use to the increasing number of statistically sophisticated researchers must be based on mathematical foundations. Used thoughtfully, the mathematics can be coupled with simplified sampling techniques, such as systematic sampling, to obtain, at minimal expense, samples of known properties. The period of transition between the less demanding idea that archival samples need not employ mathematical approaches and the more difficult, but ultimately more rewarding, implementation of mathematical approaches, will be a trying time. It is a step, however, as unavoidable to the curator of large twentieth-century collections as the leap between item-by-item inventorying and group description.

FRANK BOLES, doctoral candidate in history at the University of Michigan, is a manuscripts assistant with the Michigan Historical Collections, Bentley Historical Library. He wishes to acknowledge the assistance of Francis X. Blouin, Jr., Thomas E. Powers, and Julia Marks Young.