DISCUSSION

Joseph Waksberg, Westat, Inc.

I. Use of Loss Functions to Determine Sample Size in the Social Security Administration

This paper correctly points out that sampling textbooks and sampling theory usually start out in the middle of a problem, that is they assume there is advance knowledge of the desired variances for the sample survey under consideration. This is, of course, one of the early issues that arise in developing plans for surveys, and in my experience it is one of the most difficult ones to resolve in a satisfactory manner; mostly because the consequences of errors generally have not been quantified. Similarly, in teaching courses in sampling theory and methods, I have always found it difficult to give students a "feel" for the size of sampling errors that might be desirable under different circumstances.

I am therefore delighted to see examples in which the problem is approached rationally, and in a way that permits the power of mathematics to be used to derive the variances that appear appropriate for specific surveys. I think it would be useful to establish a file of similar cases, both to have as examples in teaching and to help a consultant direct the thinking of survey sponsors, who are generally perplexed about how to come to grips with the problem of the precision needed for their studies.

It would probably be naive to expect that the planning of all (or perhaps even most) surveys can be approached by establishing a loss function and comparing the loss with the cost of conducting surveys of various sizes. Typically, surveys are conducted to add to general knowledge, or to permit an administrator to make better decisions by use of the results, with or without other, external information. It's unlikely that in those circumstances anyone can even guess at the monetary value of the improvement in decisions, or for that matter even whether there are any improvements as a result of the surveys. However, I believe examples like the ones in the paper can help direct thinking in even these cases.

I suspect that considerable more work needs to be done on the construction of appropriate loss functions. Two aspects of the function used bothered me. First, expressing the loss function as $\Theta = E(L) + C$, implies that the cost of the survey is of the same level of importance as the same amount of money distributed in error. This implies an indifference to how an equivalent amount of money is spent. If I were administering a program and was told I could spend an additional $1,000,000 on a survey to get zero error, or risk an error of $1,000,000, I would prefer to skip the survey and just add the $1,000,000 to the money allocated. It would be interesting to give a higher weight to the cost of the survey, and see the effect on the results.

The other aspect is the one discussed in the paper, whether the loss should ignore the direction of error or just consider the situation from one of the parties concerned, and minimize errors for that party. This principle appears troublesome. If carried to a logical conclusion, it says that in the program the Federal Government should strive to make all its errors in one direction -- at the expense of the States. Even at the risk of introducing more arbitrariness in the construction of the loss function, I wonder if including the effect of errors in both directions would not be more realistic. The loss functions are bound to have a certain element of subjectivity but hopefully they can be expressed so that most analysts agree that they are reasonable.

I have one final comment about other areas in Government programs in which it would be useful to apply similar types of analyses. In the past few years a number of very large revenue sharing programs have been developed, both general and special-purpose. They involve the transfer of billions of dollars a year from the Federal Government to States and localities. They have certain characteristics in common; they rely on allocation formulas which are based on statistical data, and there has been a general reluctance to spend much money to produce current, reliable estimates of the parameters required for the formulas. The staffs of the agencies involved would perform an important public service if they established reasonable loss functions and examined the implications for survey operations, in a way similar to that done by Social Security Administration.

II. Effect of Counting Rules on Sampling Errors and Costs

Dr. Sirken's paper on counting rules and the papers published earlier on the same subject have, I think, served a useful purpose in indicating a direction in which greater flexibility can be exercised in survey planning, with a potential improvement in efficiency. I was glad to see an example which provided guidance on the conditions for which multiplicity rules would be expected to be more efficient than conventional

enumeration rules. The published papers had left me with a feeling that it would be extremely difficult to carry out the theory in practice,due to inability to estimate the necessary parameters. The example in this paper helped in this,although I still am not sure how I would tackle specific problems. Some more general guidelines on this would be helpful.

In the absence of such guidelines, or of a general body of knowledge that might become available as more surveys use multiplicity rules, I can see two important areas that lend themselves to the use of multiplicity. The first is in the estimation of a rare item in the population. Some of the earlier work by Sirken and others, using this technique, involved estimating the prevalence of certain rare health conditions. Similarly, one might think of studies aimed at estimating the characteristics of specific ethnic groups, persons with specialized educational background or training, or other low-frequency elements of the population for which establishment of a separate sampling frame is not possible. Such surveys are generally considered as comprising a screening operation to identify the subset of a sample that belongs to the group being studied, and an enumeration phase. Multiplicity rules should have a useful effect on the screening operation, certainly in cases where the enumeration can be performed at the initial sample unit, but even where additional travel is necessary for interviewing purposes. Obviously, this can only be done where there is a reasonable assurance that response errors would be kept under control.

The second area is where multiplicity rules serve to reduce response errors. Dr. Sirken mentioned one example in which this happens, estimating births and deaths. Another example involving work done at the Census Bureau, may be instructive. This was developed with a simpler concept in mind, but it can be viewed as an example of a counting rule application.

The example parallels the problem of measuring births and deaths, but with application to housing. As part of the past few Censuses of Housing, there have been sample surveys designed to measure, among other things, the change in housing arising from mergers,in which two or more apartments in a building are combined to form a single unit, or from growth when one housing unit is subdivided into several units. Associating such changes with unique housing units, although theoretically possible, is difficult for interviewers to perform properly. It is much more accurate to view the building as a whole and compare the number of units now present, with the number at an earlier point in time. The Census practice has thus been to select a sample of housing units,and essentially to perform a complete enumeration on all buildings in which the sample units are located. In estimating universe totals, the probabilities of selection are, of course, taken into account.

This can be considered an example of a multiplicity rule. It is a rather simple type since each enumeration unit is linked to only one network. At the Census Bureau, this procedure has been viewed as an example of PPS selection rather than as a case of multiplicity, but in this simple case, the two merge.

III. Optima and Proxima in Linear Sample Designs

There isn't very much that I can add to Dr. Kish's paper. He has shown, in a rather elegant manner, that a common approach is possible for analysis of what I had always treated as somewhat different problems.

Based on my own experiences, I suspect that the paper will be of greater interest to teachers of courses in sampling methods than to practicing statisticians. Working with problems of sample design, it did not take very long for me to realize that it is rare to get good estimates of either unit costs or variance, and possibly even more important that the models used for cost functions are crude approximations to reality. I imagine that most statisticians quickly become accustomed to dealing with approximations and don't worry too much about it.

I've found it more difficult to get a "feel" for this across to students. It's obviously important for them to learn about optimization and the applicable formulae. However, they also need to find out that in the real world, crude guesses about the parameters are frequently necessary. I have never been satisfied with the vague words I have had to use in indicating that considerable variation from the optimum allocation is usually possible without great increases in variances. There are few actual examples in the literature. The Hansen-Hurwitz-Madow textbook does have several tables for specific designs indicating the range of variation in parameters that can exist without important additions to variance. I can think of very few other examples.

A comprehensive discussion of this subject is thus very welcome.

THE ROLE OF THE SOCIAL SECURITY NUMBER IN MATCHING ADMINISTRATIVE AND SURVEY RECORDS

JOSEPH STEINBERG
Chairman
Bureau of Labor Statistics

General Session Introduction..... 127

Current Population Survey Reporting of Social Security Numbers -- LINDA VOGEL and TERRY COBLE, Social Security Administration..... 130

Searching for Missing Social Security Numbers -- BETH KILSS and BARBARA TYLER, Social Security Administration..... 37

Validating Reported Social Security Numbers -- CYNTHIA COBLEIGH and WENDY ALVEY, Social Security Administration..... 145

Selected Bibliography on the Matching of Person Records from Different Sources..... 151

Discussion -- ROGER HERRIOT, U. S. Bureau of the Census..... 155

GENERAL INTRODUCTION TO THE SESSION ON THE ROLE OF THE SOCIAL SECURITY NUMBER IN MATCHING ADMINISTRATIVE AND SURVEY RECORDS

Because of the interrelatedness of the papers given at this session, the authors felt that readers of the proceedings would find it easier to follow the presentations if an introduction were provided first. To this end, a number of the remarks made by the individual speakers (including remarks made by the session chairman, Joseph Steinberg) have been brought together here.

GENERAL BACKGROUND

The Social Security Administration (SSA) and the Census Bureau have, for quite some time, engaged in joint efforts to improve the quality of statistical output in the areas of income distribution and redistribution. One of these studies, which is currently underway, involves matching information on the March 1973 Current Population Survey (CPS) with earnings and benefit data from Social Security records. The work being reported on at the session was done in connection with this "1973 Match Project."

1973 Match Project.--The 1973 Match Project differs in several respects from earlier linkages between Census Bureau surveys and Social Security administrative information. Three of the major differences are:

1. The sample involved, consisting, as it does, of over 100,000 individuals 14 years or older, is many times larger than that used in any previous joint project. (Matching studies made to evaluate decennial census data [e.g., 60,104,139] have been on the order of one-fifth as large or less. 1/Previous linkages completed between CPS and SSA [e.g., 102] have been based on samples only about one-sixth as large.)

2. The process used to bring together the data from the various sources is more automated than formerly. This was one reason a larger sample of cases could be matched. Also, the fact that the major components to be linked are machine-readable promises to make it possible to publish at least some of the principal findings from the project in 1975. (However, as the second paper at the session makes clear, there are still certain manual-clerical steps which are essential.)

3. Prior joint Census Bureau-SSA "exact" match studies have focused principally on the analysis of response, nonresponse, and coverage errors. In the 1973 work, considerable emphasis is also being placed on obtaining a microdata file in which CPS reporting has been "corrected" or, more properly, calibrated. What is planned are not only modifications to the survey income amounts made possible by the presence of a comparable administrative figure obtained by means of an exact match (calibration), but also adjustments will be introduced by using synthetic or "statistically" matched information. (In the past, the statistical matching work done with the CPS has essentially been conducted independently of exact matching efforts. 2/)

The subject matter content of the 1973 Match Project is quite similar to that in earlier CPS-SSA linkages. The items being extracted from SSA's benefit and earnings files, for example, are about the same as in the efforts directed by Joseph Steinberg [e.g., 118,121] when he was at Social Security. Also, just as in some of the previous studies, information from income tax returns will eventually be included on the files, making a three-way linkage. The Internal Revenue Service (IRS) data that will be available, 3/ however, is far more limited than in the past (so limited, in fact, that it will be necessary to supplement it with IRS data introduced by using statistical matching).

Confidentiality Arrangements.--One of the things that the 1973 Match Project has in common with earlier linkage efforts is the great care that is being taken to insure the confidentiality of the shared information. The laws and regulations under which the three agencies operate impose very definite restrictions on such exchanges, and special procedures have been followed throughout so as to adhere to these provisions.

Information on the confidentiality requirements which governed prior linkage projects involving the Current Population Survey can be found in [102] and [119], which were available as handouts at the session. 4/ The 1973 work has operated under procedures which are at least as stringent as those imposed in the past. This is particularly the case for the IRS data. The details of the 1973 arrangements were also available as a session handout and are incorporated in Roger Herriot's discussion comments which appear at the end of the proceedings for this session.

SESSION FOCUS AND ORGANIZATION

The Social Statistics Section has had over 20 invited and contributed papers5/ the last decade or so devoted, either in whole or in part, to matching and data linkage. This, however, is the first time an entire session has been given to a matching study in which the primary piece of identifying information was the social security number (SSN).

Matching with the SSN.--The problems which arise when using the SSN to link Current Population Survey interview schedules to Social Security records differ in degree, but not in kind, from the problems other "matchmakers" have had. The three major factors to consider still are: (1) reporting differences in the identifying information being used to bring the files together, (2) omissions or incompleteness in the identifiers, and, finally, (3) nonuniqueness of identifiers.

In the 1973 study, as in prior CPS-SSA linkages, the chief difficulty that had to be faced was incompleteness in the identifying information. The first two papers at this session describe the situation that existed in this regard and what has been done about it so far. Next in importance were reporting errors in the social security number or in the other identifiers (name and date of birth, etc.). These are the subject of the third paper at the session.

The problem of nonuniqueness also exists with the social security number. It is estimated that more than six million people have two or more SSN's. In well over half of these cases, SSA has cross-referenced the numbers so the multiple reports for an individual can be brought together rather routinely. For most of the remaining persons, the numbers were issued in the early days of the program and probably are no longer active.6/ There are also some instances in which more than one person uses the same SSN. Fortunately, however, this situation is quite rare. 7/ Thus, compared to reporting errors and omissions in the identifying information, the nonuniqueness of the SSN's poses a relatively minor problem for the 1973 Match Project. In any event, the papers given at the session do not deal directly with the procedures that will be followed to mitigate its effects.

Nature of Papers.--The papers are reports on work in progress. For the most part, they are descriptive and nontheoretical. No attempt has been made in the presentations to set forth in a systematic way all of the procedures that have been followed in the 1973 Match Project. Just some of the important highlights which were felt to be of general interest have been given.

At the session itself, extensive tabular material supporting the results in the papers was provided as a handout. For reasons of space, these tables cannot be included here; however, they are available on request. 8/ The papers, as shown in these proceedings, follow the remarks of the participants quite closely, except for comments which appear in footnotes. The footnotes have been used to introduce parenthetical information which, in many instances, was not part of the actual presentations, to clarify points about which questions were raised during the general discussion which followed the talks, and to cite the relevant literature when this could not be conveniently done in any other way.

ACKNOWLEDGEMENTS

The authors would like to conclude this introduction by acknowledging the extensive assistance given them by Fritz Scheuren, who organized the session, and H. Lock Oh, who prepared the tabular material. Denton Vaughan oversaw the production of the copy for these proceedings on the IBM/360 Administrative Terminal System (ATS). Lois Gale, Alda Seubert, Catherine Murphy, and Tillie Mazor adapted their typing skills to specialized ATS procedures necessary for putting the material in galley form. Shirley Carter provided valuable technical assistance in preparation of the charts.

The authors would also like to thank Joseph Steinberg for his thoughtful participation as chairman and Roger Herriot, who is in charge of the Census side of the 1973 Match Project, for his role as discussant.

FOOTNOTES

1/ The citations given in square brackets here and elsewhere refer to references listed in a bibliography on matching studies which was handed out at the session and which is included in these Proceedings.

2/ For a brief historical sketch of the statistical (and exact) matches which have been done with the CPS , see Benjamin Okner, "Data matching and merging: an overview," in the Annals of Economic and Social Measurement, vol. 3, 1974, pp. 347-352.

3/ For this project, IRS made available to the Census Bureau magnetic tape abstracts of limited income information from tax returns, subject to the confidentiality arrangements discussed in this introduction and in remarks of the discussant. The dollar items abstracted consisted of total income, salaries and wages, dividends, and

interest. Codes were also included to indicate the type of return filed (e.g., joint, surviving spouse, etc.), the types of schedules used (e.g., Schedules C, D, F, etc.), and the number of exemptions claimed.

4/ Also handed out at the session were two other reports which deal more generally with various aspects of Social Security's statistical research work. These were a paper by Joseph Steinberg and Heyman Cooper, "Social Security statistical data, social science research, and confidentiality," which appeared in the 1967 Social Security Bulletin (pp. 2-14) and a 1973 SSA publication entitled Some Statistical Research Resources Available at the Social Security Administration.

For an historical summary of the Census's general provisions with regard to confidentiality, see the address given by the Bureau's current director, Vincent Barabba, which appears elsewhere in these Proceedings. See also, Robert Davis, "Confidentiality and the Census, 1790-1929," in Records, Computers and the Rights of Citizens, Report of the Secretary's Advisory Committee on Automated Personal Data Systems, U.S. Department of Health, Education and Welfare, 1973.

5/ Of particular note is the 1963 session, "Matching of medical, social and economic records for research purposes," [39,95,107]; and, also, the 1965 session entitled, "Matching of Census and vital records in social and health research: problems and results," [58,91,94,98, 112,113].

6/ It is possible, by request, and for good cause, to have more than one number; however, individuals may forget that they have a number or forget what it is, and apply again. When such inadvertent multiple issuances occur, routine administrative processes usually detect them eventually.

7/ The most important cases where more than one person is using the same SSN arise because social security numbers have been employed on occasion by advertisers in promotional schemes. Perhaps the best known such instance is the number 078-05-1120. It first appeared on a sample social security number card contained in wallets sold nationwide in 1938. Many people who purchased the wallets assumed the number to be their own. It was subsequently reported thousands of times on employers' quarterly reports; 1943 was the high year, with almost 6,000 wage earners listed as owning the number. Even today the number is still being reported at least 10 times a quarter.

8/ To obtain the tabulations, write to Dr. Benjamin Bridges, Chief, Long Range Research Branch Division of Economic and Long-Range Studies, Office of Research and Statistics, Social Security Administration, 1875 Connecticut Avenue, Washington, D.C. 20009.