

THE RECORDS OF A STATISTICAL SURVEY

THE primary purposes of a statistical survey are to assemble quantitative facts and to analyze and present these facts in a usable form. A survey may be organized chiefly for the collection and presentation of such data, as in the case of an official census, or the information may be collected in the course of performing some other function with the collection of statistical information entering as a by-product. In the process of collecting taxes, for example, statistics are collected to assist in determining the amount of taxes to be paid. As may be expected, the combination of several purposes in one undertaking will introduce important complications in discussing that undertaking. To avoid such complication, we shall assume that the subject of this paper is a survey for the sole purpose of collecting facts and is uncomplicated by any legal or administrative purposes.

In a statistical survey a set of definitions which determine what data are to be gathered is applied to the units which are being studied through the medium of a form of some kind. This form may be a questionnaire actually filled out by a respondent in a mail canvass, a census population schedule where the information is entered by an enumerator who asks the questions of the respondent, or merely a transcript on which information about the unit is transferred from some other source so that the data can be used in a different manner than the original arrangement would permit. Information so assembled is tabulated and analyzed to furnish facts about some problem or set of problems which require study.

In this paper the records created by such a survey will be discussed briefly and an attempt made to point out some factors which enter into a determination of their value from the viewpoint of the statistician. From the viewpoint of the historian the results would be the most important aspects of the survey. The economist and the sociologist would also attach importance to the definitions used and the units studied, in order to know how far the results can be generalized into laws and principles. The statistician is interested in all these aspects and, in addition, in the actual technique in conducting the survey so that future surveys of a similar nature can be conducted more efficiently.

Since the final results and interpretations of these results are usually included in books, monographs, or journal articles, they are subject to preservation in our libraries. This paper will therefore assume that the preservation of the published results has already been dealt with and will discuss only aspects of survey records prior to the final publication.

The records under discussion may be divided into three groups, according to the type of function which causes them to be created. These functions which form the basis of the division of the material are (1) planning, (2) actual collection of the data, and (3) processing preliminary to publication. Because of their importance, the records containing the primary data of this survey will be discussed separately.

I. Records Created in Planning

Every survey has its genesis in a problem, or set of problems, which require study. The records of the planning aspects of the survey should show why the survey was undertaken, the source and extent of the enabling means which made the survey possible, and the analysis of the problem which led to the definitions used and the manner of applying these definitions to the respondents.

In the early stages, the records will be those of outline and discussion. Books, magazine, and newspaper articles, memoranda and letters will be written about the problem, thereby emphasizing the need for information. This need will find tangible expression by the securing of enabling means for conducting the survey. For example, discussions of the statement that the housing facilities of a large segment of the American people were inadequate led, in part, to federal legislation for taking a national census of housing in 1940.

Since the books, magazines, and newspapers which open the public discussion of the problem are of a general historical nature, we need not include them as records of the survey. For our purposes the records of the survey may begin with the outlines or legislative materials prepared to secure approval of the undertaking. A summary of the problem and the outline of the plan of action as conceived before the project is undertaken should be included. In the case of the census of housing, for example, this part of the material would include the legislation, along with the congressional hearings and other reports which helped formulate this legislation and effect its passage.

When the project is accepted a number of important decisions must

be made. The various aspects of the problem must be discussed and those elements of the problem which are measurable, either directly or indirectly, must be segregated. Definitions of the material to be studied will be fixed and the appropriate sources of information chosen. A method will be determined as to how the definitions are to be applied to the sources of information and the entire framework will be crystallized in a questionnaire or transcription sheet of some type.

The significant reports, conference minutes, and memoranda which indicate the decisions made in the formulation of the questionnaire and the outlining of the procedure to be followed complete the important records of this phase of the survey.

II. *Records Created During Collection of the Data*

The records created during collection of the data may be divided into three groups: The first group should show clearly exactly what was done during the taking of the survey; the second group should contain the information collected and will be discussed in a separate section of this paper; the third group should include the operating records which will reveal the administrative details of the survey. The operating records are those created in the establishing and functioning of the organization and include the records of personnel, rate of work, finance, equipment and supplies, and other material pertaining directly to the operations of the organization.

The importance of the records which show the exact procedures followed arises out of the fact that the plans made at the inception of the survey are not always carried out to the letter. Modifications made in the field are doubly important for they may alter the results obtained and the interpretations of these results, and they may also indicate problems which should be considered in planning later surveys. This group of records should include all instructions to respondents, enumerators, transcribers, supervisors, and other personnel, memoranda, or letters modifying these instructions, public relations material, copies of all forms used in the survey, and reports made by the personnel, other than the regular operating reports. A copy of every form, even down to a specially printed envelope, should be included so that, after other records are destroyed, the needs and procedures of this survey can be reconstructed.

The importance of retaining the detailed operating records must be carefully considered, for one can never be sure what type of in-

formation will be needed in planning other studies in the future. Every survey, especially when there is an innovation in technique or coverage, should be considered an experiment contributing information for the planning and conduct of future surveys of a related nature.

The records relating to personnel, including data on rate of work, to equipment, and to costs should not be destroyed until thoroughly summarized. The additional cost of preparing such summaries will usually be repaid by savings resulting from the increased knowledge of the operations. Frequently, administrative summaries prepared in the course of conducting the inquiry are sufficiently adequate for preservation, permitting the detailed reports to be destroyed. In a survey of a repetitive nature, however, it seems advisable not only to summarize these records, but even to retain them until the plans for the next survey are completed. The greater detail makes possible research in techniques which may lead to additional savings.

III. *Records Created During Processing*

When data have been collected, the desired information cannot be derived immediately. Errors exist which can be ascertained and corrected. The results may be seriously affected unless this is done. The answers to certain questions may have been made in verbal terms, which do not lend themselves to immediate or obvious groupings. Such answers must also be analyzed to check the uniformity of the response. Some information may have been collected in great detail, but for tabulation purposes it may be desired to group the information into a relatively small number of classes.

Examination of the forms for completeness, for consistency, and for reasonableness is called editing. The combination of information into certain groups, each with a common characteristic, is known as classification. The translation of certain information into symbols, usually numbers, is called coding. Tabulation is the process of summarizing the data into convenient form for use. For convenience, all these operations are included in the term "processing."

At this stage of the survey we again have three groups of records. As above, the first group includes all the instructions, memoranda and similar documents which show exactly what was done. The second group will be comprised of all the tabulation sheets, summary sheets, and other work sheets which contain the information from which the final results were derived. The operating records which

show the rate at which the work progressed, with what equipment, and at what cost, will form the third group.

Another set of records which may be called "intermediate records" may be included if a punched card technique or some similar technique is employed for tabulation. The existence of such a set of records will enter into the consideration of the disposition of the original questionnaires.

As in the case of the records created during the collection of the data, it is extremely important to know exactly what went on and under what conditions. As before, the records which show the procedures should be retained and the operating records at least summarized if retention is not practical.

If all the details of the tabulations are published, so that no information in addition to that published can be derived from the work sheets, it is not necessary to retain that group of material. If, however, there is more detailed information on the work sheets than published, these data should be retained until any purpose which this additional detail may serve is satisfied. The permanent retention of such records, of course, will depend upon the value of the unpublished information on these sheets.

It may be pointed out that if any summary statistics are published, such as averages, estimates of errors, or similar computations, and the detailed steps in the derivation of these statistics are not included in the final publication, the work sheets showing the detailed steps should be retained to avoid the need for duplicating the work to substantiate the results and to insure the derivation of comparable statistics for some other set of data.

The Original Questionnaires

A statistical survey may be considered as a factual picture for a given population at a given period of time. When that period has passed, the dynamics of living create new details which obscure the old in the memories of the respondents and make the data on the questionnaires a set of records which cannot be duplicated and, therefore, very valuable. The first point in judging a set of such original records is how much information is contained on the questionnaires and how good a picture this information will furnish at a future date. Another consideration involves the existence of the same information on some other set of records. If information identical in time and

place is duplicated on two sets of records, neither is as valuable as each would be if the other were not available.

Tabulation usually requires the transfer of information to an intermediate set or sets of records, such as tally sheets when hand tabulation is used, or punched cards and result sheets for larger studies. The existence of this duplicate material will, of course, detract from the value of the original records. However, a complete transfer of information is not usual and the original questionnaires usually contain either more detail or more data than are found on the derived records.

Where the information has been transferred completely to tabulation sheets and all possible tabulations have been published, there is no need to keep any records other than the final report. Where this complete transfer of information has not taken place, the value of a set of records then depends, in part, upon the information it contains that is found on no other record. Thus, the value of tabulation sheets will in turn depend on the information which has not been published. The value of punched cards or other intermediate records of the same type depends upon the information on the cards which has not been tabulated. The value of the original questionnaires depends upon the information which has not been transferred to any of the other forms.

The existence of duplicate records requires additional space and the problem often arises as to which of two sets of records is to be preserved. Possible future needs and the work involved in setting up the various sets of derived records must enter into consideration. For example, the question may arise whether to destroy the original questionnaires or a set of punched cards containing almost all the information which appears on these questionnaires. The arguments which favor retention of the questionnaires are the presence of more information, and the greater accuracy of the information, since errors enter in transferring to the derived records. Sometimes the information on the punched cards is not as detailed as possible, but has been entered in group codes. At a later date it may be desired to regroup the data, which can usually be done only from the information on the original questionnaire. The chief argument in favor of retention of the punched cards lies in the fact that use of the information on the questionnaires when needed will probably involve the preparation of a set of punch cards, possibly duplicates of those which were de-

stroyed. Thus, keeping the punched cards will result in a saving of time and money when more tabulations are needed. The argument that the punched cards have a much shorter useful life may be offset by the fact that they can easily be reproduced by mechanical means.

An alternative by means of which information on a set of original records can be retained is reproduction on microfilm in such a manner that reference can easily be made to the records when necessary. This method is a means of saving space and is likely to be very useful. Due consideration should be given to possible future uses when this alternative is considered. The ease of reference and the procedures for future tabulation should be studied before such records are transferred to microfilm and destroyed. Where documents are of value mainly as evidence, microfilm will probably be a satisfactory means of reproduction.

Instead of retaining the complete file of case records, an acceptable alternative may be found in the retention of a sample of these records and the destruction of the remainder. Tabulations based on a carefully selected sample chosen with due consideration of future needs, will usually give the needed information within adequate limits of error. A technique of this kind may also permit the retention of a sample of the tabulation cards.

In this connection it is interesting to note that in the taking of the 1940 census of population, a five per cent sample was automatically segregated in the course of enumeration mainly for the purpose of securing additional data. After the 1930 census it was found necessary to destroy all the punch cards used in tabulating that census. However, if the cards for this five per cent sample can be retained after the regular census work is over, they will provide a means for making special tabulations as problems arise during the next decade. If this procedure is followed it will add immeasurably to the value of the information gathered in the largest statistical survey undertaken in the United States.

In conclusion, it may be said that since no two surveys are identical, certain records should be retained to tell the story of each survey, to contribute to the analysis of the results when questions arise, and to furnish information for later surveys of a similar nature.

When the survey is completed there will be assembled reports, memoranda, minutes, correspondence, instructions, questionnaires,

tabulation and work sheets, records of production, cost, purchases, etc., and, possibly, transcription sheets and punched cards.

The material retained should show the original objective and plan of the survey, the instructions as given and as carried out, and the type of forms used, so that the detailed procedure can be reproduced if necessary. A detailed summary should be made of the operating records so that the rate of work of the personnel, the cost of the survey, and the equipment used can be studied in detail. Tabulation sheets need be retained only if they show more detail than the published material. Finally, but not least important, the questionnaires, or copies of the questionnaires, should be retained for a reasonable length of time. If it is a simple form with only a few questions, and all the possible tabulations have been made, there is no need to keep the information after publication. When there is more than one set of records containing the same information, the question as to relative value is a rather important one and a decision as to which shall be retained cannot be generalized here.

Finally, any reports, summaries or evaluations of the various aspects of the survey made upon completion of a job should also be included. In fact, such reports might well be required of anyone in a technical or administrative capacity, and, if such reports are too voluminous, a summary should be prepared to round out the records of the statistical survey.

MORRIS B. ULLMAN

Bureau of the Census