Statistical Disclosure Limitation: A Primer

This chapter provides a basic introduction to the disclosure limitation techniques which are used to protect statistical tables and microdata. It uses simple examples to illustrate the techniques. Readers who are already familiar with the methodology of statistical disclosure limitation may prefer to skip directly to Chapter III, which describes agency practices, Chapter IV which provides a more mathematical discussion of disclosure limitation techniques used to protect tables, or Chapter V which provides a more detailed discussion of disclosure limitation techniques applied to microdata.

A. Background

One of the functions of a federal statistical agency is to collect individually identifiable data, process them and provide statistical summaries to the public. Some of the data collected are considered proprietary by respondents. Agencies are authorized or required to protect individually identifiable data by a variety of statutes, regulations or policies. Cecil (1993) summarizes the laws that apply to all agencies and describes the statutes that apply specifically to the Census Bureau, the National Center for Education Statistics, and the National Center for Health Statistics. Regardless of the basis used to protect confidentiality, federal statistical agencies must balance two objectives: to provide useful statistical information to data users, and to assure that the responses of individuals are protected.

Not all data collected and published by the government are subject to disclosure limitation techniques. Some data on businesses collected for regulatory purposes are considered public. Some data are not considered sensitive and are not collected under a pledge of confidentiality. The statistical disclosure limitation techniques described in this paper are applied whenever confidentiality is required and data or estimates are to be publicly available. Methods of protecting data by restricting access are alternatives to statistical disclosure limitation. They are not discussed in this paper. See Jabine (1993) for a discussion of restricted access methods. All disclosure limitation methods result in some loss of information, and sometimes the publicly available data may not be adequate for certain statistical studies. However, the intention is to provide as much data as possible, without revealing individually identifiable data.

The historical method of providing data to the public is via statistical tables. With the advent of the computer age in the early 1960's agencies also started releasing **microdata files**. In a microdata file each record contains a set of variables that pertain to a single respondent and are related 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 that individual data items cannot be uniquely associated with a particular respondent. A new method of releasing data has been introduced by the National Center for Education Statistics (NCES) in the 1990's. Data are provided on diskette or CD-ROM in a secure data base system with access programs which allow

users to create special tabulations. The NCES disclosure limitation and data accuracy standards are automatically applied to the requested tables before they are displayed to the user.

This chapter provides a simple description of the disclosure limitation techniques which are commonly used to limit the possibility of disclosing identifying information about respondents in tables and microdata. The techniques are illustrated with examples. The tables or microdata produced using these methods are usually made available to the public with no further restrictions. Section B presents some of the basic definitions used in the sections and chapters that follow: included are a discussion of the distinction between tables of frequency data and tables of magnitude data, a definition of table dimensionality, and a summary of different types of disclosure. Section C discusses the disclosure limitation methods applied to tables of counts or frequencies. Section D addresses tables of magnitude data, section E discusses microdata, and Section F summarizes the chapter.

B. Definitions

Each entry in a statistical table represents the aggregate value of a quantity over all units of analysis belonging to a unique statistical cell. For example, a table that presents counts of individuals by 5-year age category and the total annual income in increments of $10,000 is comprised of statistical cells such as the cell {35-39 years of age, $40,000 to $49,999 annual income}. A table that displays value of construction work done during a particular period in the state of Maryland by county and by 4-digit Standard Industrial Code (SIC) groups is comprised of cells such as the cell {SIC 1521, Prince George's County}.

B.1. Tables of Magnitude Data Versus Tables of Frequency Data

The selection of a statistical disclosure limitation technique for data presented in tables (**tabular data**) depends on whether the data represent frequencies or magnitudes. Tables of **frequency count data** present the number of units of analysis in a cell. Equivalently the data may be presented as a percent by dividing the count by the total number presented in the table (or the total in a row or column) and multiplying by 100. Tables of **magnitude data** present the aggregate of a "quantity of interest" over all units of analysis in the cell. Equivalently the data may be presented as an average by dividing the aggregate by the number of units in the cell.

To distinguish formally between **frequency count data** and **magnitude data**, the "quantity of interest" must measure something other than membership in the cell. Thus, tables of the number of establishments within the manufacturing sector by SIC group and by county-within-state are frequency count tables, whereas tables presenting total value of shipments for the same cells are tables of magnitude data. For practical purposes, entirely rigorous definitions are not necessary. The statistical disclosure limitation techniques used for magnitude data can be used for frequency data. However, for tables of frequency data other options are also available.

B.2. Table Dimensionality

If the values presented in the cells of a statistical table are aggregates over two variables, the table is a **two-dimensional** table. Both examples of detail cells presented above, {35-39 years of age, $40,000-$49,999 annual income} and {SIC 1521, Prince George's County} are from two-dimensional tables. Typically, categories of one variable are given in columns and categories of the other variable are given in rows.

If the values presented in the cells of a statistical table are aggregates over three variables, the table is a **three-dimensional** table. If the data in the first example above were also presented by county in the state of Maryland, the result might be a detail cell such as {35-39 years of age, $40,000-$49,999 annual income, Montgomery County}. For the second example if the data were also presented by year, the result might be a detail cell such as {SIC 1521, Prince George's County, 1990}. The first two-dimensions are said to be presented in rows and columns, the third variable in "layers".

B.3. What is Disclosure?

The definition of disclosure given in Chapter I, and discussed further below is very broad. Because this report documents the methodology used to limit disclosure, the focus is on practical situations. Hence, the concern is only with the disclosure of confidential information through the public release of data products.

As stated in Lambert (1993), "disclosure is a difficult topic. People even disagree about what constitutes a disclosure." In Chapter I, the three types of disclosure presented in Duncan, et. al (1993) were briefly introduced. These are identity disclosure, attribute disclosure and inferential disclosure.

Identity disclosure occurs if a third party can identify a subject or respondent from the released data. Revealing that an individual is a respondent or subject of a data collection may or may not violate confidentiality requirements. For tabulations, revealing identity is generally not disclosure, unless the identification leads to divulging confidential information (attribute disclosure) about those who are identified.

For microdata, identification is generally regarded as disclosure, because microdata records are usually so detailed that the likelihood of identification without revealing additional information is minuscule. Hence disclosure limitation methods applied to microdata files limit or modify information that might be used to identify specific respondents or data subjects.

Attribute disclosure occurs when confidential information about a data subject is revealed and can be attributed to the subject. Attribute disclosure may occur when confidential information is revealed exactly or when it can be closely estimated. Thus, attribute disclosure comprises identification of the subject and divulging confidential information pertaining to the subject.

Attribute disclosure is the form of disclosure of primary concern to statistical agencies releasing tabular data. Disclosure limitation methods applied to tables assure that respondent data are published only as part of an aggregate with a sufficient number of other respondents to prevent attribute disclosure.

The third type of disclosure, **inferential disclosure**, occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of home. As purchase price of home is typically public information, a third party might use this information to infer the income of a data subject. In general, statistical agencies are not concerned with inferential disclosure, for two reasons. First a major purpose of statistical data is to enable users to infer and understand relationships between variables. If statistical agencies equated disclosure with inference, no data could be released. Second, inferences are designed to predict aggregate behavior, not individual attributes, and thus are often poor predictors of individual data values.

Table 1: Example Without Disclosure

Number of Households by Heated Floorspace and Family Income
(Million U.S. Households)

	1990 Family income							
Heated Floor Space sq ft	Total	Less than $5000	$5000 to $9999	$10000 to $14999	$15000 to $24999	$25000 to $34999	$35000 to $49999	$50000 or more
Fewer than 600	8.0	1.5	1.9	1.6	1.5	.8	.5	.3
600 to 999	22.5	2.0	3.7	4.1	5.5	3.4	2.7	1.2
1000 to 1599	26.5	1.1	3.2	3.2	5.2	5.1	5.5	3.3
1600 to 1999	12.6	.3	1.0	1.1	2.2	2.3	2.6	3.1
2000 to 2399	9.0	Q	.5	.6	1.3	1.3	2.3	2.8
2400 to 2999	7.8	.2	.3	.5	1.0	1.4	1.7	2.7
3000 or more	7.4	Q	.2	.3	.7	1.0	1.3	3.8

NOTE: Q -- Data withheld because relative standard error exceeds 50%.
SOURCE: "Housing Characteristics 1990", Residential Energy Consumption Survey, Energy Information Administration, DOE/EIA-0314(90), page 54.

C. Tables of Counts or Frequencies

The data collected from most surveys about people are published in tables that show counts (number of people by category) or frequencies (fraction or percent of people by category). A portion of a table published from a sample survey of households that collects information on energy consumption is shown in Table 1 on the previous page as an example.

C.1. Sampling as a Statistical Disclosure Limitation Method

One method of protecting the confidentiality of data is to conduct a sample survey rather than a census. Disclosure limitation techniques are not applied in Table 1 even though respondents are given a pledge of confidentiality because it is a large scale **sample** survey. Estimates are made by multiplying an individual respondent's data by a sampling weight before they are aggregated. If sampling weights are not published, this weighting helps to make an individual respondent's data less identifiable from published totals. Because the weighted numbers represent all households in the United States, the counts in this table are given in units of millions of households. They were derived from a sample survey of less than 7000 households. This illustrates the protection provided to individual respondents by sampling and estimation.

Additionally, many agencies require that estimates must achieve a specified accuracy before they can to be published. In Table 1 cells with a "Q" are withheld because the relative standard error is greater than 50 percent. For a sample survey accuracy requirements such as this one result in more cells being withheld from publication than would a disclosure limitation rule. In Table 1 the values in the cells labeled Q can be derived by subtracting the other cells in the row from the marginal total. The purpose of the Q is not necessarily to withhold the value of the cell from the public, but rather to indicate that any number so derived does not meet the accuracy requirements of the agency.

When tables of counts or frequencies are based directly on data from all units in the population (for example the 100-percent items in the decennial Census) then disclosure limitation procedures must be applied. In the discussion below we identify two classes of disclosure limitation rules for tables of counts or frequencies. The first class consists of special rules designed for specific tables. Such rules differ from agency to agency and from table to table. The special rules are generally designed to provide protection to data considered particularly sensitive by the agency. The second class is more general: a cell is defined to be sensitive if the number of respondents is less than some specified threshold (the threshold rule). Examples of both classes of disclosure limitation techniques are given in Sections II.C.2 and II.C.3.

C.2. Special Rules

Special rules impose restrictions on the level of detail that can be provided in a table. For example, Social Security Administration (SSA) rules prohibit tabulations in which a detail cell is equal to a marginal total or which would allow users to determine an individual's age within a five year interval, earnings within a $1000 interval or benefits within a $50 interval.

Tables 2 and 3 illustrate these rules. They also illustrate the method of restructuring tables and combining categories to limit disclosure in tables.

Table 2: Example -- With Disclosure

Number of Beneficiaries by Monthly Benefit Amount and County

	Monthly Benefit Amount						
County	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Total
A	2	4	18	20	7	1	52
B	--	--	7	9	--	--	16
C	--	6	30	15	4	--	55
D	--	--	2	--	--	--	2

SOURCE: Working Paper 2.

Table 2 is a two-dimensional table showing the number of beneficiaries by county and size of benefit. This table would not be publishable because the data shown for counties B and D violate Social Security's disclosure rules. For county D, there is only one non-empty detail cell, and a beneficiary in this county is known to be receiving benefits between $40 and $59 per month. This violates two rules. First the detail cell is equal to the cell total; and second, this reveals that all beneficiaries in the county receive between $40 and $59 per month in benefits. This interval is less than the required $50 interval. For county B, there are 2 non-empty cells, but the range of possible benefits is from $40 to $79 per month, an interval of less than the required $50.

To protect confidentiality, Table 2 could be restructured and rows or columns combined (sometimes referred to as "rolling-up categories"). Combining the row for county B with the row for county D would still reveal that the range of benefits is $40 to $79. Combining A with B and C with D does offer the required protection, as illustrated in Table 3.

Table 3: Example -- Without Disclosure

Number of Beneficiaries by Monthly Benefit Amount and County

	Monthly Benefit Amount						
County	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Total
A and B	2	4	25	29	7	1	68
C and D	--	6	32	15	4	--	57

SOURCE: Working Paper 2.

C.3. The Threshold Rule

With the threshold rule, a cell in a table of frequencies is defined to be **sensitive** if the number of respondents is less than some specified number. Some agencies require at least 5 respondents in a cell, others require 3. An agency may restructure tables and combine categories (as illustrated above), or use cell suppression, random rounding, controlled rounding or the confidentiality edit. Cell suppression, random rounding, controlled rounding and the confidentiality edit are described and illustrated below.

Table 4 is a fictitious example of a table with disclosures. The fictitious data set consists of information concerning delinquent children. We define a cell with fewer than 5 respondents to be sensitive. Sensitive cells are shown with an asterisk.

C.3.a. Suppression

One of the most commonly used ways of protecting sensitive cells is via **suppression**. It is obvious that in a row or column with a suppressed sensitive cell, at least one additional cell must be suppressed, or the value in the sensitive cell could be calculated exactly by subtraction from the marginal total. For this reason, certain other cells must also be suppressed. These are referred to as **complementary** suppressions. While it is possible to select cells for complementary suppression manually, it is difficult to guarantee that the result provides adequate protection.

Table 4: Example -- With Disclosure

Number of Delinquent Children
by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	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 headings are fictitious.

Table 5 shows an example of a system of suppressed cells for Table 4 which has at least two suppressed cells in each row and column. This table appears to offer protection to the sensitive cells. But does it?

Table 5: Example -- With Disclosure, Not Protected by Suppression

Number of Delinquent Children
by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	D_1	D_2	D_3	20
Beta	20	D_4	D_5	15	55
Gamma	D_6	10	10	D_7	25
Delta	D_8	14	7	D_9	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.
SOURCE: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

The answer is no. Consider the following linear combination of row and column entries: Row 1 (county Alpha) + Row 2 (county Beta) - Column 2 (medium education) - Column 3 (high education), can be written as

$$(15 + D_1 + D_2 + D_3) + (20 + D_4 + D_5 + 15) - (D_1 + D_4 + 10 + 14) - (D_2 + D_5 + 10 + 7) = 20 + 55 - 35 - 30.$$

This reduces to $D_3 = 1$.

This example shows that selection of cells for complementary suppression is more complicated than it would appear at first. Mathematical methods of linear programming are used to automatically select cells for complementary suppression and also to **audit** a proposed suppression pattern (eg. Table 5) to see if it provides the required protection. Chapter IV provides more detail on the mathematical issues of selecting complementary cells and auditing suppression patterns.

Table 6 shows our table with a system of suppressed cells that does provide adequate protection for the sensitive cells. However, Table 6 illustrates one of the problems with suppression. Out of a total of 16 interior cells, only 7 cells are published, while 9 are suppressed.

Table 6: Example -- Without Disclosure, Protected by Suppression

Number of Delinquent Children
by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	D	D	D	20
Beta	20	10	10	15	55
Gamma	D	D	10	D	25
Delta	D	14	D	D	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.
SOURCE: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

C.3.b. Random Rounding

In order to reduce the amount of data loss which occurs with suppression, the U.S. Census Bureau has investigated alternative methods to protect sensitive cells in tables of frequencies. Perturbation methods such as random rounding and controlled rounding are examples of such alternatives. In **random rounding** cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down.

For this example, it is assumed that each cell will be rounded to a multiple of 5. Each cell count, X, can be written in the form

$$X = 5q + r,$$

where q is a nonnegative integer, and r is the remainder (which may take one of 5 values: 0, 1, 2, 3, 4). This count would be rounded up to 5*(q+1) with probability r/5; and would be rounded down to 5*q with probability (1-r/5). A possible result is illustrated in Table 7.

Because rounding is done separately for each cell in a table, the rows and columns do not necessarily add to the published row and column totals. In Table 7 the total for the first row is 20, but the sum of the values in the interior cells in the first row is 15. A table prepared using random rounding could lead the public to lose confidence in the numbers: at a minimum it looks as if the agency cannot add. The New Zealand Department of Statistics has used random rounding in its publications and this is one of the criticisms it has heard (George and Penny, 1987).

Table 7: Example -- Without Disclosure, Protected by Random Rounding

Number of Delinquent Children
by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	0	0	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	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 headings are fictitious.

C.3.c. Controlled Rounding

To solve the additivity problem, a procedure called **controlled rounding** was developed. It is a form of random rounding, but it is constrained to have the sum of the published entries in each row and column equal the appropriate published marginal totals. Linear programming methods are used to identify a controlled rounding for a table. There was considerable research into controlled rounding in the late 1970's and early 1980's and controlled rounding was proposed for use with data from the 1990 Census, (Greenberg, 1986). However, to date it has not been used by any federal statistical agency. Table 8 illustrates controlled rounding.

One disadvantage of controlled rounding is that it requires the use of specialized computer programs. At present these programs are not widely available. Another disadvantage is that controlled rounding solutions may not always exist for complex tables. These issues are discussed further in Chapters IV and VI.

C.3.d. Confidentiality Edit

The **confidentiality edit** is a new procedure developed by the U.S. Census Bureau to provide protection in data tables prepared from the 1990 Census (Griffin, Navarro, and Flores-Baez, 1989). There are two different approaches: one was used for the regular decennial Census data (the 100 percent data file); the other was used for the long-form of the Census which was filed by a sample of the population (the sample data file). Both techniques apply statistical disclosure limitation techniques to the microdata files before they are used to prepare tables. The adjusted files themselves are not released, they are used only to prepare tables.

Table 8: Example -- Without Disclosure, Protected by Controlled Rounding

Number of Delinquent Children
by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	0	5	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	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 headings are fictitious.

First, for the 100 percent microdata file, the confidentiality edit involves "data swapping" or "switching" (Dalenius and Reiss, 1982; Navarro, Flores-Baez, and Thompson, 1988). The confidentiality edit proceeds as follows. First, take a sample of records from the microdata file. Second, find a match for these records in some other geographic region, matching on a specified set of important attributes. Third, swap all attributes on the matched records. For small blocks, the Census Bureau increases the sampling fraction to provide additional protection. After the microdata file has been treated in this way it can be used directly to prepare tables and no further disclosure analysis is needed.

Second, the sample data file already consists of data from only a sample of the population, and as noted previously, sampling provides confidentiality protection. Studies showed that this protection was sufficient except in small geographic regions. To provide additional protection in small geographic regions, one household was randomly selected and a sample of its data fields were blanked. These fields were replaced by imputed values. After the microdata file has been treated in this way it is used directly to prepare tables and no further disclosure analysis is needed.

To illustrate the confidentiality edit as applied to the 100 percent microdata file we use fictitious records for the 20 individuals in county Alpha who contributed to Tables 4 through 8. Table 9 shows 5 variables for these individuals. Recall that the previous tables showed counts of individuals by county and education level of head of household. The purpose of the confidentiality edit is to provide disclosure protection to tables of frequency data. However, to achieve this, adjustments are made to the microdata file before the tables are created. The following steps are taken to apply the confidentiality edit.

Table 9: Fictitious Microdata

All Records in County Alpha Shown
Delinquent Children

Number	Child	County	HH education	HH income	Race
1	John	Alpha	Very high	201	B
2	Jim	Alpha	High	103	W
3	Sue	Alpha	High	77	B
4	Pete	Alpha	High	61	W
5	Ramesh	Alpha	Medium	72	W
6	Dante	Alpha	Low	103	W
7	Virgil	Alpha	Low	91	B
8	Wanda	Alpha	Low	84	W
9	Stan	Alpha	Low	75	W
10	Irmi	Alpha	Low	62	B
11	Renee	Alpha	Low	58	W
12	Virginia	Alpha	Low	56	B
13	Mary	Alpha	Low	54	B
14	Kim	Alpha	Low	52	W
15	Tom	Alpha	Low	55	B
16	Ken	Alpha	Low	48	W
17	Mike	Alpha	Low	48	W
18	Joe	Alpha	Low	41	B
19	Jeff	Alpha	Low	44	B
20	Nancy	Alpha	Low	37	W

NOTES: HH indicates head of household. Income given in thousands of dollars.

1. Take a sample of records from the microdata file (say a 10% sample). Assume that records number 4 and 17 were selected as part of our 10% sample.

2. Since we need tables by county and education level, we find a match in some other county on the other variables race, sex and income. (As a result of matching on race, sex and income, county totals for these variables will be unchanged by the swapping.) A match for record 4 (Pete) is found in County Beta. The match is with Alfonso whose head of household has a very high education. Record 17 (Mike) is matched with George in county Delta, whose head of household has a medium education.

   In addition, part of the randomly selected 10% sample from other counties match records in county A. One record from county Delta (June with high education) matches with Virginia, record number 12. One record from county Gamma (Heather with low education) matched with Nancy, in record 20.

3. After all matches are made, swap attributes on matched records. The adjusted microdata file after these attributes are swapped appears in Table 10.

Table 10: Fictitious Microdata

Delinquent Children -- After Swapping
Only County Alpha Shown

Number	Child	County	HH education	HH income	Race
1	John	Alpha	Very high	201	B
2	Jim	Alpha	High	103	W
3	Sue	Alpha	High	75	B
4*	**Alfonso**	**Alpha**	**Very high**	**61**	**W**
5	Ramesh	Alpha	Medium	72	W
6	Dante	Alpha	Low	103	W
7	Virgil	Alpha	Low	91	B
8	Wanda	Alpha	Low	84	W
9	Stan	Alpha	Low	75	W
10	Irmi	Alpha	Low	62	B
11	Renee	Alpha	Low	58	W
12*	**June**	**Alpha**	**High**	**56**	**B**
13	Mary	Alpha	Low	54	B
14	Kim	Alpha	Low	52	W
15	Tom	Alpha	Low	55	B
16	Ken	Alpha	Low	48	W
17*	**George**	**Alpha**	**Medium**	**48**	**W**
18	Joe	Alpha	Low	41	B
19	Jeff	Alpha	Low	44	B
20*	**Heather**	**Alpha**	**Low**	**37**	**W**

* Data: first name and education level swapped in fictitious microdata file from another county.
NOTES: HH indicates head of household. Income given in thousands of dollars.

4. Use the swapped data file directly to produce tables, see Table 11.

The confidentiality edit has a great advantage in that multidimensional tables can be prepared easily and the disclosure protection applied will always be consistent. A disadvantage is that it does not look as if disclosure protection has been applied.

Table 11: Example -- Without Disclosure, Protected by Confidentiality Edit

Number of Delinquent Children
by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	13	2	3	2	20
Beta	18	12	8	17	55
Gamma	5	9	11	0	25
Delta	14	12	8	1	35
Total	50	35	30	20	135

SOURCE: Fictitious microdata. Data only for County Alpha shown in Table 10.

D. Tables of Magnitude Data

Tables showing magnitude data have a unique set of disclosure problems. Magnitude data are generally nonnegative quantities reported in surveys or censuses of business establishments, farms or institutions. The distribution of these reported values is likely to be skewed, with a few entities having very large values. Disclosure limitation in this case concentrates on making sure that the published data cannot be used to estimate the values reported by the largest, most highly visible respondents too closely. By protecting the largest values, we, in effect, protect all values.

For magnitude data it is less likely that sampling alone will provide disclosure protection because most sample designs for economic surveys include a stratum of the larger volume entities which are selected with certainty. Thus, the units which are most visible because of their size, do not receive any protection from sampling. For tables of magnitude data, rules called **primary suppression rules** or **linear sensitivity measures**, have been developed to determine whether a given table cell could reveal individual respondent information. Such a cell is called a **sensitive** cell, and cannot be published.

The primary suppression rules most commonly used to identify sensitive cells by government agencies are the (n,k) rule, the p-percent rule and the pq rule. All are based on the desire to make it difficult for one respondent to estimate the value reported by another respondent too closely. The largest reported value is the most likely to be estimated accurately. Primary suppression rules can be applied to frequency data. However, since all respondents contribute the same value to a frequency count, the rules default to a threshold rule and the cell is sensitive if it has too few respondents. Primary suppression rules are discussed in more detail in Section VI.B.1.

Once sensitive cells have been identified, there are only two options: restructure the table and collapse cells until no sensitive cells remain, or cell suppression. With cell suppression, once the sensitive cells have been identified they are withheld from publication. These are called **primary suppressions**. Other cells, called **complementary suppressions** are selected and suppressed so that the sensitive cells cannot be derived by addition or subtraction from published marginal totals. Problems associated with cell suppression for tables of count data were illustrated in Section II.C.3.a. The same problems exist for tables of magnitude data.

An administrative way to avoid cell suppression is used by a number of agencies. They obtain written permission to publish a sensitive cell from the respondents that contribute to the cell. The written permission is called a "waiver" of the promise to protect sensitive cells. In this case, respondents are willing to accept the possibility that their data might be estimated closely from the published cell total.

E. Microdata

Information collected about establishments is primarily magnitude data. These data are likely to be highly skewed, and there are likely to be high visibility respondents that could easily be identified via other publicly available information. As a result there are virtually no public use microdata files released for establishment data. Exceptions are a microdata file consisting of survey data from the Commercial Building Energy Consumption Survey, which is provided by the Energy Information Administration and two files from the 1987 Census of Agriculture provided by the Census Bureau. Disclosure protection is provided using the techniques described below.

It has long been recognized that it is difficult to protect a microdata set from disclosure because of the possibility of matching to outside data sources (Bethlehem, Keller and Panekoek, 1990). Additionally, there are no accepted measures of disclosure risk for a microdata file, so there is no "standard" which can be applied to assure that protection is adequate. (This is a topic for which research is needed, as discussed in Chapter VII). The methods for protection of microdata files described below are used by all agencies which provide public use data files. To reduce the potential for disclosure, virtually all public use microdata files:

1. Include data from only a sample of the population,
2. Do not include obvious identifiers,
3. Limit geographic detail, and
4. Limit the number of variables on the file.

Additional methods used to disguise high visibility variables include:

1. Top or bottom-coding,
2. Recoding into intervals or rounding,
3. Adding or multiplying by random numbers (noise),
4. Swapping or rank swapping (also called switching),

5. Selecting records at random, blanking out selected variables and imputing for them (also called blank and impute),
6. Aggregating across small groups of respondents and replacing one individual's reported value with the average (also called blurring).

These will be illustrated with the fictitious example we used in the previous section.

E.1. Sampling, Removing Identifiers and Limiting Geographic Detail

First: include only the data from a sample of the population. For this example we used a 10 percent sample of the population of delinquent children. Part of the population (County A) was shown in Table 9. Second: remove obvious identifiers. In this case the identifier is the first name of the child. Third: consider the geographic detail. We decide that we cannot show individual county data for a county with less than 30 delinquent children in the population. Therefore, the data from Table 4 shows that we cannot provide geographic detail for counties Alpha or Gamma. As a result counties Alpha and Gamma are combined and shown as AlpGam in Table 12. These manipulations result in the fictitious microdata file shown in Table 12.

In this example we discussed only 5 variables for each child. One might imagine that these 5 were selected from a more complete data set including names of parents, names and numbers of siblings, age of child, ages of siblings, address, school and so on. As more variables are included in a microdata file for each child, unique combinations of variables make it more likely that a specific child could be identified by a knowledgeable person. Limiting the number of variables to 5 makes such identification less likely.

E.2. High Visibility Variables

It may be that information available to others in the population could be used with the income data shown in Table 12 to uniquely identify the family of a delinquent child. For example, the employer of the head of household generally knows his or her exact salary. Such variables are called **high visibility** variables and require additional protection.

E.2.a. Top-coding, Bottom-coding, Recoding into Intervals

Large income values are **top-coded** by showing only that the income is greater than 100 thousand dollars per year. Small income values are **bottom-coded** by showing only that the income is less than 40 thousand dollars per year. Finally, income values are **recoded** by presenting income in 10 thousand dollar intervals. The result of these manipulations yields the fictitious public use data file in Table 13. Top-coding, bottom-coding and recoding into intervals are among the most commonly used methods to protect high visibility variables in microdata files.

Table 12: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited
Delinquent Children

Number	County	HH education	HH income	Race
1	AlpGam	High	61	W
2	AlpGam	Low	48	W
3	AlpGam	Medium	30	B
4	AlpGam	Medium	52	W
5	AlpGam	Very high	117	W
6	Beta	Very high	138	B
7	Beta	Very high	103	W
8	Beta	Low	45	W
9	Beta	Medium	62	W
10	Beta	High	85	W
11	Delta	Low	33	B
12	Delta	Medium	51	B
13	Delta	Medium	59	W
14	Delta	High	72	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

Table 13: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited, Income Top, Bottom and Recoded
Delinquent Children

Number	County	HH education	HH income	Race
1	AlpGam	High	60-69	W
2	AlpGam	Low	40-49	W
3	AlpGam	Medium	<40	B
4	AlpGam	Medium	50-59	W
5	AlpGam	Very high	>100	W
6	Beta	Very high	>100	B
7	Beta	Very high	>100	W
8	Beta	Low	40-49	W
9	Beta	Medium	60-69	W
10	Beta	High	80-89	W
11	Delta	Low	<40	B
12	Delta	Medium	50-59	B
13	Delta	Medium	50-59	W
14	Delta	High	70-79	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

E.2.b. Adding Random Noise

An alternative method of disguising high visibility variables, such as income, is to add or multiply by random numbers. For example, in the above example, assume that we will add a normally distributed random variable with mean 0 and standard deviation 5 to income. Along with the sampling, removal of identifiers and limiting geographic detail, this might result in a microdata file such as Table 14. To produce this table, 14 random numbers were selected from the specified normal distribution, and were added to the income data in Table 12.

Table 14: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited, Random Noise Added to Income
Delinquent Children

Number	County	HH education	HH income	Race
1	AlpGam	High	61	W
2	AlpGam	Low	42	W
3	AlpGam	Medium	32	B
4	AlpGam	Medium	52	W
5	AlpGam	Very high	123	W
6	Beta	Very high	138	B
7	Beta	Very high	94	W
8	Beta	Low	46	W
9	Beta	Medium	61	W
10	Beta	High	82	W
11	Delta	Low	31	B
12	Delta	Medium	52	B
13	Delta	Medium	55	W
14	Delta	High	61	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

E.2.c. Swapping or Rank Swapping

Swapping involves selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all other variables. Swapping (or switching) was illustrated as part of the confidentiality edit for tables of frequency data. In that example records were identified from different counties which matched on race, sex and income and the variables first name of child and household education were swapped. For purposes of providing additional protection to the income variable in a microdata file, we might choose instead to find a match in another county on household education and race and to swap the income variables.

Rank swapping provides a way of using continuous variables to define pairs of records for swapping. Instead of insisting that variables match (agree exactly), they are defined to be close

based on their proximity to each other on a list sorted by the continuous variable. Records which are close in rank on the sorted variable are designated as pairs for swapping. Frequently in rank swapping, the variable used in the sort is the one that will be swapped.

E.2.d. Blank and Impute for Randomly Selected Records

The blank and impute method involves selecting a few records from the microdata file, blanking out selected variables and replacing them by imputed values. This technique is illustrated using data shown in Table 12. First, one record is selected at random from each publishable county, AlpGam, Beta and Delta. In the selected record the income value is replaced by an imputed value. If the randomly selected records are 2 in county AlpGam, 6 in county Beta and 13 in county Delta, the income value recorded in those records might be replaced by 63, 52 and 49 respectively. These numbers are also fictitious, but you can imagine that imputed values were calculated as the average over all households in the county with the same race and education. Blank and impute was used as part of the confidentiality edit for tables of frequency data from the Census sample data files (containing information from the long form of the decennial Census).

E.2.e. Blurring

Blurring replaces a reported value by an average. There are many possible ways to implement blurring. Groups of records for averaging may be formed by matching on other variables or by sorting the variable of interest. The number of records in a group (whose data will be averaged) may be fixed or random. The average associated with a particular group may be assigned to all members of a group, or to the "middle" member (as in a moving average.) It may be performed on more than one variable, with different groupings for each variable.

In our example, we illustrate this technique by blurring the income data. In the complete microdata file we might match on important variables such as county, race and two education groups (very high, high) and (medium, low). Then blurring could involve averaging households in each group, say two at a time. In county Alpha (see Table 9) this would mean that the household income for the group consisting of John and Sue would be replaced by the average of their incomes (139), the household income for the group consisting of Jim and Pete would be replaced by their average (82), and so on. After blurring, the data file would be subject to sampling, removal of identifiers, and limitation of geographic detail.

F. Summary

This chapter has described the standard methods of disclosure limitation used by federal statistical agencies to protect both tables and microdata. It has relied heavily on simple examples to illustrate the concepts. The mathematical underpinnings of disclosure limitation in tables and microdata are reported in more detail in Chapters IV and V, respectively. Agency practices in disclosure limitation are described in Chapter III.