Introduction

A. Subject and Purposes of This Report

Federal agencies and their contractors who release statistical tables or microdata files are often required by law or established policies to protect the confidentiality of individual information. This confidentiality requirement applies to releases of data to the general public; it can also apply to releases to other agencies or even to other units within the same agency. The required protection is achieved by the application of statistical disclosure limitation procedures whose purpose is to ensure that the risk of disclosing confidential information about identifiable persons, businesses or other units will be very small.

In early 1992 the Statistical Policy Office of the Office of Management and Budget convened an **ad hoc** interagency committee to review and evaluate statistical disclosure limitation methods used by federal statistical agencies and to develop recommendations for their improvement. Subsequently, the **ad hoc** committee became the Subcommittee on Disclosure Limitation Methodology, operating under the auspices of the Federal Committee on Statistical Methodology. This is the final report of the Subcommittee.

The Subcommittee's goals in preparing this report were to:

- update a predecessor subcommittee's report on the same topic (Federal Committee on Statistical Methodology, 1978);

- describe and evaluate existing disclosure limitation methods for tables and microdata files;

- provide recommendations and guidelines for the selection and use of effective disclosure limitation techniques;

- encourage the development, sharing and use of software for the applications of disclosure limitation methods; and

- encourage research to develop improved statistical disclosure limitation methods, especially for public-use microdata files.

The Subcommittee believes that every agency or unit within an agency that releases statistical data should have the ability to select and apply suitable disclosure limitation procedures to all the data it releases. Each agency should have one or more employees with a clear understanding of the methods and the theory that underlies them.

To this end, our report is directed primarily at employees of federal agencies and their contractors who are engaged in the collection and dissemination of statistical data, especially those who are directly responsible for the selection and use of disclosure limitation procedures. We believe that the report will also be of interest to employees with similar responsibilities in other organizations that release statistical data, and to data users, who may find that it helps them to understand and use disclosure-limited data products.

B. Some Definitions

In order to clarify the scope of this report, we define and discuss here some key terms that will be used throughout the report.

B.1. Confidentiality and Disclosure

A definition of **confidentiality** was given by the President's Commission on Federal Statistics (1971:222):

> [Confidential should mean that the dissemination] of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited and that the data are immune from legal process.

The second element of this definition, immunity from mandatory disclosure through legal process, is a legal question and is outside the scope of this report. Our concern is with methods designed to comply with the first element of the definition, in other words, to minimize the risk of **disclosure** (public identification) of the identity of individual units and information about them.

The release of statistical data inevitably reveals some information about individual data subjects. Disclosure occurs when information that is meant to be treated as confidential is revealed. Sometimes disclosure can occur based on the released data alone; sometimes disclosure results from combination of the released data with publicly available information; and sometimes disclosure is possible only through combination of the released data with detailed external data sources that may or may not be available to the general public. At a minimum, each statistical agency must assure that the risk of disclosure from the released data alone is very low.

Several different definitions of disclosure and of different types of disclosure have been proposed (see Duncan and Lambert, 1987 for a review of definitions of disclosure associated with the release of microdata). Duncan et al. (1993: 23-24) provide a definition that distinguishes three types of disclosure:

> **Disclosure** relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (**identity disclosure**), sensitive information about a data subject is revealed through the released file (**attribute disclosure**), or the released data make it possible to

determine the value of some characteristic of an individual more accurately than otherwise would have been possible (**inferential disclosure**).

In the above definition, the word "data" could have been substituted for "file", because each type of disclosure can occur in connection with the release of tables or microdata. The definitions and implications of these three kinds of disclosure are examined in more detail in the next chapter.

B.2. Tables and Microdata

The choice of statistical disclosure limitation methods depends on the nature of the data products whose confidentiality must be protected. Most statistical data are released in the form of tables or microdata files. Tables can be further divided into two categories: tables of frequency (count) data and tables of magnitude data. For either category, data can be presented in the form of numbers, proportions or percents.

A microdata file consists of individual records, each containing values of variables for a single person, business establishment or other unit. Some microdata files include explicit identifiers, like name, address or Social Security number. Removing any such identifiers is an obvious first step in preparing for the release of a file for which the confidentiality of individual information must be protected.

B.3. Restricted Data and Restricted Access

The confidentiality of individual information can be protected by restricting the amount of information in released tables and microdata files (**restricted data**) or by imposing conditions on access to the data products (**restricted access**), or by some combination of these. The disclosure limitation methods described in this report provide confidentiality protection by restricting the data.

Public-use data products are released by statistical agencies to anyone without restrictions on use or other conditions, except for payment of fees to purchase publications or data files in electronic form. Agencies require that the disclosure risks for public-use data products be very low. The application of disclosure limitation methods to meet this requirement sometimes calls for substantial restriction of data content, to the point where the data may no longer be of much value for some purposes. In such circumstances, it may be appropriate to use procedures that allow some users to have access to more detailed data, subject to restrictions on who may have access, at what locations and for what purposes. Such restricted access arrangements normally require written agreements between agency and users, and the latter are subject to penalties for improper disclosure of individual information and other violations of the agreed conditions of use.

The fact that this report deals only with disclosure limitation procedures that restrict data content should not be interpreted to mean that restricted access procedures are of less importance.

Readers interested in the latter can find detailed information in the report of the Panel on Confidentiality and Data Access (see below) and in Jabine (1993b).

C. Report of the Panel on Confidentiality and Data Access

In October 1993, while the Subcommittee was developing this report, the Panel on Confidentiality and Data Access, which was jointly sponsored by the Committee on National Statistics (CNSTAT) of the National Research Council and the Social Science Research Council, released its final report (Duncan et al., 1993). The scope of the CNSTAT report is much broader than this one: disclosure limitation methodology was only one of many topics covered and it was treated in much less detail than it is here. The CNSTAT panel's recommendations on statistical disclosure limitation methods (6.1 to 6.4) are less detailed than the guidelines and recommendations presented in this report. However, we believe that the recommendations in the two reports are entirely consistent with and complement each other. Indeed, the development and publication of this report is directly responsive to the CNSTAT Panel's Recommendation 6.1, which says, in part, that "The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies."

D. Organization of the Report

Chapter II, "Statistical Disclosure Limitation Methods: A Primer", provides a simple description and examples of disclosure limitation techniques that are commonly used to limit the risk of disclosure in releasing tables and microdata. Readers already familiar with the basics of disclosure limitation methods may want to skip over this chapter.

Chapter III describes disclosure limitation methods used by twelve major federal statistical agencies and programs. Among the factors that explain variations in agencies' practices are differences in types of data and respondents, different legal requirements and policies for confidentiality protection, different technical personnel and different historical approaches to confidentiality issues.

Chapter IV provides a systematic and detailed description and evaluation of statistical disclosure limitation methods for tables of frequency and magnitude data. Chapter V fulfills the same function for microdata. These chapters will be of greatest interest to readers who have direct responsibility for the application of disclosure limitation methods or are doing research to evaluate and improve existing methods or develop new ones. Readers with more general interests may want to skip these chapters and proceed to Chapters VI and VII.

Due in part to the stimulus provided by our predecessor subcommittee's report (which we will identify in this report as Working Paper 2), improved methods of disclosure limitation have been developed and used by some agencies over the past 15 years. Based on its review of these methods, the Subcommittee has developed guidelines for good practice for all agencies. With separate sections for tables and microdata, Chapter VI presents guidelines for recommended practices.

Chapter VII presents an agenda for research on disclosure limitation methods. Because statistical disclosure limitation procedures for tabular data are more fully developed than those for microdata, the research agenda focuses more on the latter. The Subcommittee believed that a high priority should be assigned to research on how the quality and usefulness of data are affected by the application of disclosure limitation procedures.

Two appendices are also included. Appendix A contains technical notes on practices the statistical agencies have found useful in extending primary suppression rules to other common situations. Appendix B is an annotated bibliography of articles about statistical disclosure limitation published since the publication of Working Paper 2.

E. Underlying Themes of the Report

Five principal themes underlie the guidelines in Chapter VI and the research agenda in Chapter VII:

- There are legitimate differences between the disclosure limitation requirements of different agencies. Nevertheless, agencies should move as far as possible toward the use of a small number of standardized disclosure limitation methods whose effectiveness has been demonstrated.

- Statistical disclosure limitation methods have been developed and implemented by individual agencies over the past 25 years. The time has come to make the best technology available to the entire federal statistical system. The Subcommittee believes that methods which have been shown to provide adequate protection against disclosure should be documented clearly in simple formats. The documentation and the corresponding software should then be shared among federal agencies.

- Disclosure-limited products should be auditable to determine whether or not they meet the intended objectives of the procedure that was applied. For example, for some kinds of tabular data, linear programming software can be used to perform disclosure audits.

- Several agencies have formed review panels to ensure that appropriate disclosure limitation policies and practices are in place and being properly used. Each agency should centralize its oversight and review of the application of disclosure limitation methods.

- New research should focus on disclosure limitation methods for microdata and on how the methods used affect the usefulness and ease of use of data products.