

Session 2

DISCLOSURE LIMITATION METHODOLOGY

RESTRICTED DATA VERSUS RESTRICTED ACCESS:
A PERSPECTIVE FROM
PRIVATE LIVES AND PUBLIC POLICIES

George T. Duncan
H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213
Phone/FAX: (412) 268-2172/7036
email: George.Duncan@cmu.edu

1994 July 8

A paper presented to the Council of Professional Associations on Federal Statistics Seminar on New Directions in Statistical Methodology, Bethesda, MD, 1994 May 25-26. This paper draws directly on Duncan, G., Jabine, T., and de Wolf, V. (eds.) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, the final report of the Panel on Confidentiality and Data Access of the National Research Council and the Social Science Research Council. Thanks are extended to the panel members for their many contributions to the report. Special thanks go to Thomas Jabine and Virginia de Wolf for both their contributions to the report and for thoughts on this paper. It is dedicated to the memory of Roger Herriot, who in his work in the federal statistical system demonstrated so clearly the value of innovative thinking.

RESTRICTED DATA VERSUS RESTRICTED ACCESS: A PERSPECTIVE FROM
"PRIVATE LIVES AND PUBLIC POLICIES"

George T. Duncan¹
Carnegie Mellon University

1. Stewardship of Statistical Agencies.

A statistical agency is more an art museum than a confessional booth. Certainly the three institutions are similar in eliciting valuables under pledges of protective stewardship—indeed both the survey respondent and the penitent entrust their personal information. But more consequentially, the statistical agency shares only with the art museum a commitment to responsible dissemination to the legitimately curious. Alike, the statistical agency and the art museum must address the tension between protection and access.

Long before statistical agencies had ever sponsored a survey to obtain personal facts, the cloak of confidentiality had been extended in a religious setting. In 1215, the Lateran IV Council decreed that "all the faithful, of both sexes, when they have reached the age of discretion, are to confess all their sins at least once a year to their own priest." (Bok 1983: 78) Traditionally, the received confession is treated as protected personal information, with the priest serving as an instrument of God. On the statistical front, it was not until 1890 that U.S. census legislation required census workers to swear under oath not to disclose census data except to their superiors. Likewise, art museums view protection of their treasured works as an essential function. Motivating the extension of protection by all three is a pragmatic footing: without assurances of security, each would be severely hampered in obtaining the largely voluntary contributions they require.

How does each institution protect its data? The priest silent to the curious is honorable. Contrarily, the art museum hidden to the inquisitive is ineffectual. Likewise, the statistical agency in secreting its data fails its mission.

¹This paper draws directly on Duncan, G., Jabine, T., and de Wolf, V. (eds.) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, the final report of the Panel on Confidentiality and Data Access of the National Research Council and the Social Science Research Council. Thanks are extended to the panel members for their many contributions to the report. Special thanks go to Thomas Jabine and Virginia de Wolf for both their contributions to the report and for thoughts on this paper. It is dedicated to the memory of Roger Herriot, who in his work in the federal statistical system demonstrated so clearly the value of innovative thinking.

Restricted Data versus Restricted Access

Whether for museums or statistical agencies, the dual role of protection and dissemination is challenging, but these two pillars cannot be compromised without risking institutional collapse. Original microdata as collected from statistical surveys can no more be provided to all who might want it than the new Andy Warhol museum in Pittsburgh could freely hand over one of his renderings of Marilyn Monroe.

Generically, two dissemination strategies are possible: provide the good in restricted form, i.e., as a transformation, to a quite general audience without preconditions on use, or provide access to the good itself, but only to a restricted audience under restricted conditions. For art museums, the first strategy calls for providing reproductions, while the second strategy calls for guarded galleries. For a statistical agency, the first strategy results in dissemination of restricted data. The second strategy results in restricted access. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics (1993), the report of the National Research Council/Social Science Research Council Panel on Confidentiality and Data Access, explores these two strategies in its Chapter 6. The purpose of this paper is to provide some perspective on the ideas and recommendations of the report on these topics of restricted data and restricted access.

2. Restricted data

Restricted data is a confidentiality-motivated transformation of the original data; it results from the application of a statistical disclosure limitation technique. Before releasing a microdata file, for example, a statistical agency might go beyond removing explicit identifiers like name, address, and Social Security number. To limit disclosure risk, the agency could, for example, give people's ages in five-year intervals rather than by the exact date of birth.

Private Lives and Public Policies gives an overview of some key concepts and techniques of disclosure limitation:

- Disclosure risk, including identity, attribute, and inferential disclosure
- Statistical procedures for disclosure limitation, both for microdata and for tabular data
- Impact of improved computer and communications technology
- Recent research on disclosure limitation

A review and evaluation of statistical disclosure limitation techniques and their

Restricted Data versus Restricted Access

application is given in the Report on Statistical Disclosure Limitation Methodology (1994) and in Dalenius (1988) (also see, Fienberg 1993), so the treatment here will not be detailed.

Disclosure risk

As explored in Duncan and Lambert (1989), 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 released data makes it possible to infer the value of an attribute of a data subject more accurately than otherwise would have been possible (inferential disclosure).

Statistical procedures for disclosure limitation

Statistical disclosure limitation techniques involve transformations of data to limit the risk of disclosure. Use of such a technique is often called masking the data, because it is intended to hide characteristics of data subjects. Some statistical disclosure limitation techniques are designed for data accessed as tables (tabular data), some are designed for data accessed as records of individual data subjects (microdata), and some are designed for data accessed as computer databases. Common methods of masking tabular data are deleting table entries (cell suppression) and altering table entries (random error, or noise introduction). Common methods of masking microdata are deleting identifiers, dropping sensitive variables, releasing only a small fraction of the data records, and grouping data values into categories. Direct access of computer databases raises new statistical disclosure limitation issues which are only recently being addressed (see, e.g., Duncan and Mukherjee 1992; Keller-McNulty and Unger 1993).

In the case of a public-use microdata file, statistical disclosure limitation techniques can be classified into five broad categories (Duncan and Pearson, 1991):

1. Collecting or releasing only a sample of the data: For example, the Bureau of the Census first released a public-use microdata file with a 1-in-1000 sample from the 1960 Census of Population and Housing.
2. Including simulated data: This technique has not been implemented, but it is conceptually akin to including several identical limousines in a motorcade that is under threat of terrorist attack.

Restricted Data versus Restricted Access

3. "Blurring" of the data by grouping or adding error to the individual values: Presenting subjects' ages in 10-year intervals is an example of grouping. A microdata file prepared by the Census Bureau for the National Opinion Research Center from the 1980 census masked census tract characteristics (e.g., percentage of blacks, unemployment rate) by adding random noise (Kim 1990).
4. Excluding certain attributes: Information on a doctoral graduate field of specialization might be omitted.
5. Swapping of data by exchanging the values of certain variables between data subjects: The value of some sensitive variable could be exchanged for that in, say, an adjacent record.

For data released as tables, the blurring and swapping techniques described above have been used. Three other statistical disclosure limitation techniques are unique to tables (Cox 1980):

1. Requiring each marginal total of the table to have a minimum count of data subjects.^a
2. Using a "concentration" rule, also known as the (N, K)-rule, where N entities do not dominate K percent of a cell; for example, requiring that the reported aspects of two dominant businesses in a cell comprise no more than a certain percentage of a cell.
3. Using controlled rounding of table entries to perturb entries while maintaining various marginal totals.

Statistical disclosure limitation practices of federal statistical agencies

The practices of federal statistical agencies regarding statistical disclosure limitation is well-covered in Jabine (1993b), a paper commissioned by the Panel on Confidentiality and Data Access. Based on a detailed study of twelve statistical agencies, their basic finding is that, although most have standards, guidelines, or formal review mechanisms, there is great diversity in policies, procedures, and practices among them.

This finding provides the basis for the Panel's first recommendation in this area (all eight recommendations are given for convenience in the Appendix):

Recommendation 6.1. The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work

Restricted Data versus Restricted Access

on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies. Major statistical agencies should actively encourage and participate in scholarly statistical research in this area. Other agencies should keep abreast of current developments in the application of statistical disclosure limitation techniques.

Beginnings have been made in implementing this recommendation. In early 1992 the Statistical Policy Office convened an ad hoc interagency committee of ten persons to be chaired by Nancy Kirkendall of the U. S. Energy Information Administration. The mandate of the committee was 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. Its final product, the Report on Statistical Disclosure Limitation Methodology, notes, "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.'" In the report's Chapter VII, a research agenda is laid out for disclosure limitation methodology. A reasonable expectation is that further progress on dissemination will be made by the dissemination of the Subcommittee's report, the presentations at the Council of Professional Associations on Federal Statistics (COPAFS) Seminar on New Directions in Statistical Methodology, and publications in the OMB Statistical Policy Working Paper series.

The Panel was concerned with the impact of statistical disclosure limitation procedures on the quality of the data as it is disseminated to data users. Statistical disclosure methods can hide or distort relations among study variables and result in analyses that are incomplete or misleading. Because of this possibility, policy researchers have expressed serious reservations about the implementation of statistical disclosure limitation (e.g., Smith 1991). Further, data masked by some disclosure limitation methods can only be analyzed accurately by researchers who are highly sophisticated methodologically. Based on these findings, the panel made the following recommendation:

Recommendation 6.2. Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers. If the data cannot be

Restricted Data versus Restricted Access

accurately analyzed using standard statistical software, the agency should make appropriate consulting and software available.

Unfortunately, this recommendation has yet to be addressed, or to appear on the research agenda of statistical agencies. The Report on Statistical Disclosure Limitation Methodology is moot on this topic.

Given the potential difficulties that certain statistical disclosure limitation techniques can cause for analysts, it is important that federal statistical agencies involve data users in selecting such procedures. As Greenberg (1991:375) notes, "survey sponsors and data users must contribute to the decision making process in identifying areas in which some completeness and/or accuracy can be sacrificed while attempting to maintain as much data quality as possible." These thoughts led to the Panel's third recommendation:

Recommendation 6.3. Each statistical agency should actively involve data users from outside the agency as statistical disclosure limitation techniques are developed and applied to data.

Steps toward implementation of this recommendation are being made through the inclusion of individuals outside the agency on microdata review panels. It remains to be seen whether the views of data users will be adequately represented.

Finally, over the past thirty years various agencies have released public-use microdata files successfully. Based on experience, such data dissemination has met a two-pronged test: (1) the microdata files have been useful to researchers and policy analysts and (2) confidentiality has been protected. Based on this finding, the panel made a final recommendation in this area:

Recommendation 6.4. Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.

Given an increased public concern over privacy and confidentiality issues,

Recommendation 6.4 presents a real challenge to statistical agencies. Far easier it would be to turn inward and protective. To do so, however, would be to abdicate the statistical agency's responsibility to provide the data a democratic society needs.

The panel noted that expansion of the number and richness of public-use microdata files to be disseminated would be better justified if all users were

Restricted Data versus Restricted Access

subject to sanctions for disclosure of information about individually identifiable data subjects. Reference was made to a recommendation, in another chapter, as follows:

Recommendation 5.3 There should be legal sanctions for all users, both external users and agency employees, who violate requirements to maintain the confidentiality of data.

3. Restricted access: Administrative procedures to protect confidentiality

Procedures for providing restricted access to data typically establish eligibility requirements for access and impose a variety of conditions governing the purposes for which the data can be used, which organizations and individuals can have access, the location of access, physical security measures, and the retention and disposition of initial and secondary data files.

Arrangements for providing restricted access to federal data for statistical purposes do exist. Jabine (1993a) provides 19 examples, including both interagency data sharing and arrangements with data users external to the federal government.

Interagency data sharing

There have been instances of agreements to permit interagency sharing of identifiable, or potentially identifiable, personal records for statistical purposes. Some of the instances involved transfers of administrative records; others involve transfers of data collected in statistical surveys. As identified in Private Lives and Public Policies, the mechanisms used to insure confidentiality in a selected set of instances included the following:

- Making data users in the receiving agency special sworn employees of the sharing agency
- Restricting further dissemination of data and follow up with respondents
- Periodic on-site inspections of the receiving agency's security measures by the sharing agency
- Regular review of the benefits of the sharing arrangement
- Written agreement that a specified data match would not be used for any other purpose and that the receiving agency would return the shared data

Restricted Data versus Restricted Access

file when the match was completed

- Minimizing the possibility of using linked data to identify an individual in a public-use file and then using the survey information in the identified individual's record for administrative purposes by data masking

In general, an obvious requirement for interagency data sharing is that the statutory requirements for confidentiality of all of the agencies involved must be observed. A second requirement is that the transfer of data among agencies must be consistent with statements made to data providers when the data were obtained from them.

Developing arrangements for interagency data sharing can be a complex and time-consuming process, especially if more than two agencies are involved or if novel applications of the data are planned. New initiatives are likely to pose new legal, ethical, administrative, and policy questions. The expected benefits in cost savings or better quality data must be substantial to justify the level of effort and perseverance needed to find acceptable answers. It helps if the proposed data-sharing arrangements offer benefits to all of the parties concerned.

The success of the instances examined in efficiently using data resources while protecting confidentiality support the panel's first recommendation regarding restricted access.

Recommendation 6.5. Federal statistical agencies should strive for a greater return on public investment in statistical programs through carefully controlled increases in interagency data sharing for statistical purposes and expanded availability of federal data sets to external users.

Full realization of this goal will require legislative changes, as discussed in Chapter 5 of Private Lives and Public Policies, but much can be accomplished within the framework of existing legislation.

External data users

The availability of high-speed computers and sophisticated analytic techniques and software have generated vastly increased appetites for federal statistical data. In many cases if the data are restricted sufficiently to ensure confidentiality, the released data will not satisfy the needs of users. Appropriate to such cases, several modes of restricted access for external data users have

Restricted Data versus Restricted Access

been developed by statistical agencies. Some of the important features of these access modes are eligibility criteria, location of access, cost and convenience for agencies and users, and methods of protecting confidentiality. Particular modes of restricted access include the following:

- Use of a fellows program with access at the agency's central facility, for a limited term, and only for projects that the host agency deems to be of interest
- Remote access to computer databases with automated screening of batch process programs
- Use of encrypted CD-ROM products which have statistical software that is restricted so as to prevent the user from obtaining unencrypted individual records or statistics that would tend to disclose individual information.
- Release of microdata under licensing agreements that provide for special sworn employee status, authorize unscheduled site visits to the data user, provide for prepublication review by the disseminating agency, and require return or destruction of the data when the research is completed.
- Ease on-site access of data users by providing access at agency regional centers.

Given this history and the value to society of broad dissemination of federal statistical data, the panel made the following two recommendations:

Recommendation 6.6. Statistical agencies, in their efforts to expand access for external data users, should follow a policy of responsible innovation. Whenever feasible, they should experiment with some of the newer restricted access techniques, with appropriate confidentiality safeguards and periodic reviews of the costs and benefits of each procedure.

Recommendation 6.7. In those instances in which controlled access at agency sites remains the only feasible alternative, statistical agencies should do all they can to make access conditions more affordable and acceptable to users, for example, by providing access at dispersed agency locations and providing adequate user support and access to computing facilities at reasonable cost.

Restricted Data versus Restricted Access

Finally the panel supported archiving of important statistical data:

Recommendation 6.8. Significant statistical data files, in their unrestricted form, should be deposited at the National Archives and eventually made available for historical research uses.

This recommendation is intended to cover statistical databases from censuses and surveys and those, like the Statistics of Income and the Continuous Work History Sample databases, that are derived from administrative records. The panel was purposely not specific as to the content of such archived databases and the length of time for which confidentiality restrictions should continue to apply. Some databases, like the economic and population censuses, might include explicit identification of data providers. Others, especially those based on samples, might not include names and addresses, but would not be subject to statistical disclosure limitation procedures of the kind that are applied to public-use microdata sets for contemporary use.

4. Conclusions

There is an inverse relationship between restrictions on data and restrictions on access: as data restrictions increase, fewer restrictions on access are needed and vice versa. A given level of confidentiality can be achieved with various combinations of restricted data and restricted access. Just as an art museum may sell reproductions, provide carefully monitored access to galleries, and allow qualified art historians considerable latitude in examination of a work, a statistical agency must choose an appropriate mix of data products to disseminate that will serve the needs of their various data users. A strong beginning has been made by the federal statistical system in developing a research and implementation agenda for restricted data. This is evident from the important contribution of the Report on Statistical Disclosure Limitation. I ponder the contribution that might be made through a comparable effort in developing a research and implementation agenda for restricted access. No less, I ponder the restricted data and restricted access procedures that will be required to ensure data access with confidentiality in the computer databases of the Global Information Infrastructure.

Restricted Data versus Restricted Access

REFERENCES

Bok, S. (1983) Secrets: On the Ethics of Concealment and Revelation New York: Random House.

Dalenius, T. (1988) Controlling Invasion of Privacy in Surveys Department of Development and Research, Statistics Sweden.

Duncan, G. T., Jabine, T., and de Wolf, V. (1993) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics Washington, DC: National Academy Press.

Duncan, G. T. and Lambert, D. (1986) The risk of disclosure for microdata. Journal of Business and Economic Statistics 7(2):207-217.

Duncan, G. T. and Mukherjee, S. (1992) Confidentiality protection in statistical databases: a disclosure limitation approach. Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute.

Fienberg, S. (1993) Conflicts between the needs for access to statistical information and demands for confidentiality. Technical Report #577, Department of Statistics, Carnegie Mellon University.

Jabine, T. (1993a) Procedures for restricted data access. Journal of Official Statistics 9(2):537-589.

Jabine, T. (1993b) Statistical disclosure limitation practices of United States statistical agencies. Journal of Official Statistics 9(2):427-454.

Keller-McNulty, S. and Unger, E. (1993) Database systems: inferential security. Journal of Official Statistics, 9(2)475-499.

Kim, J. (1990) Masking Microdata for National Opinion Research Center. Final Project Report. Bureau of the Census, Washington, D.C.

Report on Statistical Disclosure Limitation Methodology (1994) Statistical Policy Office, Office of Management and Budget, Washington, DC.

Smith, J. P. (1991) Data confidentiality: a researcher's perspective. American Statistical Association 1991 Proceedings of the Social Statistics Section. Alexandria, VA: American Statistical Association.

Restricted Data versus Restricted Access

APPENDIX. Recommendations

Recommendation 6.1. 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. Major statistical agencies should actively encourage and participate in scholarly statistical research in this area. Other agencies should keep abreast of current developments in the application of statistical disclosure limitation techniques.

Recommendation 6.2. Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers. If the data cannot be accurately analyzed using standard statistical software, the agency should make appropriate consulting and software available.

Recommendation 6.3. Each statistical agency should actively involve data users from outside the agency as statistical disclosure limitation techniques are developed and applied to data.

Recommendation 6.4. Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.

Recommendation 6.5. Federal statistical agencies should strive for a greater return on public investment in statistical programs through carefully controlled increases in interagency data sharing for statistical purposes and expanded availability of federal data sets to external users.

Recommendation 6.6. Statistical agencies, in their efforts to expand access for external data users, should follow a policy of responsible innovation. Whenever feasible, they should experiment with some of the newer restricted access techniques, with appropriate confidentiality safeguards and periodic reviews of the costs and benefits of each procedure.

Recommendation 6.7. In those instances in which controlled access at agency sites remains the only feasible alternative, statistical agencies should do all they can to make access conditions more affordable and acceptable to users, for example, by providing access at dispersed agency locations and providing adequate user support and access to computing facilities at reasonable cost.

Restricted Data versus Restricted Access

Recommendation 6.8. Significant statistical data files, in their unrestricted form, should be deposited at the National Archives and eventually made available for historical research uses.

Statistical Disclosure Limitation Methodology

by

Nancy J. Kirkendall, Energy Information Administration

Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology was released in May 1994. This working paper reflects the efforts of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology. I was the chair of the Subcommittee. The other members are William Arends, National Agricultural Statistics Service; Lawrence Cox, Environmental Protection Agency; Virginia de Wolf, Bureau of Labor Statistics; Arnold Gilbert, Bureau of Economic Analysis; Thomas Jabine, Committee on National Statistics; Mel Kollander, Environmental Protection Agency; Donald Marks, Department of Defense; Barry Nussbaum, Environmental Protection Agency; and Laura Zayatz, Bureau of the Census.

Working Paper 22 presents a basic introduction to statistical disclosure limitation, describes the methods used by 12 Federal Statistical Agencies, provides more detail on techniques used to protect tables and microdata, and discusses needed research. It also presents the Subcommittee's recommendations. The previous *Statistical Policy Working Paper* on the subject of disclosure limitation was *Statistical Policy Working Paper 2*, which was published in 1978. While *Working Paper 22* is an update of *Working Paper 2* in some sense, one of our primary purposes was to summarize and describe the current techniques which are used to protect data, and to make recommendations concerning what the subcommittee felt should be done. It is primarily intended to serve as a practitioner's handbook.

The purpose of this paper is to summarize the information and the recommendations made in *Working Paper 22*.

Disclosure Limitation

"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."¹

The historical method of providing data to the public is via tables. Beginning in 1962 with the advent of the computer age, agencies also started releasing microdata files. In a microdata file, each record contains a set of variables that pertain to a single respondent. The variables relate to that respondent's reported values. However, there are no identifiers on the file, and the data may be disguised in some way to make sure they do not reveal the respondent's identity.

¹*Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, p. 1.

For our purposes there are two types of disclosure. **Identity disclosure**, occurs when a specific respondent can be identified from the data. Identity disclosure is particularly important to microdata files, and the solution is to limit or modify the identifying information on the file. **Attribute disclosure** occurs when confidential information about the respondent is revealed. This type of disclosure is particularly important to tables (where it is assumed that one might know if a person is represented in the table), and the solution is to make sure a sufficient number of respondents contribute to each cell in the table.

A distinction is also made between tables of **frequency data** and tables of **magnitude data**. A simple example illustrates the difference. Assume that a survey provides data on a person's profession, his salary, and the county in which he lives. Let us assume that in Franklin county, we had the following three respondents who reported that they were doctors.

**Example Cell in Profession x County table
{Doctors, Franklin county}.**

Number	Count	Salary
1	1	\$600,000
2	1	\$ 75,000
3	1	\$ 75,000
Total	3	\$750,000

With this example, if we publish the total for counts (3), we say we have count data. If we publish the percent of people surveyed who were doctors, we say we have frequency data. With frequency or count data every respondent contributes exactly the same amount to the cell, and methods of identifying sensitive cells depend only on the number of respondents contributing to a cell.

On the other hand, the salaries are called magnitude data. Here the respondent's contribution to the cell total depends on his reported value. Let us assume that the two doctors who are less well paid are local general practitioners, and the third is a heart surgeon who works in the city, but lives in Franklin County. Publishing the total salary would allow each of the local doctors to make a very good estimate for the salary of the heart surgeon. If they can estimate his salary "too closely", we would say that we have attribute disclosure. Thus, for tables of magnitude data, the method of determining sensitive cells depends on the values reported by each respondent.

In the next few sections of this paper, we will illustrate the methods used to protect data and present the Subcommittee's recommendations. Section 1 concerns tables of frequency or count data; Section 2 tables of magnitude data; and Section 3 microdata. Section 4 is a summary.

1.0 Tables of Frequency (Count) Data

A cell in a table of frequencies or counts is sensitive if there are too few respondents. The methods used to protect such cells include:

1. Collapse categories (combine rows or columns).
2. Suppression.
3. Controlled (random) rounding.
4. Confidentiality edit.

Both collapsing categories and suppression are widely used by Federal agencies, and have been for years. Random rounding and controlled rounding have not actually been used by Federal agencies. The confidentiality edit is a new method which was used to protect tables from the 1990 decennial Census.

Assume that cells are defined to be sensitive if they have three or fewer respondents. The following table is an example we will use to illustrate different ways of protecting the sensitive cells. The cells which are sensitive are printed in bold with an asterisk.

Table 1 -- Example -- with Disclosure

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	1*	3*	1*	20
B	20	10	10	15	55
C	3*	10	10	2*	25
D	12	14	7	2*	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

1.1 Combine categories

As noted above, one way of protecting the sensitive cells is to combine rows and/or columns. In the following table, the education levels are combined into two categories. Clearly, the result is that there are no sensitive cells. However, a lot of information is lost.

**Table 2 -- Example Without Disclosure
Protection Provided by Combining Rows or Columns**

Household Head Education Level			
County	Low/Medium	High/Very High	Total
A	16	4	20
B	30	25	55
C	13	12	25
D	26	9	35
Total	85	50	135

1.2 Suppression

The second method of providing protection is to simply withhold from publication the sensitive cells and a combination of other cells in each row and column so that it is not possible to derive the value of the sensitive cells by subtraction using the published marginal totals. Clearly, we need at least two suppressed cells in every row and column, but is that enough? The answer is no, and here is the counter example.

**Table 3 -- Example With Disclosure
Protection Not Provided By Suppression**

Household Head Education Level					
County	Low	Med	High	Very High	Total
A	15	S ₁	S ₂	S ₃	20
B	20	S ₄	S ₅	15	55
C	S ₆	10	10	S ₇	25
S	S ₈	14	7	S ₉	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

To show that this table still contains disclosures, consider the sum of row 1 and row 2 minus the sum of column 2 and column 3. This reduces to the following equation:

$$(15 + S_1 + S_2 + S_3) + (20 + S_4 + S_5 + 15) - (S_1 + S_4 + 10 + 7) - (S_2 + S_5 + 10 + 7) = 20 + 55 - 35 - 30$$

or

$$S_3 = 1$$

This illustrates that selection of cells for complementary suppression is not a trivial matter. Methods of linear programming are used to select the set of cells which are "optimal" in some sense and which protect the sensitive cells. The following table with suppressions does protect the sensitive cells.

Table 4 -- Example Without Disclosure Protection Provided by Suppression

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	S	S	S	20
B	20	10	10	15	55
C	S	S	10	S	25
D	S	14	S	S	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

This example leads to the first of our recommendations. When suppression is used to protect tabular data, whether frequency or magnitude data, the table with suppressions should be *audited*. Auditing involves applying a linear programming algorithm to calculate the largest value a suppressed cell can take and the smallest value it can take. If the largest value and the smallest value are equal, the cell total is revealed exactly. If they are "too close" then the cell value can be estimated "too closely".

1.3 Random Rounding or Controlled Rounding

With random or controlled rounding, each cell count is rounded using some base value. In the following example, the base value is 5. In this case each cell count can be written as $X = 5q + r$. For random rounding each cell is rounded at random. This cell would be rounded up with probability $r/5$, and down with probability $1-r/5$. The problem with this procedure is that tables do not add, as illustrated in the Table 5.

**Table 5 -- Example Without Disclosure
Protection Provided by Random Rounding**

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	0	0	0	20
B	20	10	10	15	55
C	5	10	10	0	25
D	15	15	10	0	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

Random rounding has been used by Statistics Canada and was used by the New Zealand Department of Statistics before they moved to controlled rounding. The New Zealand Department of Statistics moved to controlled rounding primarily because users complained that the randomly rounded tables did not add (George and Penny, 1987.)

Controlled rounding is like random rounding except that a linear programming method is used to impose the constraint that the table must add. Controlled rounding was a topic of research during the 1980's, and for two dimensional tables and most three dimensional tables current methods work very well. It was proposed for use with the 1990 decennial census (Greenberg, 1986), but has not yet been used by any Federal statistical agency. An example of our table protected with controlled rounding is presented below.

**Table 6 -- Example Without Disclosure
Protection Provided by Random Rounding**

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	0	5	0	20
B	20	10	10	15	55
C	5	10	10	0	25
D	10	15	5	5	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

1.4 Confidentiality Edit

All of the above methods are applied to a specific table. If the table is changed in some way, or another table containing data from the same data file is constructed, another detailed analysis must follow to assure that consistent protection is applied.

The confidentiality edit is a new method which was developed at the U. S. Census Bureau and used to protect tables from the 1990 Census (Griffin, Navarro, and Flores-Baez, 1989). With this method the original microdata file is manipulated (much as it would be if it were going to be released for public use). After manipulation the microdata file can be used directly to make tables. Other tables made from the same manipulated microdata file will also be protected, and the protection will be consistent. The approach described below was used for the regular decennial Census data file (the 100 percent data file), it uses a microdata protection technique called "data swapping" or "switching" (Dalenius and Reiss, 1982).

To apply the confidentiality edit the following steps are applied.

1. Take a random sample of records from the microdata file;
2. Find a match with them in some other county, based on a set of key variables;
3. Swap all other variables on the matched records;
4. Make tables

After the confidentiality edit, our table might appear as below.

**Table 7 -- Example Without Disclosure
Protection Provided by Confidentiality Edit**

Household Head Education Level

County	Low	Med	High	Very High	Total
A	13	2	4	2	20
B	18	12	8	17	55
C	5	9	11	0	25
D	14	12	8	1	35
Total	50	35	30	20	135

The only disadvantage I have seen quoted is that the table does not look as if disclosure limitation has been applied.

1.5 Recommendation

While each of these methods has advantages and disadvantages, the Subcommittee was unable to determine which of these methods were preferable in terms of the level of protection applied, and the usefulness of the result. Our recommendation is that further research should be done to address this question, and the result widely disseminated.

2.0 Tables of Magnitude Data

For tables of magnitude data only two methods can be used to protect sensitive cells. They are combining categories, and suppression. Each has the same strengths and weaknesses as discussed above, and if suppression is used the table should be audited. For tables of magnitude data, the new question is how to identify sensitive cells?

We indicated above that the respondents' reported values are used. In fact, cells are identified as sensitive if a simple linear combination of respondent level data is positive. The linear equation is called a linear sensitivity rule and the coefficients depend on the specific rule used and the parameters chosen. There are three rules which are commonly used:

(n,k) rule -- a cell is sensitive if n respondents contribute k% or more to the cell total;

p-percent -- a cell is sensitive if the published total can be used to estimate any respondent's data more accurately than p-percent;

pq -- like the p-percent rule, but acknowledges that before data are published, common knowledge allows estimation of any respondents' data to within q percent ($q > p$).

Recommendations

The Subcommittee's recommendations for tables of magnitude data are:

1. Only subadditive linear sensitivity measures should be used to identify sensitive cells. Subadditivity is a mathematical property that assures that if two or more cells are not sensitive, then their sum (union) is not sensitive either. Fortunately, all three commonly used linear sensitivity rules are subadditive.
2. The committee prefers the p-percent or pq rules as providing more consistent protection.
3. Suppression or collapsing categories are the only accepted methods of protecting sensitive cells.
4. The parameter values used in practice should not be revealed.
5. Tables containing suppressions should be audited.

For tables of magnitude data research is needed into identifying summary statistics to publish as a replacement for a sensitive cell total. If it could be shown that the summary statistics do not reveal individual data, they could be used instead of suppression and provide users with more information.

3.0 Microdata

For tables, we have associated "disclosure" with the publication of "sensitive cells", and have justified a simple way to identify which cells are sensitive. Once that is done, several approaches have been used to protect the sensitive cells. Unfortunately, for microdata files there is no standard agreed to definition of what constitutes "disclosure", other than uniquely identifying an individual in a data file.

The following four common ways to protect microdata files are used by virtually every agency which releases microdata files.

1. Use only a sample of the population. (A sample protects an individual's data, because it is not generally known whether or not a particular individual is included in the file.)
2. Remove obvious identifiers (eg. name, address, social security number).
3. Limit geographic detail (detailed data about an individual from too small a geographic region increases the risk of identification.)
4. Top code, bottom code and/or recode continuous high visibility variables. (Recoding continuous variables essentially makes them discrete. The larger values are shown only as greater than some number, the smaller values are shown as less than some number, and the intermediate values are assigned to a range.)

Salary is an example of a high visibility continuous variable. It may take many different values, and either very large ones, very small ones, or very precise recording of the value may reveal a respondent's identity. (Like our highly paid heart surgeon.) Other ways of protecting microdata are also applied to high visibility continuous variables. They include:

5. Masking (add or multiply by random numbers);
6. Swapping or rank swapping (find two records which match on a selected set of variables and exchange (swap) the remaining variables);
7. Blank and impute for randomly selected records. (randomly select a set of records, eliminate specific reported variables and replace them by imputed values);
8. Blurring -- aggregate values across small groups of respondents. (find a group of respondents, average some of their variables, and replace the reported values by the average.)

Recommendations

The subcommittee could only make one fairly obvious recommendation for protecting microdata files.

Remove direct identifiers and limit other identifying information.

Research is needed into defining disclosure or an unacceptable likelihood of disclosure for microdata files. Another area of needed research is into the impact of disclosure limitation techniques on the usefulness of the resultant data file. The subcommittee believes that research into these topics was of the highest priority.

4.0 General Recommendations and Summary

In addition to the specific recommendations above, the subcommittee had the following general recommendations. Agencies should

1. Seek advice from respondents and data users. Respondents should be asked about variables they consider sensitive and those they do not consider sensitive. It would be better if agencies applied disclosure limitation methods only to variables considered sensitive by respondents. Data users should be offered the opportunity to comment on disclosure limitation methods. Agencies should use this information in selecting the disclosure limitation methods to use.
2. Centralize review of disclosure limited products within an agency. A centralized review of disclosure limited products assures consistency in the application of disclosure limitation within an agency. In addition, a centralized review provides greater assurance that the data are adequately protected.
3. Share software and methodology. Agencies need to help each other to assure consistency in practice, and to make more advanced methodology and software widely available.
4. Agencies which release the same or similar data sets should cooperate in the application of disclosure limitation to those data sets. If there is no coordination, it is more likely, for example, that cells selected for complementary suppression by one agency, might not be suppressed by the other agency. This would lead to disclosure.

This paper has provided an elementary description of statistical disclosure limitation methodology and the principle recommendations of the Subcommittee on Statistical Disclosure Limitation Methodology. *Working Paper 22* provides considerably more detail on statistical disclosure limitation methodology, agency practices and needed research. It also provides an extensive annotated bibliography. The Subcommittee hopes that you find the information useful.

References

- Cox, L. H., Johnson, B. McDonald, S. Nelson, D. and Vazquez, V. (1985) "Confidentiality Issues at the Census Bureau," Proceedings of the Bureau of the Census First Annual Research Conference, Bureau of the Census, Washington D.C. pp 199-218.
- Dalenius, T. and Reiss, S. P. (1982). "Data Swapping: A Technique for Disclosure Control". Journal of Statistical Planning and Inference, Vol 6, pp. 73-85.
- George, J. A. and Penny, R. N. (1987), "Initial Experience in Implementing Controlled Rounding for Confidentiality Control, Proceeding of the Bureau of Census Third Annual Research Conference, Bureau of the Census, Washington DC., pp 253-262.
- Greenberg, B. (1986), "Designing a Disclosure Avoidance Methodology for the 1990 Decennial Census," presented at the 1990 Census Data products Fall Conference, Arlington, VA.
- Griffin, R. A., Navarro, A. and Flores-Baez, L. (1989), "Disclosure Avoidance for the 1990 Census, Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 516-521.

Discussion of Presentations on Statistical Disclosure Methodology¹

Stephen E. Fienberg²

1 Prologue

This past weekend my wife and I were attending a Bat Mitzvah and the daughter of our friends read her portion of the Torah from the *Book of Numbers* dealing with the census of the Israelites in the desert. As I listened to her, I read this passage from the bible again with special care with the hope of some divine inspiration for my discussion of the two papers presented today. Let me share with you what I learned about disclosure limitation.

First, the census seemed to be much easier to take than we have found to be the case in modern times in the United States. There is no mention of an undercount, differential or otherwise, although women and children were intentionally omitted from the count. It turns out that there were 603,550 Israelites aged 20 and above, and the bible gives various breakdowns of these totals, without any reference to or apparent concern for confidentiality.

¹Presented at "Seminar on New Directions in Statistical Methodology," sponsored by the Council of Professional Associations on Federal Statistics, Bethesda, MD, May 25-26, 1994

²Stephen E. Fienberg is Maurice Falk Professor of Statistics and Social Science at Carnegie Mellon University, Pittsburgh, PA, 15213. The preparation of this discussion was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to York University.

Second, the 12 tribes were organized around the tabernacle and in that sense we could think of the tribes as corresponding to geographic areas. Part of the reported data goes down to subgroups whose order of magnitude is a few thousand. Clearly, this would not meet the Census Bureau requirement for the release of identification of geographic codes for microdata sets where the threshold is now 100,000.

Third, while the bible contains no cross-tabulations as we know them today, it does include considerable information that could be displayed in cross-classified form. But even the smallest numbers reported, e.g., the 273 for the number of first born of the Levites, would not seem to provide an example requiring cell suppression.

Fourth, the bible actually releases the names of several individuals who participated in the census, especially the names of a number of the tribal leaders and their sons. This suggests that the Israelites didn't have any hang-up about the issue of uniques in the population for the release of census data. The idea seems to be that there is the need to distinguish whether or not the release is in fact harmful. After all, everyone knew that Moses, Aaron, and a number of others were included in the census and what their demographic classifications were. Therefore, identifying them by name did not compromise them in any way.

Finally, there were few or no subsequent releases from the biblical census

so we don't have much evidence about how the Israelites would have treated concerns about confidentiality. We do know, however, that there was nothing corresponding to Title 13 in the Torah or in the commentaries such as the Talmud.

Having noted all of this in the way of prefatory remarks, let me now turn to the two papers in this session.

2 Duncan on *Private Lives and Public Policies*

George Duncan has summarized the major recommendations from Chapter 6 of *Private Lives and Public Policies*, a report issued by the NRC-SSRC panel he chaired as they relate to statistical procedures to protect confidentiality. His paper begins with a discussion of a 1215 Latern IV Council decree on confidentiality and quickly shifts to statistical agencies' dual role of protector and disseminator of data. He then takes up the panel's themes of restricted data (via some transformation) versus restricted access. To do this, he needs to define disclosure and, in keeping with the literature, discusses this at three levels: individual disclosure, attribute disclosure, and inferential disclosure, and he lists some standard techniques for providing restricted data to achieve disclosure avoidance. This material is a brief introduction to that which is covered in much greater detail in chapter II of the draft Federal Committee on Statistical Methodology Working Paper 22, *Statistical Disclosure Limitation Methodology*, described by Nancy Kirkendall. In my remarks I will focus

on the panel's recommendations regarding restricted data and those aspects of the topic dealt with in Working Paper 22.

Because of the great diversity in policies and practices of the statistical agencies (documented in the panel report and in chapter III of Working Paper 22) the panel recommended that OMB should continue to coordinate research work on disclosure limitation and disseminate the results widely. The existence of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology and its recently released working paper represent OMB's and the agencies' positive response. The panel's second recommendation relates to agency assessments of the impact of their own data disclosure limitation techniques and Working Paper 22 remains silent on the matter, a point to which I will return in a few moments.

A few years ago I argued that the statistical agencies in the U. S. clearly were using techniques that were too conservative, i.e., that they erred too much on the side of restricting data in order to ensure that guarantees of confidentiality are not compromised as opposed to increasing the extent and utility of released data. I was immediately challenged and I offered as evidence to support my proposition the total absence of anecdotes where, despite agency actions, confidentiality had been breached. Agencies must remember that they are only public protectors and not owners of the data and they need to involve users in the choice of disclosure avoidance procedures. This is the third of the panel's recommendations and this, according to Duncan, is in the process of implementation by a number of agencies. The panel's final

recommendation encouraged the continued widespread release of microdata sets.

I have watched the NRC/SSRC panel from conception through the completion of its report. While the four recommendations I have singled out here from Chapter 6 of the report sound much like apple pie and motherhood, they and the other recommendations of the panel are clearly designed to move the practice of statistical data disclosure forward and encourage the development of a statistical basis for confidentiality practices. I heartily recommend the report and its companion volume of technical commissioned papers which appeared as a special issue of the *Journal of Official Statistics* in the fall of 1993.

3 Kirkendall on *Statistical Disclosure Limitation Methodology*

Nancy Kirkendall has described some of the ideas and materials from Working Paper 22 of the Federal Committee on Statistical Methodology Subcommittee on Disclosure Limitation Methodology, an activity which she chaired. This working paper needs to be considered against the backdrop of an earlier Federal Committee on Statistical Methodology working paper on the topic issued in 1978. What we have here is a major update with considerable detail and an extensive annotated bibliography. Depending on how we approach the topic, we find both good news and bad.

First, the good news. Much has happened in the intervening 16 years. The earlier working paper was technically innovative and it served as a catalyst to the development of new disclosure limitation methodology, especially in such agencies as the Bureau of the Census, but also by those in universities such as George Duncan and my former colleague Diane Lambert, and by Tore Dalenius, my fellow discussant today. The new working paper documents many of these advances and the extent of the research developed is impressive. So too are the advances in the uses of disclosure limitation methodology by federal statistical agencies. The current agency practices, as described in chapter III of Working Paper 22, are far more advanced thanks both to the methodological developments and to attendant advances in computation. In these senses, the new working paper represents a major progress report on the health of the federal statistical system.

Next, the bad news. I found the new working paper disappointing, largely because it represents an intellectual backsliding from the innovative stance staked out by its predecessor and because of its failure to adopt what I would argue is a badly needed statistical foundation for the very methods whose cause it advances. Let me explain.

Chapter II of the report captures some the current discussions in the literature about the the definitions of disclosure, but it fails to build on Dalenius' statistical definition of disclosure that formed the foundation for the structure of the 1978 paper. As a consequence, we have descriptions of methodology

such as cell suppression which, while seemingly advanced, represent mathematics but not statistics. The techniques have been honed so that they can be implemented for large collections of cross-classifications utilizing linear programming and other techniques but we are never told, either by those who developed the approach or by the Subcommittee preparing this working paper, what statistical criteria the methods attempt to optimize and the extent to which they succeed. Thus we are told, for example, about the need to keep the values of n and p in the cell suppression rules confidential, but there is no recognition that statistical learning by those outside the agency might easily make such a statement essentially moot. Similarly, in the discussion of three-way and multiway cross-classifications, there is no recognition of relevant statistical methodology that might inform the very methods under discussion such as the probabilistic theory for Fréchet bounds on cell values (e.g., see Kwerel, 1983). When we get to the discussion of research issues relating to cell suppression, we find more of the same: advances in optimization of network flow methods, more elaborate computer programs, faster software. Where is the statistics in statistical disclosure limitation methodology? Where is the recognition that the data collected by statistical agencies is not error free? I contend that this very measurement error ultimately drives the statistical properties of attempts to compromise otherwise confidential data and disclosure limitation methodology to counter such attempts.

Nancy Kirkendall presented an example of an application of cell suppression which produces through complementary suppressions the following table (in which S stands for a suppressed cell and x a released cell):

S_1	x	x	S_2	S_3
S_4	x	x	S_5	x
x	S_6	x	x	S_7
x	S_8	x	x	S_9

She uses this to illustrate the need for auditing tables prior to release since the cell with entry S_3 can be determined via the other cells. It is interesting to note that all of this is related to the theory of existence of maximum likelihood estimates under quasi-independence for two-way tables. (e.g. see Chapter 5 of Bishop, Fienberg, and Holland, 1975). That those developing methods in this area seem unaware of such links to the statistical literature serves to reinforce my point on the need to make statistical disclosure methods more statistical.

I have a similar reaction to the briefer materials described in the Working Paper on data swapping, especially as it was implemented in the 1990 decennial census. This method grew out of a novel notion suggested by Tore Dalenius, but there appears to be little recognition by those who implemented the approach regarding the effect that the method has had on the utility of the resulting data, for example, as it is to be used for enforcement of the

Voting Rights Act. I understand that considerable effort went into some of these considerations in advance, but we have little documentation and no post-censal evaluations.

The Working Paper also places what I believe to be a misguided emphasis on "population uniques." As various authors have noted, uniqueness in the population is a necessary but not sufficient condition for identity disclosure, and there is no reason to believe that identity disclosure necessarily compromises confidentiality guarantees. My example of the identity release in the biblical census I believe makes this point well. The Working Paper relegates the more interesting and more important statistical problems of inferential disclosure and measuring disclosure risk to the research agenda.

Finally, the report tries to make a clear demarcation between methods for microdata and methods for tabulations. What it fails to recognize is that many examples of tabulations are in fact restricted microdata. For example, tables of counts are microdata in which either the original variables are categorical or are continuous but have been disguised through the use of conversion through categories, and where the data have been truncated by the dropping of variables. Surely there should be some linkage between the methods for microdata and for tabulations. This is less a criticism of Working Paper 22 than it is of the state of the art of research on disclosure limitation. (See the related remarks in Fienberg, 1994).

There are interesting statistical ideas and proposals for a unified theory

of disclosure control in the research literature, such as those captured by the papers by Fuller, Lambert, Little, and Rubin in the recent special issue of JOS on privacy and confidentiality, but these are not given appropriate coverage in the Working Paper nor are they reflected in agency thinking. Perhaps this simply reflects the lag between research and practical implementation. Despite such shortcomings, Working Paper 22 is an excellent summary both of current methods and practices in the agencies. The Subcommittee should be applauded for its efforts.

4 Restricted Access or Expanded Access?

George Duncan's second major topic was the NRC/SSRC panel's recommendations on administrative procedures to protect confidentiality. The panel has emphasized the role of interagency data sharing as well as technological aids that facilitate such access. While the need for such restricted access clearly will continue, I believe that the future will be one of expanded rather than restricted access. Working Paper 22 is especially helpful in this regard. Chapter V on "Methods for Public-use Microdata Files" provides a concise primer on the developments in this area.

I've mentioned the role of technology in restricted access, but technology is even more important when we come to expand access. A number of federal statistical agencies are playing leadership roles in this regard. Nancy Kirkendall referred to the innovative approach being explored by the National

Center for Educational Statistics, but there are many other examples. For example, micro-data from the 1990 decennial census are currently available over the Internet via the Consortium for International Earth Science Information Network (CIESIN) in Michigan. Further, the Bureau of the Census has created SIPP-On-Call, a new interactive approach to allow access to files from the Survey of Income and Program Participation over the Internet. Special user-friendly access is available via Gopher or NSF's Mosaic. Even *Wired* magazine, in its June 1994 issue, describes such access to its readers and points out that one also has on-line access to the Privacy Act and Title 13 as hypertext documents!

The new world of immediate user and intruder access over the "information highway" will place greater demands on released microdata and it will test, in new ways, the appropriateness of disclosure limitation methods both for the preservation of confidentiality and for the increased utility of the released data. This, I predict, will be a major topic for the next Federal Committee Subcommittee effort in this area and I expect that new statistical approaches to disclosure limitation will accompany these emerging changes.

5 Summary

There is much meat for statistical thought in *Private Lives and Public Policies*, the report of the NRC/SSRC panel, and in both the original Working Paper No. 2 and the recently released Working Paper No. 22, *Statistical Disclosure Limitation Methodology*, produced under the sponsorship of the

Federal Committee on Statistical Methodology. I fully expect that the next COPAFS-sponsored seminar on new statistical methodology, will highlight new advances in this area that build on the substantial contributions to date, that will also better link to statistical ideas, and that will report on the enhanced utility of released data resulting from these new developments.

6 References

- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. N. (1975). *Discrete Multivariate Analysis: Theory and Practice* MIT Press: Cambridge, MA.
- Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10, in press.
- Kwerel, S. M. (1983). Fréchet bounds. In *Encyclopedia of Statistical Sciences*. Vol. 3, (S. Kotz and N. L. Johnson, eds.), Wiley: New York, 202-209.

Tore Dalenius
DISCUSSION

Introduction

Before around 1970, the main direction of the methodological development was on the development of survey designs enhancing the efficiency, i.e. increasing the amount of information provided by a survey by other means than increasing the size of the survey.

Around 1970, a decisive change may be observed. The attention of the survey statisticians was now gradually directed towards how to recognize and hopefully address the problem of invasion of privacy. To address that problem, it proved necessary to apply methods which in fact served to reduce the amount of information made available. The subject of this meeting – Disclosure Limitation Methodology – reflects the above-mentioned change of methodological direction.

Dr. Duncan's presentation is based on ch. 6 of the book "Private Lives and Public Policies". This chapter considers two main options for protection of the confidentiality of released data: providing 'restricted data' and providing 'restricted access'. Dr. Kirkendall's presentation is based on the report Statistical Disclosure Limitation Methodology, prepared by a subcommittee of the Federal Committee on Statistical Methodology. This report is limited to disclosure limitation by means of 'restricted data'. Obviously, both documents are *final* products, a fact of relevance for the shaping of my discussion.

In what follows, I will first discuss selected aspects of restricted data and restricted access, respectively, to be followed by brief accounts of some additional aspects.

SELECTED ASPECTS OF RESTRICTED DATA

1. Two Classes of Data

Dr. Duncan and Dr. Kirkendall discuss in some detail two classes of data, viz. tabular data (frequency data and magnitude data), and microdata.

2. Frequency Data

The data to be restricted are represented by a table $T(N)$ with $R \times C$ cells. The restriction is achieved by a two-step procedure:

- i. the sensitive cells of $T(N)$, if any, are identified by subjecting the table to a threshold rule: cells with a small number of data subjects (such as $n = 3$) are considered sensitive;
- ii. next, some cells are combined, suppressed or rounded.

3. Magnitude Data

Typically these data are non-demographic, such as income or sales, accounted for by a table $T(X)$ with $R \times C$ cells. The variable X has in most cases a skew distribution: a small number of data subjects may account for a large proportion of the cell values. These cells may accordingly be sensitive, i.e. make it possible to link the cells with the data subjects accounted for, that is, to identify the data subjects. Hence, some kind of a restriction has to be applied to these cells.

The restriction of the data is achieved by:

- i. first identifying cells to which a small number of data subjects contribute a large percentage of the cell value - this may be done by using the p percent rule, the pq rule, or the (n, k) rule, also called the 'dominance rule';
- ii. next, these cells are subjected to restrictions, such as top-coding.

4. Microdata

Most releases of microdata are made up by a set of records with data about *individuals*. Only in exceptional cases do the data refer to business establishments.

Before the records can be released, formal identifiers must be removed ('deidentified'). But it may still be possible to link a record with a data subject: unique combinations of data concerning some attributes may serve as 'quasi-identifiers'. Hence additional restrictions are necessary, such as:

- i. sampling;
- ii. excluding data for one or more variables;
- iii. representing the data by broad classes; age may for example be represented by an interval (age class);
- iv. releasing data only for large populations; and
- v. confidentiality edit of the data,

to give but five examples.

SELECTED ASPECTS OF RESTRICTED ACCESS

5. A Wide Class of Procedures

Dr. Duncan includes in this class several disclosure limitation approaches. Common to them is that the statistical agency establishes eligibility requirements for the data users who are to be included in the group of users given access. I will briefly consider four procedures.

6. Interagency Data Sharing

This term is used to denote two cases:

- i. transfer of *administrative* data from a government agency to the statistical agency; and
- ii. transfer of *statistical* data from a government agency to the statistical agency.

7. Swearing In of Users

Formally, this kind of restricted access means that potential users are given status of employees of the statistical agency concerned, either at the main office, or at some local office near the place where the potential users live.

Clearly, the statistical office will have an opportunity of critically assessing the users' research projects and also the merits of the users.

8. Site Inspection

Assume that there is a government agency with authority to inspect how a statistical agency performs with respect to protection of the confidentiality of the data to be released. Then this "control agency" may implement a scheme for inspection of the performance of the statistical agency.

The scheme may call for inspection every k th month. A better scheme would, however, call for inspection at dates chosen at random. This would make it impossible for the statistical agency to perform well during an inspection but not between inspections.

SELECTED MISSING TOPICS

9. The Coverage of the Two Presentations

It goes without saying that it is possible to identify topics which have not been presented, or possibly only touched upon. I will provide three such examples.

10. Example No. 1 – Schemes for Rounding

Rounding the counts in a table may be carried out in several ways. The main ways are related to:

- i. the choice of a base different from the standard $b = 5$;
- ii. the simultaneous use of more than one base, especially if the table is large (many rows and columns);
- iii. rounding all cells in the table rather than a subset of cells; this type of scheme has in fact been proposed for use in the British population census; and
- iv. the use of deterministic rather than random rounding.

11. Example No. 2 – The Multi-Table Problem

Let T_1 be a table with no disclosure. And let T_2 be another similar table. Release of both T_1 and T_2 is not necessarily safe. Access to both tables may make it possible to derive a combined table T_3 which is disclosing.

12. Example No. 3 – Release by a Database

The statistical agencies should develop schemes for releasing statistics by means of a database. There is no reason to 'wait and see' what comes out with respect to a data superhighway.

TOPICS FOR RESEARCH AND DEVELOPMENT

13. Terminology

There is as yet no generally agreed upon terminology in the area under consideration here. It suffices to mention the following facts:

- i. privacy is defined in a great many different ways;

- ii. confidentiality is sometimes viewed as 'anonymity'; and
- iii. what in the two presentations is called 'disclosure limitation' was called 'disclosure avoidance' in the 1978 report; an alternative term is 'disclosure control', which I prefer.

It is indeed high time to develop a standard terminology.

14. A Catalogue of Potential Research Topics

In the report from the subcommittee there are some suggestions about research topics. But additional topics are needed. I will suggest one topic, viz. design of microdata about business establishments.

15. Inventory and Analysis of Sensitive Topics

In the last two decades, the non-response rate in surveys has shown a tendency to grow, possibly reflecting an increasing unwillingness to answer questions about sensitive topics.

In my view, the survey statisticians should process surveys already carried out and generate an *inventory* of sensitive topics which may explain the development. Such an inventory would be useful in the design of future surveys, by drawing the statisticians' attention to the need for special measures (such as special measurement methods) to improve the rate of cooperation.

The inventory should be *analyzed* to identify groups of data subjects with very large non-response rates. Such groups may then be singled out for special action.

CONCLUDING REMARKS

By way of presenting a summary of my views about the two presentations, I want to say that I have found them very informative and helpful. Dr. Duncan and Dr. Kirkendall are to be congratulated to the contributions they and their cooperators have made.