Archival Sampling

By PAUL LEWINSON [1]

National Archives

Introduction

THIS paper discusses one of several methods of selecting Government records for archival preservation. *Selection* is the basic archival task; without it there would be *records* but no *archives*. It is through selective appraisal that many records are disposed of and a few (those of enduring value) are retained. Those retained are archives. Records — the total documentation of an organization's operations — are unmanageable by the small body of employees in the typical archival institution and are unusable by researchers. Archives — the selected records of enduring value — are manageable and usable. A number of factors enter into the judgment that records have enduring value, and appraisal tests may be applied in various ways. In this paper, however, only one important technique in the selection of records for archival preservation is discussed — the "sampling" of those records.[2]

Definitions and Materials

Since all archives consist of records that have been selected for preservation, archival sampling must be defined in a special way that sets it apart from other methods of selection. At the same time it must be made clear by the definition that archival sampling is not simply a variant either of ordinary sampling, such as is encountered (for example) in the commercial world, or of statistical sampling, although it may include either. When we speak of sampling in the archival context, in short, we are using a convenient term, borrowed from the common vocabulary by way of analogy and given a special technical significance.

The dictionary defines a sample, as the word is commonly used, as "a part of anything presented for inspection, or shown as evidence of the quality of the whole; a specimen." This definition clearly covers the kind of sampling involved, for instance, when a

[1] The writer is Chief Archivist of the Industrial Records Division of the National Archives. His paper discusses National Archives practices but is applicable to other institutions that have to deal with large groups of records.

[2] T. R. Schellenberg, *The Appraisal of Modern Public Records* (National Archives *Bulletin* No. 8, Oct. 1956), should be read in connection with this paper.

sample of grain is taken from a bin in order to establish the grade of the whole lot. Records, too, are sometimes selected for preservation as evidence of the quality of some large body of records. But while a commercial sample is so chosen as to be typical in every way of the whole lot from which it comes, an archival sample may be chosen either for that purpose or because of some particular significance that it has over and above its specimen character.

The statistician defines sampling more strictly as "a procedure by which information of measurable reliability is obtained from only a part of the total population." Again, archival sampling may be so carried out that it, too, gives information about a "population" or "universe" of records, but it will rarely be the case that this information is "of measurable reliability" according to the strict standards of the statistician.[3]

The thread that runs through these two definitions may be picked up in the phrases "evidence of the quality of the whole" (from the dictionary definition), and "information . . . from only a part of the total population" (from the technical definition). There is an implication here that sampling is done with reference to a body of material that hangs together, that forms a unity, that can be dealt with as having some characteristic meaning or meanings.

We may then define archival sampling as follows: "Sampling of Government archives consists in the selection of some part of a body of homogeneous records, so that some aspect of the Government's work or the information received or developed by the Government may be represented or illustrated thereby."

This definition is broader than the ordinary definition in that it envisages an archival sample as one that may *illustrate some aspect* of Government work, rather than *represent* it; and it is broader than the statistical definition in that it requires no *measurable reliability* of the sample.

Our definition also distinguishes archival sampling from other techniques of archival *selection* by limiting archival sampling to the selection of *some part of a body of homogeneous records.* Thus, the archivist is not sampling when he determines that among all the records of a Government bureau those of its top officials should be selected for preservation, or that in a classified general file certain file classes should be selected. In these cases he is not selecting some part of a body of homogeneous records. Both the total

[3] The definition is quoted from W. Edwards Deming, *A Brief Statement on the Uses of Sampling in Censuses of Population, Agriculture, Public Health, and Commerce* (Lake Success, N. Y., UNESCO, 1948), 15 p.

documentation of a Government agency and the general file are heterogeneous and varied bodies of records, save in the most general and superficial sense. The archivist *is* sampling when he selects for preservation a certain number of applications from among all applications, a certain number of case files from among all the case files of a court or board, or all the files of some regional offices from among the documentation of all the regional offices of an agency.

In the preparation of this paper 68 disposition case folders were studied. They represent the response of all the record branches and the Records Management Office of the National Archives and Records Service to a request for job folders covering all cases in which sampling, defined much as it has been defined here, was used.[4] From among these disposition jobs examples can be drawn to illustrate the concept of the "homogeneous body of records" from which samples may be selected for preservation.

Sources for Sampling

Case Files

The most frequent sources of sampled records are case files. As defined in the draft "Glossary of Records Terminology" (National Archives and Records Service, Jan. 1956), a case file is

> a body of records, kept together, dealing with a particular transaction or with closely related transactions; originally, such a body of records pertaining to a judicial or quasi-judicial case, but — increasingly — pertaining to an administrative decision or series of decisions (as in a personnel case file), or to a work project or series of work projects (as in a loan case file or a construction case file).

Traditionally, a case file is the documentation of a dispute between parties that is heard before a court of law. Case files in this restricted sense have been subject to the sampling process in the National Archives. Thus, bankruptcy case files are disposed of, except where they have a specified procedural, real property, or large economic interest.[5] Similarly, criminal case files of the U. S. District Courts are disposed of, except for 12 classes chosen for

[4] Most of the folders were for disposal jobs, but some were for accessions. In dealing with these jobs in this paper, no attempt is made to give subtotals under the various headings of the discussion (since many of the jobs provide for more than one sampling operation or provide for samples that could as readily be classified one way as another). Under these circumstances, no collation of the various kinds of sampling jobs leading to a total of 68 is possible. Jobs especially worth studying in detail are cited in the text or identified in footnotes.

[5] General Schedule No. 6, Job 247-9. In 1930 there were 26,355 bankruptcy cases in the United States; in 1950, there were 9,162.

their historical significance, their juridical importance, their bearing on social and economic conditions, and other reasons.[6]

Similar to court case files are the case files that develop in connection with the proceedings of the Government's many quasi-judicial bodies. These are the boards and commissions — operating less formally than the courts and usually confining themselves to finding facts rather than to deciding legal points — that settle disputes between conflicting parties or grant or deny permission to act in some regulated field of economic activity. Quasi-judicial (or administrative) tribunals may be permanent, as in the case of the National Labor Relations Board, whose "unfair labor practices" case files from the field offices are accessioned on a sampling basis.[7] Or they may be temporary, as in the case of the Office of Price Administration, of whose nearly 1 million enforcement case files all but 3,000 were disposed of.[8]

Other than quasi-judicial agencies make administrative determinations that produce case files in matters that do not involve disputes between parties or regulatory action. There are loan case files, for example, such as those taken as a sample from the records of the Farm Security Administration and its predecessors.[9] There are also project case files such as those for construction projects, sampled from the Farm Security Administration files[10] and for various kinds of individual human welfare projects, sampled from the files of the Veterans' Administration.[11] Perhaps we might also include in the sampling of this class of case files the retention of the complete documentation of a single naval vessel in order to exemplify ships' records.[12]

Submissions

A second distinguishable class of homogeneous bodies of records that may be sampled may be designated as "submissions." This term is intended broadly to cover many types of records, having in common the characteristic that they give the Government requested or required information. The term could cover registrations, as to establish a permissible rent under wartime regulations.[13] Or it might

[6] Job II-NNA-944.
[7] Jobs 346-S19 and II-NNA-674.
[8] Job 347-S204.
[9] Job 446-138.
[10] Job 447-120.
[11] Rehabilitation case files, Job 351-173 for headquarters cases and Job 348-96 for field cases.
[12] Job 445-218; the corresponding disposal job is No. 345-S193.
[13] Job 347-S21.

cover applications (as for permission to act in some regulated field or for a benefit of some sort), returns (as are required of taxpayers), or schedules or questionnaires (answers to very formalized inquiries).

Except as submissions may form the initial basis of case files, they have not often been sampled by the National Archives. One job is cited in the above paragraph. Certain approved "dealer," "utility," and "manufacturer" files that are in effect applications to take part in the rural electrification program have been sampled.[14] Individual income-tax returns were once the subject of an elaborate sampling process, but the sample was later disposed of, and further sampling was abandoned.[15] No samples have been taken of schedules or questionnaires, except as they occur in case files [16] of which they are a part.

Miscellaneous

We come last to a number of instances of sampled records that are neither case files nor submissions and that defy classification. They are useful to consider for the light they throw on the kinds of records for which sampling may be appropriate. According to our definition these kinds of records must all be in some sense "homogeneous bodies."

We find that correspondence files have been sampled when the correspondence is of a particular kind, as inquiries and complaints from mortgage-burdened home owners, congressional correspondence, or outgoing correspondence during a very special period.[17] Administrative and progress reports have also been sampled, as in the case of the Commodity Credit Corporation and the Housing and Home Finance Agency; and research reports, in certain Reconstruction Finance Corporation cases.[18] The complete documentation of a number of OPA local boards was retained as a sample of the records of all such boards.[19] Finally, we find samples of various

[14] Job 348-208.

[15] The sampling of income-tax returns is discussed below in this paper.

[16] The way in which submissions form the basis of or otherwise get into case file samples is illustrated in Jobs 348-S225(OHE), and 348-8(OHE), both having to do with case files resulting from applications for authority to construct.

[17] Jobs D39-132, 351-14, and III-NNR-122, respectively. See also the paragraph on Procedures, below.

[18] Jobs II-NNA-110, II-NNA-869, and II-NNA-1873, respectively.

[19] Jobs 448-62 and 448-73. See also in Record Group 188, Records of the Office of Price Administration, among the records of the record branch, the folders labeled "Historical Records Section Plans" and "Records Branch Plans"; and *Preliminary Inventory of the Records of the Price Department of the Office of Price Administration* (National Archives *Preliminary Inventory* No. 95, 1956), p. 246, entry 1349; and p. 267, appendix IV.

classes of broadcasts, sometimes in the form of recorded transcriptions, sometimes in the form of scripts.

Summary

These, then, are the sources from which samples have been drawn in National Archives practice. All of them conform to the norm of homogeneity. The *case files* of each type are homogeneous in general form, in the procedures they represent, and in the areas of activity with which they deal. The *submissions* tend to be wholly alike in format and are held together by a common procedure and subject matter. And the *miscellaneous* types of sample sources are plainly very special and individual bodies of records.

These sample sources have another common characteristic — a relatively low concentration of interest or value in comparison with their bulk but a value or interest that is worth salvaging if the bulk can be reduced. Many court or quasi-judicial case files, for example, simply decide once more in a particular instance what has long since been fixed on as a principle. Others, however, are important because they start a new way of dealing with some kind of Government regulation or because they deal with a very important question. Case files, therefore, may be valuable not in their totality but in part; they may be sampled by subclasses, or case by case on some specific grounds, or in a small random sample of a generally illustrative nature. And so with the other classes of sample sources.

Objectives of Sampling

A sample of records is preserved for its *value*, just as is any other body of records selected for archival preservation. And the values that a retained sample embody are basically the same as those of other bodies of records — evidential or informational.

Preserving Evidential Values

Evidential values inhere in records that illuminate "the functioning and organization of the Government body that produced them." Broadly speaking function and organization constitute administrative history: the origin of a Government program, its implementation in procedures, and its manner of execution.

The social or economic *origins* of a program appear in a sample "consisting of all [general loan correspondence files] . . . filed under the letter 'C' " at the Washington office of the Home Owners Loan Corporation;[21] these files are appraised as "letters of inquiry, com-

[20] Jobs 452-35 and II-NNA-1212.
[21] Job D39-132.

plaint, etc., from the general public to the President or other officials and replies" which constitute a "mass of detailed testimony to the distress of home owners during the years 1933-36 . . . of great value for the study of social and economic conditions during those years." The entire body of these letters was estimated as 1,400 linear feet; about 35 linear feet were kept.

Procedures arising out of legal and administrative problems and variations among related programs are documented by a Commodity Credit Corporation disposal job,[22] the appraisal for which states that

> The research value and historical interest of these records . . . can be met by the selection from the records of each program [of] samples that will reflect (1) the procedures and methods used . . . (2) the administrative and legal problems encountered . . . and (3) the different aspects and characteristics of each program.

Changes in procedure and policy are documented by a Bureau of Indian Affairs sample [23] that excepts from the provisions of a disposal list "correspondence for the last five working days of each month" and "correspondence for the calendar years 1933 and 1934." The "retained sample," says the appraisal, "will reflect the pattern of administration and organization . . . on a year by year basis [and] documents the activities . . . of the Bureau . . . during a very important and controversial transition period. . . ."

A perspective on the *execution* of a widespread program is preserved in an Office of Rent Stabilization sample consisting of the records of one local rent advisory board from each ORS region, chosen so as to represent "large urban centers, industrial towns, and military installations," the three types of areas affected by Federal postwar rent control.[24]

Preserving Informational Values

As records in general may be valued for "the information they contain on persons, corporate bodies, things, problems, conditions, and the like, with which a Government body dealt," so a sample of records may be chosen for its informational values.

Information on persons. — The impact of a Government program on a class of persons is documented in a number of disposition cases that selected veterans' folders dealing variously with physical, educational, and economic rehabilitation.[25] A selection of case files

[22] Job II-NNA-110. This job is referred to elsewhere in this paper.
[23] Job III-NNR-122.
[24] Job II-NNA-580.
[25] Jobs 348-96, 351-173, III-NWR-2, II-NNA-103, II-NNA-1112, and II-NNA-1317.

whose value depends on the importance of particular persons occurs in preserving, from among criminal court cases, those "involving the President and Vice-President . . . Congress . . . judges [and other federally appointed officers], agents of foreign states, prisoners of war, internees, alien enemies, and persons aiding [the last four classes]."

Information on corporate bodies. — Again, among court files, provision has been made for selecting from the mass of bankruptcy cases those (among others) that deal with "the reorganization of a corporation" or with "a railroad adjustment." [27] Some provision is made for a sample of records showing the price structure of various industries through the retention of submitted prices in an OPA schedule.[28]

Information on things. — Except for the case already mentioned in which the complete documentation of a naval vessel was retained as an illustrative sample, the sample documentation of things in the National Archives consists chiefly of records concerning buildings and their equipment. Included are "complete sets of the basic architectural drawings" of Veterans' Administration general, tuberculosis, and neuropsychiatric hospitals,[29] "the engineering records of selected projects which exemplify the most important techniques of house construction used by the Farm Security Administration in each region of the country," [30] and all "large-scale drawings (i.e., 3-inch to full size scale)" for the buildings and facilities at one War Relocation Authority center that "includes installations of all types . . . found at . . . Centers." [31]

Information on Conditions. — Problems, processes, and states of being are all covered by the term "conditions" as used in this discussion. If a sample of records has been retained for its evidential values, it usually follows that that sample has informational values as well, bearing on the origins and the problems of a Government agency in the documentation of its organization and functioning. A sample may also be taken from records that have no evidential values but that do have informational values, in order that docu-

[26] Job II-NNA-994.

[27] Job 247-9, General Schedule No. 6. [*Editor's note*: The editor sees little advantage in applying the word "sampling" to the process of selecting a whole class of records (whether case files or decimal files) on the basis of a subjective evaluation of its particular importance and with no thought of its standing in any sense for all classes.]

[28] Job 347-S196.

[29] Job II-NNA-1416.

[30] Job 447-120.

[31] Job 345-S186.

mentation on economic, geographical, political, or other conditions may be preserved. In one sampling, for instance, which provided for documenting "the different aspects and characteristics of each program" of farm price-support programs of the Commodity Credit Corporation, the program, the commodity, and the market whence the sample was to come were specified.[32]

In the case of an elaborate sampling of Farm Security Administration and Rural Resettlement Administration loan folders, provision was made for the retention of "all paid-in-full . . . loans made in 134 counties . . . selected to represent cross sections of the various farming areas in the United States." These farming areas, established by experts of the Bureau of Agricultural Economics, represent the varying geographical conditions under which farming is carried on in the different parts of the country.[33]

Another case in which geography was the basis for a sample is that of the rent case files of the Office of Housing Expeditor. Here a list of 20 cities was set up "in order to obtain a panel of area rent offices from the whole country," illustrating wartime rental conditions in all the major regions, in cities large and small, in industrial, commercial, and transportation centers, and in similar categories.[34]

As a final example may be cited certain issues in the collective-bargaining relations between employers and unions that have been singled out as criteria in the annual sampling of case files of the Federal Mediation and Conciliation Service. Among the issues so included are wages and hours; pensions, insurance, welfare, and other "fringe" benefits; the handling of grievances; and questions of union recognition and jurisdiction.[35]

Summary

The general objectives of sampling, or the general purposes that the retention of samples is to serve, range as widely as the objectives of any other kind of archival retention. A sample may be taken from a body of records for *evidential values* — for what it preserves of administrative history in general or, in the case of a particular agency or program, of its origins, organization, procedures, and functioning. In preserving a sample of records for evidential values, especially if a sample is taken at recurring inter-

[32] Job II-NNA-110.

[33] Job 446-138. See also Carl J. Kulsrud, "Sampling Rural Rehabilitation Records for Transfer to the National Archives," in *American Archivist*, 10:328-334 (Oct. 1947).

[34] Job 348-S224.

[35] Job II-NNA-2026.

vals, *informational values* will also be preserved, particularly with respect to the origin of and the changes in the agency or program. But informational values may also be the prime object of sampling: information on the geographical dispersion of a phenomenon or condition, on important types of structures or techniques, on classes of persons, on kinds of bodies corporate, on types of individuals of special interest, or on particular problems within a range of problems dealt with by Government action. When any of these objectives has been attained by the "selection for retention of some part of a body of homogeneous records," sampling may be said to have taken place.

To Sample or Not To Sample

Having defined archival sampling and noted specific cases illustrating sample sources and sampling objectives, we must now consider how to recognize disposition problems for which sampling is or is not indicated as the solution. Later a few matters of sampling technique will be discussed.

As has already been said, archival sampling is relevant only for homogeneous bodies of record material — case files, submissions, very limited and specific types of correspondence or reports, the total documentation of specimen offices out of a network of such offices, and special types of material (such as sound recordings, broadcast scripts, and other audiovisual records) that are also limited in their subject matter, origin, or nature.

Such bodies of records are appropriate for sampling if their total volume is very large compared with the importance of their content and the degree of research interest in their subject matter, or — to put it another way — if it is inconceivable that all could be kept but undesirable that none should be.

In the case of records whose value is chiefly *evidential*, if there is to be any sampling it will usually be of *typical* records that will document agency operations more precisely than do the general directives, policy statements, work statistics, and reports that are usually preserved in any event. Such samples will tend to be relatively small in total volume, although they may be large in an absolute sense where they evidence the functioning of a widespread or highly differentiated activity. In the case of records whose value is chiefly *informational*, sampling is more likely to be on the basis of the nontypical; that is, records relating to leading, important, or significant matters will be segregated from the whole body for retention. The size of such samples will depend wholly on how

many important or significant persons, places, things, conditions, or events fall within the scope of activity of the agency whose records are under consideration. Of course there have been and will continue to be many cases in which a sample is retained for both its evidential and its informational values, and in which provision is made for the preservation of both typical and significant documentation.

In the National Archives a sampling provision has on occasion been written for a job to meet an emergency situation. That is, when a disposition plan has to be drawn up quickly, as in the case of an agency being liquidated, a clause has been written in the schedule providing for the selection of a sample by agreement between the National Archives and the agency concerned before disposal is carried out. Such clauses sometimes specify the size of the sample, sometimes not;[36] they are nearly always drawn to relate to particular items in a schedule. This is in principle a poor practice, as it usually puts off the actual drawing out of the sample to a disadvantageous time — a time when knowledgeable agency personnel is no longer available to work with the archivist; but it is sometimes the only way of ensuring the preservation of some desirable part of a body of records.

Let us now see how these standards were applied in three specific cases.

A disposal schedule covering the field office records of the Price Department of the OPA provided that "all price adjustment files, price determination files, and price filings created or required to be filed as a result of OPA regulations, amendments, and orders" were to be disposed of except for "samples specified in [an] attached listing." [37]

This body of records, which falls into the class of "submissions," was homogeneous: it consisted, in effect, of price lists for goods and services filed by all kinds of business establishments as a preliminary to getting their prices approved or changed under wartime regulations. It is to be noted that the filings were not identical in form; some were simple lists (restaurant menus, for example), some were printed catalogs, some were statements in letter form. All together they constituted quite a large body of records — some 8,000 cubic feet — large both absolutely and in relation to the interest that they might be expected to have over a long period. The sample retained reduced this bulk to about 400 cubic feet; its character was

[36] See Jobs II-NNA-1873 and II-NNA-1159 respectively.
[37] Job 347-S196.

very specifically spelled out in a 29-page listing that formed part of the disposal schedule.

The listing provided for the retention, for each OPA region, of all or of varying percentages of price filings (ranging from 2 to 10%), in some cases with a minimum of 1 filing, in some with a maximum of 50 filings, by price regulation or order number, omitting entirely certain regulations of minor importance. It was intended as a means of preserving evidential values ("The retention of samples . . . is primarily for . . . providing permanent documentation of the administrative process of applying price regulations at the field level where the field was empowered to act with some degree of discretion"); but to a degree, also, as a means of preserving information ("secondarily in some cases for . . . making available economic data supplementary to those available elsewhere").

Other disposal schedules govern case files of the National Labor Relations Board.[38] While the OPA disposal job mentioned above covered a closed record group, as of the time of the agency's liquidation, the NLRB jobs have continuing application. Excepted from disposal are "certain cases to be jointly selected by the Board and the National Archives." At first these selected cases were to be 10 to 12% of the total number; later (by an amending schedule) they were reduced to 3%. Criteria for selection are set up in the schedules.

Here again we have a body of homogeneous records, of the "case file" class; in a sense infinitely large, since the Board is a permanent and not an emergency agency. The average annual intake of samples has thus far been about 25 cubic feet. The records deal with questions arising from the right of labor to organize and bargain collectively and from the prohibition of unfair labor practices on the part of labor organizations and employers.

The criteria for the selection of these case files indicate that both evidential and informational values are at stake. Thus, cases may be selected because they illustrate a "contribution to the development of methods and procedure" or for their "influence in the development of principles, precedents, or standards of judgment" (evidential); or they may be selected because of their "effect upon the national or local economy or upon the industry" or because of "the intensity of public interest" in them (informational).

[38] Jobs 346-S19 and II-NNA-674, the latter an amending schedule.

Negative Considerations

Needless to say, not all large bodies of homogeneous records are to be sampled. In some instances the entire body of records should be preserved; in others, none should be.

In the first group would fall records that represent some new departure in Government activity and records that represent some permanently important transactions. For the Federal Mediation and Conciliation Service the disposal schedule provides for the retention of all case files for the pioneer period of Federal mediation, from 1912 to a date before 1948 that has not yet been determined, but for the retention of only a sample of case files after 1948.[39] And in the instances of population census schedules and land grants all records have been and will continue to be preserved, the schedules as a sort of a "Domesday Book" of the American people, the land grants because of the permanent legal importance of titles to land.

In the group of records of which no sample need be preserved fall, on the evidential side, utterly routine records (e.g., minor "housekeeping" records, however homogeneous and large in quantity); and, on the informational side, records whose content is amply covered in other records or in nonarchival sources, records that are themselves a statistical sample, or records whose quantity is so great that a sample, if large enough to be meaningful, would be unmanageable.

It is perhaps on the grounds just given that so few submissions (one of the sources for sampling mentioned earlier) have been made the source of archival samples for informational purposes. In the case of schedules or questionnaires — very formalized answers to a Government inquiry — the body of records is itself almost always a statistical sample. Thus, for its consumer price index the Bureau of Labor Statistics collects prices in 46 cities (the 12 largest, 9 other large cities, 9 medium-size cities, and 16 small cities), covering in those cities about 60,000 retail establishments, and getting rental figures from about 30,000 tenants. A statistical sample, as has been said, is set up by elaborate technical and mathematical procedures designed to give specific information within strictly measurable limits of reliability. To reduce such a sample to a still smaller quantity of records would probably make the sample less reliable for information or limit still further the kinds of information to be derived from it, or both.

[39] Job II-NNA-2026.

[40] *Techniques of Preparing Major BLS Statistical Series* (U. S. Department of Labor *Bulletin* 1168, Dec. 1954), p. 66, 69.

Where submissions are universal, on the other hand, as in the case of the income-tax returns that are made by all individual income receivers covered by the tax laws, it can be argued that any sample large enough to serve research purposes not already served by such publications as the Bureau of Internal Revenue's annual *Statistics of Income* would itself be too large for further, private exploitation.[41] It can be argued, also, that the retention of any corporate income-tax returns would serve no research purpose that could not be met by recourse to published company reports, commercial bank research studies, and the like.[42]

Special Features

A number of sampling jobs among those examined for this discussion involve particular sampling situations that warrant specific discussion.

It was pointed out earlier that archival sampling is sometimes akin to statistical sampling when it aims at the preservation of *typical* or *representative* records, but that it is often quite different in its objectives when it aims at the preservation of *significant* and atypical records only.

The archivist has to decide whether a random sample, drawn to preserve typical or representative records, or a selected sample is needed. For a *random* sample he must make provision for "removing from the files every fifth, tenth, twentieth, etc., case depending upon the percentage" required; "a random sample in the ratio of 1 to 5"; or "10 area cases from each of the selected cities, 25 . . . from critical areas, 20 . . . from each of the 8 Litigation Offices."[43] In each instance the archivist is concerned directly with the size of the sample.

A *selected* sample of significant records, on the other hand, will require the archivist to establish criteria of significance. This has already been mentioned in connection with three sampling jobs in labor relations case files. In these jobs, after specifying that *im-*

[41] 58.6 million personal-income-tax returns were filed in 1956.

[42] Jobs 349-S167 and 352-S232 provided for the disposal of individual income-tax returns except for an elaborately worked-out sample. Three others, 445-90, 447-C1, and 447-264, covered the accessioning of the samples provided for. Two more, II-NNA-641 and II-NNA-945, canceled the sample retention provisions of the disposal jobs and disposed of the samples already accessioned. The grounds for retaining the sample were informational — to provide the means for studies of income beyond those regularly made by the Government; the grounds for the abandonment of the sampling procedure were principally the exhaustive nature of the Government's regular studies and the improbability of further private research on what was still a large body of records.

[43] Quoted from Jobs 347-S196, 350-118, and II-NNA-580.

portant case files should be drawn out for preservation, up to a stated percentage of the total, it is further set forth what the standards of importance should be. Since these case files have both evidential and informational values, they are to be regarded as important in showing the development of principles and precedents; important because of their repercussions on the economy or because of intense public interest; or important because of the issues involved. These criteria are even further refined. "Principles, precedents, or standards" are to be viewed in terms of such matters as jurisdiction, the limits of the concept of interstate commerce, the "implications of bargaining in good faith," "the unit appropriate for purposes of collective bargaining," and other matters that have special importance in the field of labor relations. Significant "issues" are also spelled out, divided — in one instance — between contract and noncontract issues, and further subdivided as matters of union security and recognition, wages and hours in all their particular ramifications, and other technical matters.

Subject-Matter Knowledge

In devising sampling schemes for records, the archivist obviously must have some special knowledge of the subject matter of the records involved. This may mean a knowledge of the area of labor relations, as in the cases just cited; a knowledge of the laws governing the agencies involved; a knowledge of the problems that have beset these agencies in their functioning; a knowledge of important specific events that affected the work and procedures of the agencies, and — sometimes in addition to, sometimes in default of these qualifications — a knowledge as to what competent authorities, what body of literature, or what college of experts may be consulted. A recent case involving scientific and technological records of the Patent Office drew both on the engineering knowledge of members of the staff and on their acquaintance with scientists and scientific organizations.[44]

Such knowledge will also aid the archivist in deciding when a sample should be changed. Among the three sampling jobs in the field of labor relations case files discussed above, for example, the second is a recasting of the first, partly because of a change in labor-relations legislation (the superseding of the Wagner by the Taft-Hartley Act) and partly because of the desirability of reducing the sample; the third has a written-in provision to cover any changes

[44] Job II-NNA-1291. Special knowledge was also called for in sampling each program of the Commodity Credit Corporation and the loan folders of the Farm Security Administration and the Rural Resettlement Administration, mentioned earlier.

in the nature and importance of the "issues" that are to be represented.[45] A knowing archivist may also find himself in a position to gear the sample he creates to existing practices or needs in some other part of the Government. This was done with certain rent control case files of the Office of Housing Expediter, by selecting as the cities whose records were to be preserved the same cities as were used in the statistical sample of the Bureau of Labor Statistics.[46]

Implementation

It will be noted that some of the National Archives disposition jobs that form the basis for this discussion provide for the retention once and for all of a sample from a closed series; others provide for a recurring sample, to be withdrawn for retention periodically as long as the agency of origin, the activity covered, or the record series continues to exist. In the case of recurring samples the archivist must give thought to the amount of archival space and care that he is pledging on a long-term basis. He must consider carefully how the accessioning of the sample increments shall be spaced in time. If he accessions small increments at frequent intervals, he may needlessly scatter a series or a record group inconveniently around his stacks. If he spaces the increments too widely in time, he may inconvenience the agency of origin. Moreover, he may not get so carefully considered a sample if the originating agency is asked to take on too large a selection job on any one occasion.

In general, other things being equal, samples should be drawn, ticketed for drawing, or covered by very specific instructions as soon as possible; preferably they should be drawn as soon as possible. Otherwise, particularly if an agency is being liquidated, the operating and professional agency staff on whom should fall the chief burden of actual sample drawing will have disappeared from the scene. Or, if the series of records is closed, knowledge of what the records are and mean will have evaporated although the agency may still exist.

It is well, also, to incorporate in a disposal list or schedule that involves sampling some provision to fix responsibility for sampling at a high enough level in the agency. This has been done in a number of the jobs here considered. In Job II-NNA-2026, for example, "each Regional Director will select for eventual transfer to the

[45] Jobs II-NNA-674, 346-S19, and II-NNA-2026.
[46] Job 348-S224.

National Archives . . ."; and in Job 346-S19 "the Regional Directors shall periodically . . . submit to the Board a list of . . . cases . . . important. . . ." The archivist may also suggest that the fixing of responsibility be incorporated in agency issuances implementing disposal lists or schedules. Especially if a sample is to be based on importance, significance, or the issues involved — if the sample is not random, that is — selection should not be left to clerical personnel.

The Statistical Sample

At the beginning of this paper a distinction was drawn between archival sampling and statistical sampling. It was pointed out that archival sampling does not rest basically on mathematical considerations and does not achieve measurable reliability in representing the universe or population it deals with. There have been no instances of statistical sampling among National Archives disposition jobs except in the case of individual income-tax returns, where the sampling requirement was later dropped. The Farm Security Administration loan folders come close to being statistical samples.

But the possibility cannot be foreclosed that some day a statistical, mathematically based sample of known reliability may be considered worth drawing for preservation. Statisticians with whom the question has been discussed do not dismiss this possibility. It is true, moreover, that archivists occasionally encounter statistical samples in their disposition work. Archivists therefore should acquaint themselves with the theory and techniques of statistical sampling by consulting such authoritative but nontechnical treatments of the subject as are available.[47]

This much may be said here.

In a statistical sample we are concerned primarily with reliability. An unreliable sample is of no use to anybody. Both mathematical logic and experimentation have shown that under certain conditions a measurable degree of reliability can be attained in a sample, and that a reliability can be obtained that will measure up to any desired degree short of absolute certainty.

[47] Suggested are Deming, *Brief Statement on Sampling*, cited in note 3 above; and a series of three articles published in the (British) Organization and Methods Division of H. M. Treasury, *O&M Bulletin*, vol. 10, nos. 2, 3, and 4 (Apr.-July 1955). Other discussions of sampling and statistical records of value to archivists are Inter-agency Records Administration Conference, *The Uses and Management of Statistical Records* (Report of the Seventh Meeting, 1945-46 Season, Mar. 22, 1946), and *Sampling Techniques in Records Preservation* (Report of the Fourth Meeting, 1948-49 Season, Dec. 17, 1948); Kulsrud, in *American Archivist*, 10:328-334 (Oct. 1947); and Morris B. Ullman, "The Records of a Statistical Survey," *ibid.*, 5:28-35 (Jan. 1942).

The reliability of a sample is expressed in terms of "confidence limits." This means that any sample drawn from a given universe will produce the same results as any other sample similarly drawn, within a tolerance of 2, 5, 10, or any other percentage of variation, depending on the confidence limits that have been built into the sample. This is, of course, a way of saying that within the stated limits the sample is representative of the universe.[48]

As an example, let us consider the individual income-tax returns that were once the subject of a National Archives disposal job with a sampling proviso. If the sample was properly drawn, it would be representative of the universe of taxpayers in that, for instance, the proportion of returns in the $5,000 to $6,000 tax bracket would be the same in the sample as in the universe except for a stated tolerance of variation. Its representative character could be experimentally tested by drawing a number of additional complete samples and seeing whether or not each sample corresponded with the others, within the stated limits. Without experimentation, however, the reliability of the sample could be prearranged and post-audited by the application of a mathematical formula.

The conditions of reliability are (1) the randomness of the sample and (2) its size in relation to its objectives.

Randomness of the Sample

The mathematics of sampling is basically the mathematics of chance, familiar in a crude way to such archivists as play poker or roll dice. It is for this reason that randomness is an absolute prerequisite to reliability; the members of the universe being sampled that are selected for the sample must be chosen on a chance basis. On commonsense grounds, it will be seen at once that any other method might produce an unrepresentative sample. For one thing, conscious or unconscious prejudice might creep into the selection of the sample. For another, the preexisting arrangement of the universe might corrupt the sample, as in the following instance cited by Deming: "For example, a procedure that calls for the selection of the person listed on line 1 of every [census] sheet . . . would be seriously biassed in favor of heads of families, if heads of families [were] enumerated first." [49] In other words, the sample would be

[48] Deming, as reported in the IRAC Conference on sampling techniques, cited above, makes these points thus: "When a sample is statistically designed, you can be sure that it will deliver the precision that is required, or very close to it. Moreover . . . after a statistical sample is drawn it is possible to discover . . . whether the sample is precise or wide of the mark . . . Sampling error is expressed in terms of a band of error (such as 5 percent) within which the result of a complete tally . . . would fall."

[49] Deming, *Brief Statement on Sampling*, p. 14.

unrepresentative by being heavily weighted with older people and males.

But there is a more fundamental reason for insisting on the random choice of a sample. This lies in the fact that the mathematics of sampling is the mathematics of chance (or probability), and therefore the reliability of any sample is untestable and undemonstrable unless it is a random sample to begin with, save by the processing of a large number of parallel samples, by comparing the sample results with a complete count of the universe, or by completely subjective methods. To make such tests would vitiate the economies of sampling.[50]

Size of the Sample

The archivist, if concerned with a statistical sample at all, will be particularly concerned with the size required for reliability. This will be the case whether he is participating in the actual sampling of records to be retained or whether he is deciding on the value of an existing sample created in connection with some official statistical program whose records he is appraising. He will want to know how big a statistical sample should be to be worth preserving on grounds of its reliability. He will not wish to preserve a larger sample.

On the face of it, one might suppose that the bigger the universe, the bigger the sample must be. This, however, is not true. The size of the sample is determined by (1) *how varied the universe is* in terms of the factors whose magnitudes or relationships are being studied; and (2) *how much reliability (or representativeness) is required* of the result of the sampling process.

The Variation Factor

That the mere size of the universe does not itself govern the size of a reliable sample can be demonstrated on purely commonsense grounds by taking a hypothetical example. Let us suppose that the whole population of the United States — this time in the usual sense of "inhabitants" — is exactly alike in all respects. If this were the case, quite obviously a sample of one individual would be completely reliable and representative. And this would be equally true if the population of the United States were 160 million, 200 million, or 500 million. It is because the population of the United

[50] This statement does not take into account either stratification or the exceptional cases in which a carefully safeguarded and handpicked sample may produce valid results. On such refinements the reader may consult almost any textbook of statistics.

States varies so much in so many ways that in actuality it takes many more than one individual to constitute a reliable sample.

At the other extreme of our hypothesis of a uniform population lies the actual fact that every person in the United States is an individual who differs from every other person. On this basis, however, no description of the population is possible short of a complete enumeration, accompanied by individual descriptions. This is, of course, an impracticable task. Fortunately, it is also needless in all the many cases in which — for certain specific purposes, practical or intellectual — we want to know something less than everything about the population. For a housing program, for example, we might be interested only in how many young and how many old persons there are in the population; how many single, how many married; how many small, how many large families; how many urban, how many rural dwellers.

In other words, whenever we think of a large universe (and there is no point in statistics of small universes) we think in terms of *classes* of members, and this is especially true when we think statistically.

In gearing a sample to the factor of how varied the universe is, therefore, we do not consider all the variations. All universes vary infinitely: no two leaves of a tree are exactly alike, nor are any two products of even the finest machine. We consider only those variations that bear on the problem we are trying to solve by statistical analysis. In a human population, this might be a matter of age, or of age and sex, or of age, sex, and income, and so on. Further, if it is a matter of age, it will not be on the basis of all differences in age ("there is one born every minute"), but of certain age *classes* set up for their relevance to the object of our study. For some purposes, we might set up seven classes, by 10-year intervals to age 60 and one interval called "over 60"; for other purposes, we might set up only three: "under 21," "21 to 60," and "over 60."

In short, the variations we must consider in setting up a reliable sample correspond to the classes of variations that have been prescribed for the analysis; and the larger the number of such classes, the larger the sample must be.

The Reliability Factor

It is never necessary that a statistical study be completely reliable — that is, that its results tally literally 100% with actuality. In fact, such ultimate reliability is impossible, as should be clear from what has already been said. The only complete representation of the human population of the United States is that population con-

sidered in its entirety. If even one member is left out, what remains will be — albeit infinitesimally — unreliable with respect to the peculiar characteristics of that one member. Even a complete enumeration will not tally 100% with actuality, because of the inevitability of human or mechanical error in the count. Theoretically at least it is possible to conceive that a well-designed sample will come closer to 100% than an enumeration. But it can never reach 100%.

It follows logically that a sample, being always less than the universe, must be less than 100% reliable, and that — other things being equal — the reliability will decrease as the size of the sample decreases. This of course does not mean that samples are worthless. For one thing, reliability does not decrease in direct ratio to the size of the sample. For example, a well-known political poll is considered by experts to be reliable within 4% although it is believed to be based on a sample of the voting population of only 3,000 maximum. Moreover, practical guidance to practical problems — like the planning of a housing program — can be provided by something much less than dead certainty as to the statistics of housing conditions and needs. As long as the degree of error can be foreseen, the costs of solving the practical problem, including waste due to error, can be evaluated.

Depending on what the intellectual or practical problem is, then, the degree of reliability that is required can be fixed in advance. For most problems, a sample that had only a 50-50 chance of corresponding with reality would be useless; for some, a 10% chance would be better than guesswork; for others, only finer statistical results would be worth while. And when the size of the sample is being determined, the question of the degree of error, unreliability, or unrepresentativeness that is permissible in the results will take its place next to the question of the number of relevant variations in the universe under study.

The Cost Factor

From everything that has been said above, one fact must be quite plain. The more refined and reliable the sample must be, the more costly it will be to select, process, and maintain. If the matters that it is to illuminate can be arranged into few classes, and if the tolerable degree of unreliability is large, the sample can be small. If light is to be thrown on some universe by a multiple division into classes, and if the reliability must be of high degree, the sample must be large.

For this reason, when statisticians prepare a sampling scheme they limit the sample both in respect of the refinement of its division into classes and in respect of the unreliability that they will tolerate, with a view to minimum costs for their particular purpose. They may even sacrifice some refinement and reliability to conform to allowable costs, on the grounds that a measure of refinement and reliability, of known limitations, is better than no statistical data at all.

In any case, of course, much sharp and practiced judgment is required to shape a sample to the uses for which it is drawn in other respects than refinement and reliability — in respect of its overall appropriateness.

All this involves, necessarily, another consideration to which the archivist must give heed if confronted either with a statistical sample that is to be appraised or with a universe of returns, submissions, or other forms of records that might be made the basis of a statistical sampling operation for archival purposes.

The archivist must bear in mind that there is no such thing as a demonstrably universal sample that would serve all conceivable research purposes within a given universe. The overall appropriateness, the refinement, and the reliability of a sample arise out of the particular research problem that the sample is to illuminate. If the sample is on these grounds quite large, it might be useful for other research problems in a more or less limited way. If the sample is small (perhaps because little refinement or reliability was required for its original use), the chances are correspondingly small that it can be used for other than its already-exploited use.

Therefore it will ordinarily be of little statistical use to accession a body of records that is itself a small sample of some relatively large universe. It will be of little statistical use to reduce a sample by further sampling, for — as has been said — this will ordinarily reduce and restrict the sample so that it can yield only less information than it did to begin with.

And if the question arises of taking in or creating a large statistical sample, the archivist — like the statistician himself — must consider whether the cost of selecting or maintaining the sample, and the cost to possible users of processing the sample, are commensurate with the chances of use and its results. It is not at all impossible that this ratio of cost to product might in some cases turn out favorably for retention. But in each case it deserves careful scrutiny and consultation with statisticians practiced in the subject-matter with which the sample deals.