FEDERAL AGENCY PRACTICES FOR AVOIDING STATISTICAL DISCLOSURE: FINDINGS AND RECOMMENDATIONS

by

Thomas B. Jabine, Social Security Administration
John A. Michael, National Center for Education Statistics
Robert H. Mugge, National Center for Health Statistics

Statisticians are becoming increasingly concerned over the need to avoid statistical disclosure, i.e., the revelation of confidential information about identifiable (but not identified) individual persons or organizations through published statistical tables and microdata tapes (computerized records pertaining to individual statistical units). For example: a published table might indicate that all male retirees in a given community receive the maximum social security benefit, thus disclosing the benefit amount for each retiree; or a published micro-data tape might give the details of health conditions of a female who according to the tape is over 100 years of age and there is only one such individual in the identified community.

This paper reports on an effort to examine statistical disclosure in the extensive and complex statistical programs of the Federal Government. People over the nation are constantly entrusting statistical agencies with various kinds of information about themselves, on the promise that the information will be used only in anonymous form, for purposes of statistical analysis. Federal agencies have a serious obligation to protect these data from statistical as well as any other kind of unauthorized disclosure.

What are Federal statistical agencies doing to prevent statistical disclosure, how well are they succeeding at it, and what more needs to be done on a government-wide basis to minimize the possibility of statistical disclosure? To answer these questions was the charge of the Subcommittee on Disclosure-Avoidance Techniques, established early in 1976 by the Federal Committee on Statistical Methodology, which is sponsored by the Statistical Policy Division of the Office of Management and Budget.1/

The Subcommittee began its work by studying the rules, regulations, and policy statements of Federal agencies relating to statistical-disclosure avoidance. The literature was then searched and relevant articles and reports were located and studied. The Subcommittee received reports on various relevant agency experiences and discussed and analyzed them. A number of actual examples of disclosure were found and considered. (To the best of our knowledge, none of these actually caused any harm, and none have ever been noted outside of the Subcommittee and the agencies which perpetrated them. However, some of them were considered by the Subcommittee to be unacceptable.) Finally, the chapters of the final report were drafted by Subcommittee members, these were intensively reviewed by the Subcommittee and revised, and the Subcommittee reached a reasonable degree of consensus on all points in the final report, which should be ready for the printer before the end of September 1977.

The Subcommittee's report is organized as follows:

The first chapter is an introduction, explaining the charge to the Subcommittee and its auspices and operating procedures.

Chapter II tackles the definition of statistical disclosure. Various previously used definitions are cited and evaluated. A definition proposed by Dalenius is found to be most useful: "If the releases of certain statistics makes it possible to determine a particular value relating to a known individual more accurately than is possible without access to those statistics, then a disclosure has taken place."

This definition is very broad and is not intended to be the basis for agency operating decisions. But neither do the definitions implied in the laws and regulations relating to confidentiality provide such a basis. In fact, absolutist definitions are useless in identifying disclosures which might be both necessary and acceptable for a given statistical program. It must be recognized that the release of some data in potentially identifiable form is justifiable under certain circumstances. Thus, the acceptability of disclosure risk in any given situation must be evaluated.

The Subcommittee found that published tabulations present quite a different set of conditions and problems concerning statistical disclosure as compared with public-use microdata tapes. Therefore, separate presentations are made. Chapter III deals with statistical disclosure in published tabulations. Different kinds of disclosure in statistical tabulations are defined and discussed.

Disclosures may be exact or approximate; they may be probability-based or certain; they may be direct or indirect; they may depend on external or internal data analysis; and they may relate to count data or magnitude data, each having a different set of implications. Depending upon the type of disclosure and its context, dh_ dJsk of actual revelation of confidential data may be great or small, so it is necessary to evaluate these risks before deciding what steps to take. Various disclosure-avoidance techniques which may be used in the case of tabulations are described and evaluated.

Chapter IV discusses potential disclosuri and their avoidance in connnection with the fast-burgeoning Federal agency programs involving the release of public-use microdata tapes. Several factors bear upon the likelihood of a disclosure taking place through a given microdata tape--the sampling fraction used in a survey, the detail of geographical descriptors, degree of detail given on the data subject's characteristics, existence of data for the same individuals in population registers, errors or noise in the data, and the age of the data. Two classes of risk are evaluated: first the risk of disclosure about a particular individual of interest; and second, the risk of disclosure of information on some identifiable individual through a "fishing expedition." Disclosure-avoidance techniques are described and evaluated --eliminiating small-group categories, allowing no unique cases, introducing noise into the data, removing known individuals from the file, and releasing files only for controlled, restricted usage.

For many statistical programs the only sure way to eliminate the risk of disclosure completely would be by refraining from any release of microdata tapes whatsoever, and by reducing published tables to a few broad and bland ones. Yet the release of public-use microdata tapes needed by the research community, together with far more detailed published tabulations, may entail a disclosure risk which, while not absolute zero, is extremely low. Decisions must be made on the proper balance between the community's needs for statistical information relevant to public policy issues and the individual's need for confidentiality protection.

Chapter V is devoted to this crucial question of balance. It reports on the Subcommittee's vain attempts to discover any cases in which an individual has been harmed through statistical disclosure, and it describes ongoing research into the public's attitudes on these questions.

The Subcommittee found that in actual practice, agencies are rarely confronted with problems arising from statistical disclosures, or even from public fears that such disclosures might take place. On the other hand, agencies receive many complaints from data users on the restrictions to data availability resulting from disclosure-avoidance practices.

The final chapter (VI) summarizes the Subcommittee's findings and lays out its recommendations to Federal agencies on the avoidance of statistical disclosure. The draft of Chapter VI is presented below in its entirety:

CHAPTER VI - Findings and Recommendations

A. The Concept of Statistical Disclosure

Findings: Several of the major Federal statistical agencies have developed and applied a variety of disclosure avoidance techniques in connection with the release of statistical tabulations and microdata files (files of individual records with identifiers removed). However, it appears that little attention has been given to defining exactly what constitutes disclosure and how to decide which disclosures are acceptable and which are not.

A few statisticians, notably Fellegi, Hansen and Dalenius have suggested formal definitions of statistical disclosure. This Subcommittee has adopted the definition proposed by Dalenius as a framework for its discussion and review of disclosure-avoidance techniques. The Dalenius definition is broad in scope. It was not the intention of Dalenius, nor is it ours, to recommend or imply that statistical disclosure so defined should never be permitted to occur. If that position were adopted, the present output of statistical information would be drastically reduced. We have adopted this broad definition because we believe it offers the best basis to

1. Identify all potential disclosures in connection with proposed releases.

2. Decide which of these potential disclosures are unacceptable.

3. Use appropriate techniques to prevent unacceptable disclosures.

Tne formal definition of disclosure adopted by the Subcommittee appears in Chapter II, pp.17-25. It can be summarized here by saying that disclosure takes place if the release of tabulation or microdata makes it possible to determine the value of some characteristic of an individual 2/ more accurately than would otherwise have been possible.

B. Deciding What to Release

Findings

1. Federal statutes and regulations governing the release of statistical information in the form of tabulations and microdata do not and cannot provide a clear basis for deciding in each case what must be done to avoid disclosure. Agencies that address this issue are obliged to strike a balance between the requirement to protect the confidentiality of information about individuals and the need for detailed statistical information and records for public policy purposes.

2. The use of microdata files by social scientists and others has developed rapidly since 1960. Microdata file users are becoming increasingly adept at handling these files and are applying sophisticated analytical techniques to exploit them fully. This development has significantly increased the utility of statistical data bases created by Federal agencies from censuses, surveys and administrative records and promises to do so even more.

3. The Privacy Act provision concerning the "disclosure" of certain microdata files (552 a(b)(5)) is ambiguous and has resulted in considerable uncertainty as to the circumstances under which microdata files can be released.

4. The Subcommittee has identified several examples of statistical disclosure which, in our opinion, were not acceptable. Some of those involved potential disclosures of salaries or benefit amounts of specific individuals. We also find, however, that most agencies that release statistical information are becoming increasingly sensitive to the disclosure issue, and that they have adopted or are in the process of adopting policies and procedures designed to avoid unacceptable disclosure (see agency statements in Appendix A).

Recommendations

B 1. All Federal agencies releasing statistical information, whether in tabular or microdata form, should formulate and apply policies and prodcedures designed to avoid unacceptable disclosures. Because there are wide variations in the content and format of information released, the Subcommittee does not feel that it is feasible to develop a uniform set of rules, applicable to all agencies, for distinguishing acceptable from unacceptable disclosures.

In formulating disclosure avoidance policies, agencies should give particular attention to the sensitivity of different data items. Financial data such as salaries and wages, benefits, and assets and data on illegal activities and on activities generally considered to be socially sensitive or undesirable require disclosure-avoidance policies that make the risk of statistical disclosure negligible.

Agencies should avoid framing regulations and policies which define unacceptable statistical disclosure in unnecessarily broad or absolute terms. Agencies should apply a test of reasonableness, i.e., releases should be made in such a way that it is reasonably certain that no information about a specific individual will be disclosed in a manner that can harm that individual.

B 2. Special care should be taken to protect individual data when releases are based on complete (as opposed to sample) files and when data are presented for small areas.

B 3. In formulating disclosure-avoidance policies and procedures, agencies should take into account the various kinds of disclosure discussed in Chapters III and IV of this report. Thus, these policies should deal with situations which can lead to unacceptable disclosures, such as:

a. In tabulations
   (1) Empty data cells.
   (2) Cells equal to marginal totals.
   (3) Cells representing a small number of cases.
   (4) Quantity data cells dominated by one or two units.
   (5) Sets of tables from which the above situations can be arrived at by algebraic manipulation.

b. In microdata files
   (1) Files containing data for all members of a defined population.
   (2) Files with detailed geographic information.
   (3) Files with very precise information, such as exact dates of events, or exact amounts of various kinds of income or assets.
   (4) Files containing substantial amounts of information which is likely to be duplicated in external sources containing identifiers.

B 4. With respect to the release of microdata files the Subcommittee believes that

a. There should be no restrictions or conditions attached to the release of microdata files when it is reasonably certain that no information for specific individuals will be disclosed as a result. The Subcommittee has referred to files released under these conditions as public-use files.

b. Where the test for a public-use microdata file is not met, but it appears that the public interest will be served by releasing microdata files for statistical and research purposes on a restricted basis to specific users, such releases should be permitted when all of the following conditions are met.3/

   (1) The receiving organization has authority and obligation to protect the file against mandatory disclosure equivalent to that of the releasing agency.
   (2) Responsible personnel of the receiving agency are subject to meaningful sanctions for violation of confidentiality provisions.
   (3) The receiving organization agrees to:
      (a) Use the file only for statistical and research purposes.
      (b) Not attempt to identify individual data subjects for any purpose.
      (c) Not release the file to anyone else without authorization from the releasing agency.
      (d) Maintain adequate security to protect the file from inadvertent or unauthorized disclosure.
      (e) Apply agreed-on disclosure-avoidance techniques before releasing tabulations based on the file.
      (f) Destroy or return the file within a specified period of time.

B 5. With respect to the release of tabulations, a distinction between unrestricted (public-use) and restricted releases, similar to that described for microdata files in recommendation B 4, would also be appropriate. Thus, for tabulations for which the risk of statistical disclosure is deemed too great to permit release to the general public, restricted releases might be made under conditions similar to those described in paragraph b of recommendation B 4, substituting "tabulations" for "file" wherever the latter word appears.

B 6. To insure compliance with its disclosure-avoidance policies and procedures, each agency that releases statistical information should establish appropriate internal clearance procedures. There should be a clear assignment of individual responsibilities for compliance. Staff members responsible for compliance should be encouraged to become familiar with the materials summarized in this report, and to take advantage of relevant training activities (see recommendation C 2).

B 7. In order to guide their disclosure-avoidance policies, agencies should systematically document the consequences of these policies. In particular they should investigate and record:

a. The details of any cases in which data subjects or others allege that statistical disclosure has occurred.
b. Requests for tabulations and microdata files without identifiers that have been denied or only partially met because of agency disclosure-avoidance policies.

B 8. The Statistical Policy Division, OMB, should encourage agencies that release tabulations and microdata to develop appropriate

policies and guidelines for avoiding disclosure, and to review these policies periodically. To the extent feasible, SPD should help agencies to obtain technical assistance in the development of disclosure-avoidance techniques. SPD should also be prepared to assist and advise agencies in cases where unacceptable disclosures are alleged to have occurred and in cases where potential users, including other Federal agencies, feel that agency disclosure-avoidance policies are unnecessarily restrictive.

C. Disclosure-Avoidance Techniques

Findings

1. In recent years, many different effective techniques for avoiding disclosure have been developed and used. No one technique is ideal for all types of releases.

2. While these techniques have been applied in several instances in the United States and other countries, they are not generally known or accessible to many agency personnel responsible for the release of statistical information. In this report, we have tried to provide a systematic summary description of useful disclosure-avoidance techniques and references to more detailed information.

Recommendations

C 1. This report should be given wide circulation to Federal agencies that release statistical information, whether based on surveys or on program records.

C 2. Based on the material covered in this report, the Statistical Policy Division, OMB, should conduct periodic training seminars for Federal agency personnel who are responsible for developing and applying statistical disclosure-avoidance procedures. These seminars could be organized in much the same way as OMB's recent seminar on presentation of errors in statistical data. Participants would be expected to train and provide technical assistance to appropriate persons in their agencies.

C 3. Disclosure-avoidance procedures should be described, in a general way, in connection with publications or other releases of data to which the procedures have been applied. However, such descriptions should not include details whose publication would tend to reduce the degree of protection provided by the particular procedures used.

C 4. To minimize disclosure risks, agencies that release data based on samples should, where feasible, refrain from publishing information that would make it easier for others to determine which individuals were included in the sample. For example, if a sample is based on ending digits of social security numbers, the particular pattern of ending digits used to select the sample should not be published.

D. Effects of Disclosure on Data Subjects and Users

Findings

1. While we have found some examples of what we consider to be unacceptable statistical disclosures, we have not been able, in spite of a fairly systematic effort, to locate a single instance in which an individual (natural person) alleged that he or she was harmed or might be harmed in any way by statistical disclosure resulting from data released by Federal agencies. The same statement cannot be made for legal persons (corporations, partnerships, etc.) as data subjects. Several companies included in the Federal Trade Commission's Line of Business Surveys have sought legal relief from mandatory response, asserting that publication of tabulations as planned by FTC would result in damaging disclosures of individual company data.

2. There have been a number of cases in which users of data for both natural and legal persons have been unable to obtain the amount of detail desired from tabulations or microdata files because of agency disclosure-avoidance policies. Many such restrictions occur because of limitations on the minimum size (population) of geographic area which may be separately identified. In the case of microdata files, these restrictions, in addition to limiting the availability of data as such, sometimes make it impossible for the user to calculate sampling errors for the statistics of interest when such information is not provided by the releasing agency.

Recommendations

D 1. With respect to agency policies for releases, in statistical form, of information about individuals (natural persons), consideration should be given to the present apparent imbalance where there have been no instances of harm to individuals but several cases where requests for data have been denied. It is recommended that agencies review their policies to determine whether there are ways to respond more fully to user needs without violating statutory requirements or risking harm to individual data subjects. Some agencies may wish to try new data release procedures, such as controlled remote access to restricted microdata files, on a trial or experimental basis, with careful monitoring.

D 2. With respect to data for legal persons (corporations, etc.), both data subjects and data users have expressed some dissatisfaction with current agency disclosure-avoidance policies. The Subcommittee believes that continuing review of these policies is warranted, but it does not have any specific recommendations for change at this time.

E. Needs for Research and Development

Findings

1. Insufficient theoretical or empirical research has been carried out to determine the vulnerability of different classes of data to disclosure or the effects of disclosure-avoidance techniques on the utility of statistical data.

2. The Privacy Protection Study Commission 4/ has recommended, "That the National Academy of Sciences, in conjuction with the relerelevant Federal agencies and scientific and professsional organizations, be asked to develop and promote the use of statistical and procedural techniques to protect the anonymity of an individual who is the subject of any information or record collected or maintained for a research or statistical purpose."

Recommendation

E 1. The Subcommittee would welcome a program of relevant research and development in the area of disclosure-avoidance techniques. Some particular areas that deserve attention are:

a. How disclosure risks in tabulations and microdata are related to varying sampling fractions.

b. How disclosure risks are related to the number of variables in the data base and to their individual and joint distributions.

c. Software systems for providing controlled online access to microdata files.

1/ Membership of the Subcommittee included the three authors together with Richard A. Bell of the Social Security Administration; Tore E. Dalenius, consultant to the Statistical Policy Division; William J. Smith, Jr., of the Internal Revenue Service; Mervyn R. Stuckey of the Statistical Reporting Service, USDA, and Paul T. Zeisset of the Bureau of the Census. Maria Elena Gonzalez of the Statistical Policy Division worked with the Committee in her capacity as chairperson of the Federal Committee on Statistical Methodology. Michael chaired the Subcommittee. Jabine gave oversight to the project on behalf of the Federal Committee on Statistical Methodology.

2/ Except where otherwise specified, the word "individual" as used in this chapter is meant to cover all types of reporting units--natural persons, corporations, partnerships, fiduciaries, etc.

3/ The Subcommittee recognizes that some agencies cannot make this kind of restricted release under existing law.

4/ Privcy Protection Study Commission, Personal Privacy in an Information Society, Washington D.C.: U.S. Government Printing Office, 1977, p. 587.