CHAPTER II - Statistical Disclosure Limitation Methods: A Primer

This chapter provides a basic introduction to the disclosure limitation techniques that are commonly used to limit the possibility of disclosing identifying information about respondents in tables and microdata files. The techniques are illustrated with examples. The tables or microdata files 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 theses sections and subsequent chapters. It includes a discussion of the distinction between tables of frequency data and tables of magnitude data, a definition of table dimensionality, and hierarchical variables, 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. 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 it and provide statistical summaries, and/or public use microdata files to the public. Some of the data collected are considered proprietary by respondents.

On the other hand, not all data collected and published by the government are subject to disclosure limitation techniques. Some data on businesses that is collected for regulatory purposes are considered public. In addition, 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 made publicly available. 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. (See Chapter I for a brief discussion of the use of restricted access as opposed to restricted data.)

The most common method of providing data to the public is through statistical tables. With the development of powerful computers with large memory capability and high processing speeds, agencies have started providing an on-line query system with access to a statistical data base. Data users create their own tabulations by customized queries. In most of these systems only data that have already had disclosure limitation applied are available to users. If the unprotected microdata are used as the basis for a query system, disclosure limitation rules must be applied automatically to the requested tables. The concern with the later approach is that users may be able to discern confidential data if they use a sequence of queries in which disclosure limitation is applied independently.

M**icrodata files** are another way agencies attempt to provide user-friendly products. These products have become indispensable to the research community as the release of microdata files for public use has grown. 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, names, addresses and other **direct identifiers** are removed from 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.

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 categories 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). The number in the cell is the count or frequency of the number of people in the population with the cell characteristic. A table that displays value of construction work done during a particular period in the state of Maryland by county and by 4-digit North American Industry Classification System (NAICS) groups is comprised of cells such as the cell {NAICS 4231, Prince George's County}. In this case the number in the cell would be the average value (or aggregate value) of the construction work for companies in the population with the cell characteristics.

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" that applies to 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.

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 (NAICS 4231, 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 (NAICS 42, Prince George's County, 2002). The first two-dimensions are said to be presented in rows and columns, the third variable in "layers" or “pages,” with the layers being a separate table for each category of the third variable.

B.3. Hierarchical Structure of Variables

Most tables are cross tabulations of two or three classification variables such as geography. Classification variables may have a hierarchical structure. A hierarchical coding structure produces subtotals with the variable’s coding structure. For example, the North American Industry Classification System (NAICS) classification variables are variables with a hierarchical structure. Four digits industry codes can be collapsed into three digit codes for major industries and two digits for industry groups. An interior table cell might relate to a specific 4 digit NAICS code, with subtotals given by 3-digit NAICS codes, and the marginal total given by the appropriate 2-digit code. Identifying any hierarchical structure within the classification variables on a file is necessary for applying disclosure limitation techniques, and for assessing protection.

Geography is commonly referred to as a variable with a hierarchical structure. However, this may not always be technically correct depending upon the classification structure. If geography is broken down into states, regions, and national level, then geography would be a hierarchical variable because each state is classified within specific regions. However, if the geographic classification provides locality, metropolitan area, county, state, and region, then the classification may not necessarily be hierarchical because the counties, localities, and metropolitan areas may not be component parts of each other.

B.4. What is Disclosure?

Although the definition of disclosure given in Chapter I is broad, this report documents the methodology used to limit disclosure and is concerned only with the disclosure of confidential information through the public release of data products. 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 identification will automatically reveal additional attribute information that was not used in identifying the record. 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 occurs when confidential information about a person or firm's business operations is revealed or may be closely estimated. Thus, attribute disclosure comprises identification of the subject and divulging confidential information pertaining to the subject.

Attribute disclosure is the primary concern of most statistical agencies in deciding whether to release 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 disguise the attributes of a single respondent.

The third type of disclosure, **inferential disclosure**, occurs when individual 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. There are two main reasons that some statistical agencies are not concerned with inferential disclosure in tabular or micro data. 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, very little data would be released. Second, inferences are designed to predict aggregate behavior, not individual attributes, and thus are often poor predictors of individual data values. Inferential disclosure is still a concern where cases of exceptionally close statistical associations exist and regression models can be used to generate predictions. Inference disclosure is a consideration for reviewing analytical products produced from either a research data center or research project with an agency's restricted access data program. The risk of disclosure may exist in regression models that contain only fully-interactive sets of dummy variables as independent variables. In these cases, agencies need to further examine the potential disclosure risks from the use of certain regression models.

C. On-Line Query Systems

The dissemination of data through the availability of on-line query systems requires special application of disclosure limitation methods. On-line query systems may have multiple capabilities. The simplest form is where the system accesses summary files containing aggregated data that have already been tested for sensitivity and disclosure limitation methods applied. Another capability is the dissemination of tabulations from online queries of microdata files that have already been protected. Applications that access unprotected microdata can introduce a risk of identity disclosure when restricting the query to a small geographic area or category. This is of particular concern for sequences of independent queries about small geographic areas or categories. Specialized tabulations generated from queries to unprotected microdata files must pass through a series of filters where the disclosure limitation rules are applied.

Four agencies have developed on-line query systems with various capabilities for users to generate special tabulations. The Centers for Disease Control and Prevention developed "CDC Wonder" ((Wide-ranging OnLine Data for Epidemiologic Research (WONDER)) at http://www.cdc.gov/nchs/index.htm. The CDC wonder system allows users to submit queries to public-use data sets about mortality (deaths), cancer incidence, HIV and AIDS, behavioral risk factors, diabetes, natality (births), and census data on CDC's mainframe and the requested data are readily summarized. The data are previously tested for sensitivity with disclosure limitation methods applied prior to being added to the database. Users of the CDC wonder system are subject to the agency's data use restrictions that prohibits linking the data with other data sets or information for the purpose of identifying an individual. The Bureau of Labor Statistics also has an online query system available at http://www.bls.gov/data/sa.htm which allows users to access first level summary data (disclosure limitation applied) to generate customized tables.

The Economic Research Service in conjunction with the National Agricultural Statistics Service developed a system available at http://www.ers.usda.gov/Data/ARMS/ for users to generate customized data tables by accessing data from the Agricultural Resource Management Survey (ARMS) program. In the ARMS system, disclosure limitation has already been applied to the microdata. The Census Bureau developed the "American Fact Finder" available at **http://www.census.gov** that provides users with access to both summary tabular data as well as microdata files. The Advanced Query System of American Fact Finder has the sensitivity rules and disclosure methods built into the system so that queries submitted by users must pass disclosure review before the user can view the results. At the National Center for Education Statistics (NCES) all postsecondary sample survey data are available through the use of data analysis tools that produce tables up to three-dimensions and give correlation matrices. In addition, elementary and secondary level data from the National Assessment of Educational Progress (NAEP) are also available in an on-line data tool. A more detailed description of on-line query systems is contained in Chapter 4 Section C.

D. 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 below as an example.

D.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 calculated by multiplying a respondent's data by a sampling weight and then aggregating all the weighted responses. When data are used to make estimates concerning the population from which a sample is drawn, they are generally adjusted by sample weights that take into account the peculiarities of the sampling procedure. Weighted totals take the place of actual frequencies in published tables. The use of sample weights makes an individual respondent's data less identifiable from published totals when the values of the weights themselves are not disclosed. In particular, if the weighting of the survey responses is complex, the published estimate may hide the fact that there are only one or two contributors to a cell. Because the weighted numbers represent all households in the United States, the counts in Table 1 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.

Table 1: Example Without Disclosure

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

1997 Family income

Heated Floor Space sq ft	Total	Less than $10000	$10000 to $24999	$25000 to $49999	$50000 or more	Below Poverty Line	Eligible for Federal Assistance
Fewer than 600	7.9	2.9	3.1	1.6	0.3	2.7	4.9
600 to 999	21.5	4.3	8.6	6.0	2.6	4.6	10.2
1000 to 1599	30.4	2.8	9.7	10.8	7.0	3.7	9.9
1600 to 1999	15.3	.6	3.2	5.4	6.1	0.9	2.8
2000 to 2399	7.9	.2	1.2	2.5	4.0	0.3	1.1
2400 to 2999	5.3	Q	0.3	1.4	3.4	0.2	0.5
3000 or more	4.1	Q	0.3	.9	2.8	Q	0.4

NOTE: Q -- Data withheld because relative standard error exceeds 50% or fewer than 10 households were sampled.
SOURCE: "Housing Characteristics 1997", Residential Energy Consumption Survey, Energy Information Administration, DOE/EIA-0632(97), page 58.

When it is known with certainty that an individual is a study respondent, the task of identifying the person and his/her attributes is much simpler than when there is a high probability that the person is not represented in the table or microdata at all. Should the complete count data reveal that respondent to be unique using information that an individual was a respondent, his or her identity would be confirmed and their attributes revealed. Data collection based upon a sample of persons is protective because the presence of a given person's records is not certain and a respondent who appears to be unique may not be the person he/she is thought to be.

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. 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.

Sampling may lower the disclosure risks from published data depending on the sampling rate, the number and detail of variables tabulated, and whether or not there exists a public listing of the complete population from which the sample is drawn. The sample should also be free of any outlier values such as individuals or establishments with unusual characteristics. The use of sampling methodology does not ensure that the published data are free from disclosure risks and any published tables from a sample should still be reviewed.

D.2. Defining Sensitive Cells

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 to protect against the potential harm to an agency or respondent from disclosing confidential information. Such rules differ from agency to agency and from table to table. These special rules are generally designed to provide protection to data considered particularly sensitive by the agency. The second class is more general where the number in a cell is considered to represent an unacceptable disclosure risk such as: a cell is defined as sensitive if the number of respondents is less than some specified threshold (the threshold rule).

D.2.a Special Rules

Special rules impose restrictions on the level of detail that may be provided in a table. For example, Social Security Administration (SSA) rules prohibit tabulations in which a cell value inside a row or column of a table 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 is a two-dimensional table showing the number of beneficiaries by county and size of benefit. This table could not be released to the public because the data shown for counties B and D violate Social Security's disclosure rules. For county D, there is only one cell with a positive value, and a beneficiary in this county is known to be receiving benefits between $40 and $59 per month. This violates two rules. First the detailed cell is equal to the row 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 cells with positive values, but the range of possible benefits is from $40 to $79 per month, an interval of less than the required $50.

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: FCSM Statistical Policy Working Paper 2.

To protect confidentiality, Table 2 could be restructured and rows or columns combined (sometimes referred to as “rolling-up categories” or “collapsing”). 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: FCSM Statistical Policy Working Paper 2.

D.2.b. The Threshold Rule

With the threshold rule, a cell in a table of frequencies is defined as **sensitive** if the number of respondents is less than some specified number. Some agencies require at least 5 respondents in a cell, while others require 3. Under certain circumstances the number may be much larger. The choice of the minimum number is generally made in consideration of: (a) the sensitivity of the information that the agency is considering to publish, (b) the amount of protection the agency determines to be necessary given the degree of precision required to achieve disclosure.

D.3. Protecting Sensitive Cells After Tabulation

In tables of frequency data, if cells have been identified as being sensitive, the agency must take steps to protect the sensitive data. There are generally two approaches for doing this. One consists of making changes to the table itself. This is done as part of, or after tabulation. These methods include restructuring tables and combining categories (as illustrated above), cell suppression, random rounding, controlled rounding, or controlled tabular adjustment. The second approach that has evolved more recently is the application of microdata methods to the data file prior to tabulation. These methods are particularly efficient for use with on-line query systems or where multiple tables will be created from a single data file. This approach is illustrated in section D.4 of this chapter.

Table 4 is a fictitious example of a table with disclosures. The fictitious data set consists of information concerning delinquent children. Cells in Table 4 with fewer than 5 respondents are defined as sensitive and are identified with an asterisk. This table is used to illustrate cell suppression, random rounding, controlled rounding, and controlled tabular adjustment in the sections below.

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, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

D.3.a. Suppression

One of the most common methods of protecting sensitive cells is by **suppression**. 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 non-sensitive cells must also be suppressed. These are referred to as

complementary suppressions. While it is possible to select cells for complementary suppression manually, in all but the simplest of cases, it is difficult to guarantee that the result provides adequate protection.

Table 5 shows an example of a system of suppressed cells for Table 4 that has at least two suppressed cells in each row and column. This table appears to offer protection to the sensitive cells, however, a closer review shows disclosure of sensitive data still occurs

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	D1	D2	D3	20
Beta	20	D4	D5	15	55
Gamma	D6	10	10	D7	25
Delta	D8	14	7	D9	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.
SOURCE: Numbers taken from Cox, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

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 + D1 + D2 + D3) + (20 + D4 + D5 + 15) - (D1 + D4 + 10 + 14) - (D2 + D5 + 10 + 7) = 20 + 55 - 35 - 30.

This reduces to D3 = 1.

This example shows that selection of cells for complementary suppression is a complicated process. Mathematical methods of linear programming are used to automatically select cells for complementary suppression and also to **audit** a proposed suppression pattern (e.g. 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, McDonald, and Nelson (1986).. Titles, row and column headings are fictitious.

D.3.b. Random Rounding

In order to reduce the amount of data loss that occurs from suppressing sensitive cells in a table alternative data perturbation methods such as random rounding and controlled rounding are available to protect sensitive cells in tables showing frequency data. 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. (A more theoretical discussion of this method is contained in “Elements of Statistical Disclosure Control” by Leon Willenborg and Ton de Waal, 2001).

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.

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, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

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 for 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.

D.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 (see Cox and Ernst, 1982). Linear programming methods are used to identify a controlled rounding for a table. Controlled rounding is used by the Social Security Administration in statistical tables showing frequency counts. Table 8 illustrates controlled rounding where the sum of the cell values in each row and column are constrained to equal the sum of the published totals.

D.3.d. Controlled Tabular Adjustment

Controlled tabular adjustment is a relatively new approach, similar to controlled rounding, but it is most valuable when applied to tables of magnitude data. This method was initially referred to as “synthetic tabular data.” It was described as controlled tabular adjustment in subsequent work (Cox and Dandekar, 2004). For magnitude data, a linear sensitivity rule is used to determine which cells are sensitive. With controlled tabular adjustment each original sensitive value of a table is replaced with a safe value that is a “sufficient distance” away from the true value; and non-sensitive cell values are minimally adjusted to ensure that the published marginal totals are additive. A “sufficient distance” from the true value would be the value needed to be added to the cell total that would make the cell not sensitive according to the linear sensitivity rule being applied. For frequency data, most linear sensitivity rules are equivalent to a threshold rule of 3 respondents and a "sufficient distance" from the true value would involve changing the value by either 1 or 2. That is, the value of a sensitive cell would be changed to either 0 or 3. This is identical to rounding to the base 3.

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, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

Table 9 illustrates a simplified way to implement controlled tabular adjustment, as described in Dandekar (2004). The internal sensitive cells are first listed in descending order from most sensitive to least sensitive (2, 2, 1, 1). Adjustments are applied sequentially beginning with the first cell. The first cell is changed at random to 0 or 3 (by either subtracting 2, or by adding 1.) Subsequent adjustments will be implemented with alternate signs. So if the first cell is altered by adding 1, the second cell is altered by subtracting 2, the third is altered by adding 2, the last is altered by subtracting 1. Once the internal sensitive cells have been altered, no additional changes are needed in the interior non-sensitive cells (as is typically done for controlled rounding). Marginal table totals are re-computed to account for the changes made to the internal sensitive cells. These changes are needed so that the tables add. In Table 9 the marginal totals are adjusted to minimize the percent by which cells are changed. In this example, no changes are needed to the grand total.

Table 9: Example – Without Disclosure -- Protected by Controlled Tabular Adjustment

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* - 1 = 0**	3	**1* + 2 = 3**	**20 + 1 = 21**
Beta	20	10	10	15	55
Gamma	3	10	10	**2* - 2 = 0**	**25 - 2 = 23**
Delta	12	14	7	**2* + 1 = 3**	**35 + 1 = 36**
Total	50	**35 - 1 = 34**	30	**20 + 1 = 21**	135

Controlled tabular adjustments to individual cell values are shown in **Bold** font**.**

D.4. Protecting Sensitive Cells Before Tabulation

Tabular data can be protected by applying disclosure protection methods to the underlying microdata files to assure that any tables that are generated from the microdata files are fully protected. This approach is particularly efficient if there are many tabulations being created from the same data.

The Census bureau has been the leader in applying microdata methods to protect files based on the Decennial Census. Data swapping is illustrated in section II.F.2.c, and is also described in Domingo-Ferrer, (2002). The decennial Census collects basic data from all households in the U.S. It collects more extensive data via the long-form from a sample of U.S. households. Both sets of data are subjected to a data swapping procedure. This technique was used for short form data in the 1990 census, but was revised and extended to the long form data in 2000. The procedure now takes a targeted approach to swapping which increases the effectiveness of the procedure with some cost in terms of bias of variance. All Decennial tabulations come from the swapped files, this guarantees the consistency of the tables and avoids problems associated with protecting interrelated tables.

In 1990, a different procedure was used in the confidentiality edit for the sample data, called "blank and impute", see section II.F.2.d. In this technique, selected records have particular values blanked and treated as missing. Since there are usually pre-existing procedures for imputation of missing data, "blank and impute" has some advantage in economy. However, the procedure reduces effective sample size and the compensation in the calculation of variance is sometimes difficult to accomplish. In some sense, "blank and impute" is a precursor of the synthetic data techniques currently being researched at the Census Bureau and elsewhere (Raghunthan, et. al. 2003). The advantage of data swapping is that it maximizes the information that can be provided in tables. Additionally, all tables are protected in a consistent way.

E. 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 within too close of a range the values reported by the largest, most highly visible respondent. By protecting the largest reported values, we, in effect, are able to protect all values.

Linear sensitivity rules are used to identify cells that are "sensitive" and need to be protected. Recent research has focused on applying protections to the microdata file prior to tabulation. This provides a great advantage, especially if tabulations will be provided through a query system. Historically cell suppression was used to protect sensitive cells in tables. Cell suppression is done as part of the construction of a table.

E.1. Defining Sensitive Cells – Linear Sensitivity Rules

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 that are selected with certainty. Thus, the units that 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. Cells that do not pass the linear sensitivity test are defined as **sensitive** cells, and are withheld from publication.

The primary suppression rules most commonly used to identify sensitive cells by government agencies are the **(n) threshold rule**, **(n, k) rule**, and the **p-percent** or **pq** rules. See Cox, (1981). 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. The p% and pq rules default to a threshold rule of 3 when applied to count data. Primary suppression rules are discussed in more detail in Section VI.B.1.

E.2 Protecting Sensitive Cells After Tabulation

Tables for publication are populated from the microdata files. During aggregation, a linear sensitivity rule is used to identify any sensitive cells. Once sensitive cells have been identified, there are 3 options: restructure the table and collapse cells until no sensitive cells remain, use cell suppression, or apply controlled tabular adjustment. 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 C.3.a of this chapter. The same problems exist for tables of magnitude data.

Controlled tabular adjustment was illustrated for frequency data in Section C.3.d. of this chapter. For magnitude data, the "sufficient distance" is the amount that would need to be added to the cell total so that the linear sensitivity rule would classify the cell as not sensitive.

An administrative way to avoid cell suppression is used by a number of agencies. They obtain written permission, or "**informed consent**" 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 and specific authorization or consent to the agency for publicly releasing the confidential information. In this case, respondents are requested by an agency to voluntarily give their consent after being informed of the need to release the confidential information, and the proposed statistical or non-statistical use of the information. This method is most useful with small surveys or sets of tables involving only a few small cells, where only a few waivers are needed. Of course, respondents must be informed of the proposed use of the data prior to giving their consent.

E.3. Protecting Sensitive Cells Before Tabulation

There are few microdate products for establishment surveys because of the skewed nature of the population. However, applying microdata methods to protect files of establishment data prior to tabulation has simplified the protection of tabular data and provided new data products.

The Census Bureau was the first to apply microdata methods to protect establishment level data files prior to tabulation. The technique of noise addition, section II.F.2.b, has been the primary method used, in conjunction with other methods. In particular, noise addition has been used to protect quarterly workforce indicators released from the Longitudinal Employer Household Dynamics project. Magnitude data for establishments tends to be skewed and dominated by large companies. This can lead to a situation where applying linear sensitivity rules flags many cells for protection against disclosure. Noise addition adds noise to each responding establishment's data by a small percentage. The amount of the perturbation of the reported value depends on the magnitude of the reported value, and the value of the linear sensitivity rule for the cells containing that respondent's data. If a cell contains only one establishment, or if a single establishment dominates a cell, the published value in a cell will not be a close approximate to the dominant establishment's value because that value has had noise added to it. The dominant establishment's true reported value is protected by the noise addition. It is important to note that all establishments have their values multiplied by a corresponding noise factor, or adjusted weight, before the data are tabulated. The noise multipliers can be randomly assigned to control the effects of the noise on different types of cells within a table.

Noise addition was also used by the U.S. Department of Agriculture's Economic Research Service to protect the reported values in their annual Agricultural Resource Management Survey (ARMS) that is available through an on-line query system. The values are adjusted alternating between adding and subtracting noise following the order of observations in the data set so that the cell totals are approximately the same after the noise addition is applied.

The method has several advantages over cell suppression in that in provides some information in more cells of the table, and it eliminates the need to coordinate cell suppression patterns. This methodology provides consistency in the tables generated from the microdata, but it is important that the initial microdata have been sufficiently perturbed so that the tables produced are safe for release. One limitation of this methodology is that marginal values can show large changes as a result of adjusting the underlying weights. The relationship between the actual unadjusted cell values and adjusted cell values using noise addition should be reviewed prior to releasing the data.

F. Microdata

Information collected about establishments is primarily magnitude data. These data are likely to be highly skewed, and there are likely to be high risk respondents that could easily be identified via other publicly available information. As a result, special care must be taken when considering the release of microdata files containing establishment data. Examples of the public release of microdata files from establishment surveys include data from the Commercial Building Energy Consumption Survey, which is provided by the Energy Information Administration, and files from the 1997 Census of Agriculture provided by the Census Bureau. Disclosure protection is provided using the techniques described below in addition to removing variables that serve as direct identifiers of respondents to the survey.

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. A “Checklist on Disclosure Potential of Proposed Data Releases” was developed by the Confidentiality and Data Access Committee to assist agencies in reviewing the disclosure potential of proposed public use microdata files and is available for download at http://www.fcsm.gov/committees/cdac/. The Bureau of Labor Statistics, Bureau of Transportation Statistics, National Center for Health Statistics, Census Bureau, and Social Security Administration use the CDAC checklist or some modified format of the checklist for reviewing proposed data releases for any disclosure potential. The National Science Foundation also uses the CDAC checklist as guidelines for their contractors to follow when reviewing a proposed file for public release. 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, most 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 and detailed breakdown of categories within variables on the file.

Additional methods used to disguise high risk variables include:

1. Truncation of extreme codes for certain variables (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.

F.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. Second: remove identifiers that directly identify respondents such as name, address, and identification numbers. 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 9. These manipulations result in the fictitious microdata file shown in Table 10.

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 may be identified by a knowledgeable person. Limiting the number of variables to 5 makes such identification less likely.

F.2. High Risk Variables

It may be that information available to others in the population could be used with the income data shown in Table 10 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. Variables such as income, race, and age are **high risk** variables and require additional protection.

Table 10: 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	59	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.

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

In the example, large income values are **top-coded** by showing only that the income is greater than 100,000 dollars per year. Small income values are **bottom-coded** by showing only that the income is less than 40,000 dollars per year. Finally, income values are **recoded** by presenting income in 10,000 dollar intervals. The result of these manipulations yields the fictitious public use data file in Table 11. Top-coding, bottom-coding and recoding into intervals are among the most commonly used methods to protect high risk variables in microdata files.

Table 11: Fictitious Microdata -- Sampled, Identifiers Removed

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

Geographic Detail Limited 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.

F.2.b. Adding Random Noise

An alternative method of disguising high risk 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 12. To produce this table, 14 random numbers were selected from the specified normal distribution, and were added to the income data in Table 10.

Table 12: 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.

F.2.c. Data Swapping and Rank Swapping

Swapping involves selecting a sample of the records, finding a match in the database on a set of predetermined variables and swapping all other variables. Swapping is illustrated in section E.2.e. In that example records were identified from different counties that 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.

Swapping offers the opportunity to select some statistics that will be preserved through the swapping operation. This is accomplished by forcing agreement between the swapped pairs on the variables involved in those statistics. The National Institute of Statistical Sciences (NISS) has a software package which performs and analyzes data swapping in categorical data variables that is available from their website at http://www.niss.org/software/dstk.html. The NISS technique uses random swapping; this affords one the ability to quantify the effect on statistics produced from the swapped data set. For data sets with an accurate measure of record level risk, one can employ a variation, termed targeted swapping. Those records with high risk are automatically selected for pairing in the swap process. In targeted swapping, fewer records are involved and the protection level is generally higher. However, the targeted procedure is biased and the ability to present a general statement on data quality is very limited.

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 that 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.

Data Shuffling is another method for modifying micro data that has been applied to numerical data. The procedure involves two steps: first the values of the confidential variables are modified using a general perturbation technique and second, a data shuffling procedure is applied using the perturbed values of the confidential variables on the file. The perturbed values are sorted from lowest to highest value in the re-shuffled file. Then the perturbed value is replaced with the original value of the confidential variable based on the ranking of the original values from the confidential variable. Before the data are perturbed, the conditional distribution between the confidential and non-confidential variables is derived. This method preserves the rank order correlation between the confidential and non-confidential attributes, and avoids the loss in data utility that could occur from applying data swapping or rank swapping methodology. Data shuffling is discussed in more detail in Chapter V.

Data swapping was used to protect the confidentiality of the Census 2000 tabulations. The procedure was performed on the underlying microdata, and all tabulations from the 100% (short form) and from the sample (long form) data were created from the swapped files. It affected pairs of households (or partnered households) where one or both of those households had a high risk of disclosure. The set of census households that were deemed as having a disclosure risk were selected from the internal census data files. These households were unique in their geographic area (block for 100% data and block group for sample data) based on certain characteristics. The data from these households were swapped with data from partnered households that had identical characteristics on a certain set of key variables but were from different geographic locations. Which households were swapped is not public information. The swapping procedure was performed independently for the 100% data and the sample data. To maintain data quality, there was a maximum percent of records that were swapped for each state for the 100% data and another maximum percent for the sample data.

To illustrate the set of data swapping procedures that were 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 13 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 data swapping 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 data swapping procedures:

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

Table 13: Fictitious Microdata

All Delinquent Children in County Alpha

Number	Child	County	HH education	HH income	Race	Sex
1	John	Alpha	Very high	201	B	M
2	Jacob	Alpha	High	103	W	M
3	Sue	Alpha	High	75	B	F
4	Pete	Alpha	High	61	W	M
5	Ramesh	Alpha	Medium	72	W	M
6	Dante	Alpha	Low	103	W	M
7	Larry	Alpha	Low	91	B	M
8	Marilyn	Alpha	Low	84	W	F
9	Steve	Alpha	Low	75	W	M
10	Paul	Alpha	Low	62	B	M
11	Renee	Alpha	Low	58	W	F
12	Virginia	Alpha	Low	56	B	F
13	Mary	Alpha	Low	54	B	F
14	Laura	Alpha	Low	52	W	F
15	Tom	Alpha	Low	55	B	M
16	Al	Alpha	Low	48	W	M
17	Mike	Alpha	Low	48	W	M
18	Phil	Alpha	Low	41	B	M
19	Brian	Alpha	Low	44	B	M
20	Nancy	Alpha	Low	37	W	F

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

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 Alpha. 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 14.

4. Use the swapped data file directly to produce tables. See Table 15.

Applying the set of data swapping procedures has a great advantage in that multidimensional tables can be prepared easily and the disclosure protection applied will always be consistent.

Table 14: Fictitious Microdata

Delinquent Children After Swapping -- Only County Alpha Shown

Number	Child	County	HH education	HH income	Race	Sex
1	John	Alpha	Very high	201	B	M
2	Jacob	Alpha	High	103	W	M
3	Sue	Alpha	High	75	B	F
4*	**Alfonso**	**Alpha**	**Very high**	**61**	**W**	M
5	Ramesh	Alpha	Medium	72	W	M
6	Dante	Alpha	Low	103	W	M
7	Larry	Alpha	Low	91	B	M
8	Marilyn	Alpha	Low	84	W	F
9	Steve	Alpha	Low	75	W	M
10	Paul	Alpha	Low	62	B	M
11	Renee	Alpha	Low	58	W	F
12*	**June**	**Alpha**	**High**	**56**	**B**	F
13	Mary	Alpha	Low	54	B	F
14	Laura	Alpha	Low	52	W	F
15	Tom	Alpha	Low	55	B	M
16	Al	Alpha	Low	48	W	M
17*	**George**	**Alpha**	**Medium**	**48**	**W**	M
18	Phil	Alpha	Low	41	B	M
19	Brian	Alpha	Low	44	B	M
20*	**Heather**	**Alpha**	**Low**	**37**	**W**	F

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

Table 15: Table Protected By Data Swapping

Number of Delinquent Children by County and 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.

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

The blank and impute method involves deleting the values for selected variables for selected respondents from the microdata file and replacing them with values for those same variables from other respondents or through modeling. This technique is illustrated using data shown in Table 16.

Table 16: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited using Blank and Impute - Delinquent Children

Number	County	HH Education	HH Income	Race
1	AlpGam	High	61	W
2	AlpGam	Low	***63***	W
3	AlpGam	Medium	30	B
4	AlpGam	Medium	52	W
5	AlpGam	Very High	117	W
6	Beta	Very High	***52***	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	59	B
13	Delta	Medium	***49***	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.

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 1990 Census sample data files (containing information from the long form of the decennial Census).

F.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 education group, such as 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 can be subject to sampling, removal of identifiers, and limitation of geographic detail to further reduce the risk of identification.

F.2.f. Targeted Suppression

Although **suppression** is one of the most commonly used ways of protecting sensitive cells in tables, it may also be used on records in microdata files. When a record contains extreme values or unique values that cannot be adequately protected, it may be necessary to delete the single record in its entirety, or suppress the sensitive values for certain variables on the record.

G. Summary

This chapter describes the standard methods of disclosure limitation used by federal statistical agencies to protect both tables and microdata. It relies heavily on simple examples to illustrate the concepts. A consideration when evaluating different methods is that records subject to swapping, blanking and imputation, and blurring methodologies are not distinguished (or flagged) in any way on a file. This means that not only are the adjusted records protected, but a high degree of uncertainty is introduced such that whatever methods are used to isolate any particular record, the user will not be able to determine with certainty that the isolated record contains actual and not swapped, imputed or blurred values. The mathematical underpinnings of applying disclosure limitation methodology 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.