SURVEYS: A PRACTICAL APPROACH

Nicholas J. Ciancio and John L. Stover, Statistical Reporting Service, USDA

I. INTRODUCTION

This paper will present to the student and those actually involved in the collection of survey statistics some of the problems encountered when developing, performing, and analyzing a survey. These aspects are usually omitted during the formal education of a statistician. Data used in textbooks or journal articles are usually presented without reference to the day-to-day problems of its collection, or are entirely hypothetical.

Recent literature such as Sudman [20], Sukhatme and Sukhatme [21], and Sudman [14] discuss some methodology of data collection, but for the most part the literature is entirely too theoretical concerning this topic. Examples of the latter are Cameron [2], Bryant, et al. [1], and Cornfield [6].

Therefore, the object of this paper will be to take the reader through a survey step by step. We will consider the survey as consisting of three main sections:

(a) Presurvey work,
(b) Collection of the data, and
(c) Data summary and estimation.

The discussion of the above concepts will be based on two surveys conducted in California. The first is a citrus weight study which is currently in the planning stages. The second is a survey of agricultural labor wage rates, which has been conducted by the Statistical Reporting Service of the USDA on a quarterly basis for the past decade. These surveys will be further discussed throughout this paper.

II. PRESURVEY WORK

Before the collection of data can be undertaken, adequate preparation must go into the initiation of the survey. This takes months of time. In the case of certain very large surveys, this can take several years. The first and foremost problem is to determine what information the results of the survey are to produce and who will supply this data. A target population must be determined, as well as the means by which the data are to be gathered.

Upon determining the target population, a reliable list of sampling units must be compiled. These lists are seldom complete, especially if there are a large number of participants. For the citrus weight study, the sampling unit is the packinghouse. The estimated number of packinghouses in California and Arizona is 317. In the farm labor survey, the employer of agricultural laborers was determined to be the sampling unit. There are approximately 13,000 farms and agricultural services in California employing such labor.

The method by which the data are to be collected will depend upon the results desired. For the citrus weight study it was determined to collect carton weights by size and grade so that an overall uniform weight by citrus variety can be determined. In the farm labor survey, wage rates, are desired for various types of farm workers (such as field, livestock, packinghouse, machine operators, etc.) and by method of pay (cash wages only, those given housing, those given room and board, etc.). Thus, employers are asked to give wages paid and hours worked for these different classes for a particular time period.

The problem of sample design type to be used will determine the technique of calculating sample size and allocation of the sample. For these topics, the student can find numerous texts that attack the problem. This is one of the most critical portions of the planning stage. If an incorrect sample size and allocation are used, the results of the survey will have little or no significance. If stratified random (or psuedo-random) sampling is used, estimates of the variances for each of the strata are needed. Such estimates are difficult to obtain for new surveys. Usually estimates of variances can be obtained by taking proportions of the range of possible values as determined by sampling distributions. Deming [8] gives specific ratios of the range, assuming one knows the range of values and a possible distribution.

A stratified random sample of cartons by variety was used for the citrus weight study. The sample size was calculated to be 23,400 cartons across variety by year. A Neyman allocation was used to distribute the sample into the strata. On the farm labor survey, a stratified psuedo-random sample of employers is used. Stratification is based on the peak number of farm laborers employed in a given year. The total sample is currently set at 1,645 employers in California. This sample size is adjusted periodically, based on strata variances computed from previous surveys.

The aforementioned problems of sample size and allocations cannot be taken too lightly. Much approximation goes into setting the sample size for the first time. However, for an ongoing survey, estimates can be made from previous year's data. At this point, the accuracy and the percentage confidence levels of the estimate are determined. Usually the statistician has to calculate sample sizes for various confidence levels and error rates so that a complete cost analysis can be drawn up.

The budget is of prime consideration. If costs of enumeration are high, then the sample size will be low. The largest portion of the budget will be wages paid to enumerators. Other costs to be considered include mileage paid to enumerators for the use of their cars, telephone expenses, training of enumerators and clerical staff, questionnaire design and printing, field equipment, computer expenses, and so forth. Other costs, depending on the type of survey, will also need to be included. Once the budget is exhausted, the survey is complete.

After the sample size and allocation are determined, the actual drawing of the sample from

the list occurs. These sampling units can then be plotted on a map to determine the number and location of enumerators and supervisors necessary to run the survey. Hiring of enumerators is a long and tedious task, particularly when there is no list of enumerators who have worked on similar previous surveys. Even if there is a list of enumerators, new people may have to be hired. It is recommended that some sort of screening device be used, preferably a test to determine the applicant's literacy, abilities to perform simple arithmetic calculations, read maps, and follow directions. Such a test has been found to be quite useful for hiring enumerators on surveys conducted by the Statistical Reporting Service. Inherent in this process is the determination of the qualifications necessary for the position. Overall the practice of hiring involves traveling to a certain area and staying in a set place for several days to conduct interviews. Carefully written advertisements placed in local newspapers and trade publications announcing the interview time and place are helpful. Other names of potential candidates can be obtained from currently employed enumerators, and from various county and municipal agents involved in a general way with the particular industry or segment of society to be studied. These last sources should not be relied upon exclusively if the evils of racism and sexism are to be avoided.

For the citrus weight study, it was determined that 70 packinghouses would suffice to obtain the necessary data using 7 enumerators to collect the data. On the farm labor survey it has been found that 65 enumerators and 5 supervisors are needed. When dealing with such numbers of people, who are really only intermittently employed on these types of surveys, it should be remembered that extra employees should be hired. Before any given survey, and frequently during the course of the survey, enumerators will quit (due to illness, death in the family, dislike of job, etc.) or will have to be terminated for incompetence. Their workload will then have to be reassigned to the remaining employees.

During this time frame, a questionnaire must be designed, printed, and tested in the field. It is necessary that the questionnaire be worded so that it is easily understood by the enumerators and respondents. Its execution must be well thought out so that it will provide precisely the data of interest. It is also very helpful to have it in a format so that key punchers can transcribe the data to computer cards. The questionnaire should not be too long if personal interviews are to be conducted. A long and thick questionnaire will increase the rate at which respondents refuse to cooperate as well as have a detrimental effect on the quality of the data obtained towards the end of the interview due to respondent fatigue.

Once the form is finalized, it is sent to the printer. It is recommended that galley proofs be obtained and checked for errors before the entire lot of questionnaires is printed. Also, a field test is useful to solve any problems that might crop up. Another item to watch for on Nationwide surveys is different meanings ascribed to certain words in various regions. Survey designers should take care in wording the questionnaire so that each question is understood uniformly by the respondents.

Preparation of computer software is necessary before the data collection is actually underway. This should include an edit system, summary program, and possible estimation procedures. These can either be prewritten software packages or self-written, tailored to individual needs. Sample data should be entered to check all possible errors and to insure that the summary, edit, and estimation work properly. Edit limits must be determined along with input formats for data entry.

Sets of instructions are needed for office procedures. These instructions should be exhaustive for preparation of enumerator supplies and their distribution, checking in questionnaires as they arrive, the clerical edit, statistical edits prior to key punching, and so forth. Procedure must also be documented for keeping track of enumerator assignments. The clerical staff must be schooled in exactly what is expected of them.

Similarly, instructions for enumerator procedures must be developed. They must also be self-sufficient and cover as many problem situations as possible. Included in these instructions are also procedures for filling out forms relating to the specific survey (such as time and mileage sheets, accident reports, overtime pay, and so on).

The equipment needed in the field should be ordered well in advance so that it arrives before the start of the survey, and there is sufficient time to allow for lost shipments and shipment of the wrong goods. Detailed inventories should be kept at all times of the location and quantity of supplies. In an on-going survey, care should be taken to insure that supplies are returned to the control of the statistician at the end of the survey.

A kit for each enumerator is made with enough equipment to carry the enumerator through the survey. These kits can be distributed at the training schools for enumerators. These sessions should be planned so that a reasonable number of people are in attendance. For large surveys, several schools may be necessary in different locations. These functions reiterate the material covered in the instruction manuals and disseminate administrative procedures. A classroom with proper lighting, space, and air-conditioning (or heating, depending on the season) is necessary. It is well worth the cost to rent such a room rather than make do with facilities that are not adequate. Enumerators need to be notified well in advance of the time and location of their respective schools. Should the school run more than one day, motel reservations will need to be made for those enumerators living out of the area in which the school is held. While conducting the school, practice interviews are made with most problems being covered. This gives the statistician the opportunity to "weed out" those employees who will not be able to perform adequately. The time frame of the survey should be discussed so that the enumerator will have an idea of what percentage of his work should be completed by specific dates during the course of the survey. At the end of the school, assignments and supply kits can be distributed.

III. DATA COLLECTION

During the training and hiring of enumerators, supervisory enumerators are also selected. They will have the important role of coordinating between the field and the statistician. The main assignment of the supervisor is to insure that quality data is collected. This can be done by quality control checks on the enumerators. This entails great effort and time, but is money well spent. A subsample of completed questionnaires can be made, and then verified, on a survey involving counts or objective measurement. This is more difficult on an interview type survey, as the respondent will be reluctant to give another interview. In this case, the supervisor can look over completed questionnaires for internal consistency and reasonableness. Any errors found can be corrected, and the enumerators informed of their mistakes. Should there be serious problems that cannot be corrected by the enumerator, the enumerator should be terminated, his assignment picked up and redistributed to employees who can do the job.

While the survey is in progress, the main problems are to insure that: the enumerators have sufficient supplies, they know their assignments and time frame, they actually collect the data in the field (as opposed to their living-rooms), and they submit this data to the statistician. Due to time limitations, it may be necessary to have the last few days' work shuttled to the statstician rather than relying on the postal service.

Once the data is in the office, it is checked in, clerical and statistical edits are made, and then the data is submitted to key punching. The backlog in editing, especially toward the end of the survey, may be a cause of concern.

Adjustments to assignments, and the possible hiring of new enumerators (done only as a last resort) are usually made during the course of the survey by the supervisors. Basically, the supervisor makes sure the data is collected correctly and returned on time.

At the end of the survey, it is necessary to collect all unused materials for reuse. Also, an evaluation of enumerator performance, based on supervisor reports and quality of data received in the office is very helpful.

Let us assume at this point of the survey that all the data that is going to be submitted has been edited and key punched. This does not mean all of the data has been submitted to the office. A certain percentage of the sample will not be accessible during the survey, and there will be respondents who refuse to cooperate. A decision must be made in how to handle these missing reports. If the estimate has to be submitted by a certain date, then a strict time-table must be adhered to. This means if data comes in after a certain point in time, it will not be used.

IV. DATA SUMMARY

A summary of the data can now be obtained since all the submitted data has been "cleaned" both clerically and statistically. The computer summary should include the raw data in tabular form by strata. Counts should be made on the number of completed and usable questionnaires. When more than one measurement is made on one unit, an analysis of variance table is helpful.

Expansions are calculated by dividing the sample size into the population size for each strata. To calculate an estimate, strata totals are multiplied by appropriate expansion factors. Estimates for missing reports must be included in strata totals.

If the summary is estimating a total that does not yield the answer needed directly, then some statistical technique is needed to approximate the final estimate. Some techniques are regressions, time series, chartings, and so on. Standard errors can be calculated by applying the formula applicable to the sampling technique employed.

V. CONCLUSION

The previous sections just touch upon some of the practical considerations that must be made in setting up and conducting a survey. As can be seen, setting up the survey consumes most of the time involved with the survey. Usually, time restrictions make the collection and analysis of the data move quickly. These sections of a survey pertain mainly to the types of surveys that the authors have encountered. Other surveys may have unique problems not discussed.

Briefly, a summary of main problems to watch for are:

(1) Budget -- Be very careful not to overrun the allocated funds. Give yourself sufficient financial room to operate.

(2) Hiring and training of quality enumerators -- This is essential to the reliability of the data. If poor data is the foundation of a project, then nothing but trash will be obtained, no matter how sophisticated the analysis.

(3) Time schedule -- Prepare well in advance of survey starting date. Whatever can go wrong probably will, so give yourself sufficient time to deal with various crises.

(4) List building -- Constantly revise and update the universe list. Especially helpful to enumerators are telephone numbers, physical addresses, (as opposed to post office box number), who to contact if interviewing a business, and who not to contact (John Smith, Sr. might give you the data, while John Smith, Jr. might drive you off with a shotgun).

(5) Sample size -- Revise according to the list. Proper estimating techniques should also be updated.

In this brief paper, the authors have presented some real world problems. Hopefully, this will help the reader who sets up and conducts surveys.

REFERENCES

[1] Bryant, E.C., Hartley, H.O., and Jessen, R.J., (1960). "Design and Estimation in Two-Way Stratification." J.A.S.A., 55, 105-123.

[2] Cameron, J.M. (1951). "The Use of Components of Variance in Preparing Schedules for Sampling of Baled Wool." Biometrics, 7, 83-96.

[3] Church, B.M. (1954). "Problems of Sample Allocation and Estimation in an Agricultural Survey." J.R.S.S. (B), 224-235.

[4] Cochran, W.G., (1963). "Sampling Techniques." John Wiley and Sons, Inc.

[5] Cornell, F.G., (1947). "A Stratified - Random Sample of a Small Finite Population." J.A.S.A., 42, 523-532.

[6] Cornfield, J., (1944). "On Samples from Finite Populations." J.A.S.A., 39, 236-239.

[7] Cox, D.R., (1952). "Estimation by Double Sampling." Brometrika, 39, 217-227.

[8] Deming, W.E., (1960). "Sample Design in Business Research." John Wiley and Sons, Inc.

[9] Fisher, R.A., (1973). "Statistical Methods for Research Workers." 14th ed., Hafner Publishing Company, New York.

[10] Hansen, M.H., Hurwity, W.N., and Madow, W.G., (1953). "Sample Survey Methods and Theory." John Wiley and Sons, Inc., Vol. 1.

[11] Jessen, R.J., (1942). "Statistical Investigation of a Sample Survey for Obtaining Farm Facts." Iowa Agr. Exp. Sta. Res. Bull., 304, 7-59.

[12] Jessen, R.J., and Houseman, E.E., (1944). "Statistical Investigations of Farm Sample Surveys in Iowa, Florida, and California." Iowa Agr. Exp. Sta. Res. Bull., 329, 265-338.

[13] Johnson, F.A., (1943). "A Statistical Study of Sampling Methods for Tree Nursery Inventories." Journal of Forestry, 41, 674-679.

[14] McVay, F.E., (1947). "Sampling Methods Applied to Estimating Numbers of Commercial Orchards In a Commercial Peach Area." J.A.S.A. 42, 533 540.

[15] Milne, A., (1959). "The Centric Septematic Area - Sample Treated as a Random Sample." Biometrics, 15, 270-297.

[16] Osborne, J.G., (1942). "Sampling Errors of Septematic and Random Surveys of Cover-Type Areas." J.A.S.A., 37, 256-264.

[17] Politz, A., and Simmons, W., (1949). "An Attempt to Get the 'Not at Homes' Into the Sample Without Callbacks." J.A.S.A., 9-31.

[18] Snedecor, G.W., and King, A.J., (1942). "Recent Developments in Sampling for Agricultural Statistics." J.A.S.A., 37, 95-102.

[19] Sudman, S., (1972). "On Sampling of Very Rare Human Populations." J.A.S.A., 67, 335-339.

[20] Sudman, S., (1976). "Applied Sampling." Academic Press, New York.

[21] Sukhatme, P.V., and Sukhatme, B.V., (1970). "Sampling Theory of Surveys with Applications." Iowa State University Press.

[22] Tukey, J.W., (1950). "Some Sampling Simplified." J.A.S.A., 45, 501-519.

[23] Yates, F., (1949). "Sampling Methods for Censuses and Surveys." Hafner Publishing Company, New York.