Methods for Tabular Data

Chapter II presented examples of disclosure limitation techniques used to protect tables and microdata. Chapter III described agency practices in disclosure limitation. This chapter presents more detail concerning methodological issues regarding confidentiality protection in tables.

As mentioned earlier, tables are classified into two categories for confidentiality purposes: tables of frequency (or count) data and tables of magnitude data. Tables of frequency data show the percent of the population which have certain characteristics, or equivalently, the number in the population which have certain characteristics. If a cell has only a few respondents and the characteristics are sufficiently distinctive, then it may be possible for a knowledgeable user to identify the individuals in the population. Disclosure limitation methods are applied to cells with fewer than a specified **threshold number** of respondents to minimize the risk that individuals can be identified from their data. Disclosure limitation methods include cell suppression, data perturbation methods and the confidentiality edit.

Tables of magnitude data typically present the results of surveys of organizations or establishments, where the items published are aggregates of nonnegative reported values. For such surveys the values reported by respondents may vary widely, with some extremely large values and some small values. The confidentiality problem relates to assuring that a person cannot use the published total and other publicly available data to estimate an individual respondent's value too closely. Disclosure limitation methods are applied to cells for which a **linear sensitivity measure** indicates that some respondent's data may be estimated too closely. For tables of magnitude data cell suppression is the only disclosure limitation method used.

Tables of frequency data are discussed in Section A. The major methodological areas of interest are in controlled rounding and the confidentiality edit. Tables of magnitude data are discussed in Section B. This section provides some detail concerning linear sensitivity measures, auditing of proposed suppression patterns and automated cell suppression methodologies.

A. Tables of Frequency Data

Tables of frequency data may relate to people or establishments. Frequency data for establishments are generally not considered sensitive because so much information about an establishment is publicly available. Disclosure limitation techniques are generally applied to tables of frequencies based on demographic data. As discussed earlier, the most commonly used **primary disclosure rule** for deciding whether a cell in a table of frequency data reveals too much information is the "threshold rule". A cell is defined to be sensitive when the number of respondents is less than some predetermined threshold. If there are cells which are identified as being sensitive, steps must be taken to protect them.

The methods of preventing disclosure in tables of counts or frequencies were illustrated in II.C.2. Included are: combining cells, cell suppression, perturbation methods, random rounding, controlled rounding and the confidentiality edit. Combining cells is generally a judgmental activity, performed by the survey manager. There are no methodological issues to discuss. Selection of cells for complementary suppression is the same problem for both tables of frequencies and tables of magnitude data. Complementary suppression will be discussed in Section B.2 of this Chapter.

Perturbation methods include random rounding and controlled rounding as special cases, and controlled rounding is a special case of random rounding. Controlled rounding is the most desirable of the perturbation methods, because it results in an additive table (sums of row, column and layer entries are equal to the published marginal total), can always be solved for two-dimensional tables, and can generally be solved for three-dimensional tables. Section 1 provides more detail on the methodology used in controlled rounding. The confidentiality edit is a relatively new technique and was used by the Census Bureau to publish data from the 1990 Census. The confidentiality edit is discussed in Section 2.

A.1. Controlled Rounding

Controlled rounding was developed to overcome the shortcomings of conventional and random rounding and to combine their desirable features. Examples of random rounding and controlled rounding were given in II.C.2. Like random rounding, controlled rounding replaces an original two-dimensional table by an array whose entries are rounded values which are adjacent to the corresponding original values. However, the rounded array is guaranteed to be additive and can be chosen to minimize any of a class of standard measures of deviation between the original and the rounded tables.

A solution to the controlled rounding problem in two-dimensional tables was found in the early 1980's (Cox and Ernst, 1982). With this solution the table structure is described as a mathematical network, a linear programming method which takes advantage of the special structures in a system of data tables. The network method can also be used to solve controlled rounding for sets of two-dimensional tables which are related hierarchically along one dimension (Cox and George, 1989). For three-dimensional tables an exact network solution does not exist. Current methods make use of an iterative approximate solution using a sequence of two-dimensional networks. The exact solutions for two-dimensional tables and the approximate solutions for three-dimensional tables are both fast and accurate.

Current research focuses on refinements to the two-dimensional problem and solutions to the three-dimensional problem. These are described in Greenberg (1988a); Fagan, Greenberg and Hemmig (1988); Kelly, Assad and Golden (1990); Kelly, Golden, Assad and Baker (1990); and Kelly, Golden and Assad (1990c).

Although solutions to the controlled rounding problem are available, controlled rounding has not been used by U.S. government agencies.

A.2. The Confidentiality Edit

The newest approach to protection of tables of frequency data is the confidentiality edit, (Griffin, Navarro and Flores-Baez, 1989), which was illustrated in II.C.2.d. 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. In 1990 the confidentiality edit used different procedures to protect tables based on these two systems of data. For the basic decennial Census data (the 100 percent data file) a small sample of households were selected and matched with households in other geographic regions that had identical characteristics on a set of selected key variables. All variables in the matched records were interchanged. This technique is called switching. The key variables used for matching were selected to assure that Census aggregates mandated by law would be unchanged by the confidentiality edit.

The effectiveness of the data switching procedure was investigated by simulation (Navarro, Flores-Baez and Thompson, 1988). It was found that switching provides adequate protection except in areas with small populations (blocks). The solution used by the Census Bureau was to use higher sampling fractions to select households for switching in such areas.

For the sample file, consisting of the data collected on the long form, the sampling was shown to provide adequate protection except in small geographic regions (blocks). In these regions one record was selected and a sample of the variables on the record were blanked and replaced by imputed data. This procedure is called "blank and impute". Both "blank and impute" and "switching" have been suggested as methods to provide disclosure limitation to microdata files.

Once the data for the sampled households were switched in the 100% microdata file and blank and impute was done in the sample file, the files were used directly to prepare all Census tabulations. The advantage of the confidentiality edit is that it maximizes the information that can be provided in tables. Additionally, all tables are protected in a consistent way.

B. Tables of Magnitude Data

For tables of magnitude data the values reported by respondents are aggregated in the cells of a table. Examples of magnitude data are income for individuals and sales volumes and revenues for establishments. Particularly for establishments these reported values are typically highly skewed with a few very large reported values which might easily be associated with a particular respondent by a knowledgeable user. As a result, a more mathematical definition of a **sensitive cell** is needed for tables of magnitude data. For tables of frequency data each respondent contributes equally to each cell, leading to the simple threshold definition of a sensitive cell.

Mathematical definitions of sensitive cells are discussed in Section B.1 below. Once the sensitive cells are identified, a decision must be made as to how to prevent disclosure from occurring. For tables of magnitude data the possibilities include combining cells and rolling up categories, and cell suppression. All were summarized and illustrated in Chapter II.

In the combination method tables are redesigned (categories rolled-up) so there are fewer sensitive cells. Table redesign methods are useful exercises, particularly with tables from a new survey or where portions of a table contain many sensitive cells because the population is sparse. However, it is not generally possible to eliminate all sensitive cells by collapsing tables, and rigorous automated procedures for collapsing in general remain to be developed.

The historical method of protecting sensitive cells in tables of magnitude data is cell suppression. Sensitive cells are not published (they are suppressed). These sensitive suppressed cells are called **primary suppressions**. To make sure the primary suppressions cannot be derived by subtraction from published marginal totals, additional cells are selected for **complementary suppression**. Complementary suppressions are sometimes called **secondary suppressions**.

For small tables, it is possible to manually select cells for complementary suppression, and to apply audit procedures (see Section 2.a) to guarantee that the selected cells adequately protect the sensitive cells. For large scale survey publications with many related tables, the selection of a set of complementary suppression cells which are "optimal" in some sense is an extremely complex problem. Complementary suppression is discussed in Section 2.b.

Instead of suppressing data, some agencies ask respondents for permission to publish cells even though they are sensitive. This is referred to as the waiver approach. Waivers are signed records of the respondents permission to publish. 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 believe that the data are not particularly sensitive before they will sign waivers.

B.1. Definition of Sensitive Cells

The definitions and mathematical properties of linear sensitivity measures and their relationship to the identification of sensitive cells in tables were formalized by Cox (1981). This is one of the important advancements since Working Paper 2. Although the common linear sensitivity rules were known in 1978 and were used to identify sensitive cells, their mathematical properties had not been formally demonstrated. The important definitions and properties are given below.

For a given cell, X, with N respondents the respondent level data contributing to that cell can be arranged in order from large to small: $x_1 \geq x_2 \geq \ldots x_N \geq 0$. Then, an **upper linear sensitivity measure, S(X),** is a linear combination

$$S(X) = \sum_{i=1}^{N} w_i x_i$$

defined for each cell or cell union X and its respondent data $\{x_i\}$. The sequence of constants, $\{w_i\}$, is called the sequence of weights of S(X). These weights may be positive or negative. A cell or cell union X is **sensitive** if $S(X) > 0$. Note that multiplying a linear sensitivity measure by a constant yields another (equivalent) linear sensitivity measure. The linear sensitivity

measures described in this section are all normalized so that the weight multiplying x_1 is equal to 1. This normalization makes it easier to compare them.

If a respondent contributes to two cells, X and Y, then it remains a single respondent to the union of X and Y, with value equal to the sum of its X and Y contributions.

One of the properties which assists in the search for complementary cells is **subadditivity**, which guarantees that the union of disjoint cells which are not sensitive is also not sensitive. Cox shows that a linear sensitivity measure is subadditive if the sequence of weights is nonincreasing, i.e. if $w_1 \geq w_2 \geq ... \geq w_N$. Subadditivity is an important property because it means that aggregates of cells which are not sensitive are not sensitive and do not need to be tested.

Valid complementary cells have the property that their union with the sensitive cell(s) in a row, column or layer where marginal totals are published is not sensitive according to the linear sensitivity measure. A simple result is that zero cells are not valid candidates for complementary suppression as the union of a sensitive cell and a zero cell is equal to the sensitive cell, and is therefore still sensitive. Complementary suppressions may not be needed if marginal totals are not published.

The commonly used primary suppression rules are described Sections a, b, and c below. They are compared in Section d. Each of these rules involves parameters which determine the values taken by the weights, w_1 ..., w_N. Although agencies may reveal the primary suppression rule they use, they should not disclose parameter values, as knowledge of the rule and its parameters enables a respondent to make better inferences concerning the values reported by other respondents. An example is presented in Section 3.

There are three linear sensitivity measures which have been discussed in the literature and used in practical applications. These are the p-percent rule, the pq rule and the (n,k) rule. They are described below. All are subadditive, as can be seen by the fact that the weights in the equations defining S(X) are non-increasing.

B.1.a. The p-Percent Rule

Approximate disclosure of magnitude data occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is declared sensitive, if upper and lower estimates for the respondent's value are closer to the reported value than a prespecified percentage, p. This is referred to as the "p-percent estimation equivocation level" in Working Paper 2. It is more generally referred to as the **p-percent rule**, and has linear sensitivity measure,

$$S^{p\%}(X) = x_1 - \frac{100}{p} \sum_{i=c+2}^{N} x_i .$$

Here, c is the size of a coalition, a group of respondents who pool their data in an attempt to estimate the largest reported value. The cell is sensitive if $S^{p\%}(X) > 0$.

Note that if there are less than 3 respondents (N < 3) in cell X, then $S^{p\%}(X) = x_1 > 0$ and the cell is sensitive for all values of p and c.

The p-percent rule is derived as follows. Assume that from general knowledge any respondent can estimate the contribution of another respondent to within 100-percent of its value. This means that the estimating respondent knows that the other respondents' values are nonnegative and less than two times the actual value. For the p-percent rule, it is desired that after the data are published no respondent's value should be estimable more accurately than within p percent (where p < 100).

It can be shown that the coalition including the second largest respondent is in a position to estimate the value of x_1 most accurately, and that if x_1 is protected, so are all the smaller respondents. Thus, it suffices to provide the protection to the largest respondent, and to assume that the estimating party is a coalition of the second largest respondent and the next largest c-1 respondents. As the coalition respondents may estimate each of x_{c+2}, ..., x_N to within 100 percent, they have an estimate for the sum of these smallest respondents, E, such that

$$| \sum_{i=c+2}^{N} x_i - E | \leq \sum_{i=c+2}^{N} x_i .$$

They can estimate the value of x_1 by subtracting the value they reported to the survey ($\sum_{i=2}^{c+1} x_i$) and their estimate of the smaller respondent's total, E, from the published total. The error in this estimate will be equal to the error in estimating E, which is less than or equal to $\sum_{i=c+2}^{N} x_i$.

The requirement that this estimate be no closer than p-percent of the value of x_1 (p < 100) implies that

$$\sum_{i=c+2}^{N} x_i \geq \frac{p}{100} x_1 .$$

This can be rewritten as the linear sensitivity rule above.

B.1.b. The pq Rule

In the derivation for the p-percent rule, we assumed that there was limited prior knowledge about respondent's values. Some people believe that agencies should not make this assumption. In the pq rule, agencies can specify how much prior knowledge there is by assigning a value q which represents how accurately respondents can estimate another respondent's value before any data are published (p < q < 100). Thus, there is an improved estimate, E_2, of $\sum_{i=c+2}^{N} x_i$ with the property that

$$\left| \sum_{i=c+2}^{N} x_i - E_2 \right| \leq \frac{q}{100} \sum_{i=c+2}^{N} x_i .$$

This leads directly to a more accurate estimate for the largest respondent's value, x_1. The requirement that this estimate be no closer than p-percent of the value of x_1 implies that

$$\frac{q}{100} \sum_{i=c+2}^{N} x_i \geq \frac{p}{100} x_1 .$$

This can be rewritten as the linear sensitivity rule

$$S^{pq}(X) = x_1 - \frac{q}{p} \sum_{i=c+2}^{N} x_i .$$

Note that the pq rule (sometimes called a prior-posterior ambiguity rule) and the p-percent rule are identical if the ratio q/p, the "information gain", is equal to 100/p. In the table below we use the ratio q/p as a single parameter for the pq rule. If users fix a value for p and a value for q < 100, the pq rule is more conservative (will suppress more cells) than the p-percent rule using the same value of p.

Note that if there are fewer than 3 respondents (N < 3), then $S^{pq} = x_1 > 0$ and cell X is sensitive for all values of c and q/p.

Most frequently the pq rule is given with the size of a coalition equal to 1. In this case the linear sensitivity rule is given by

$$S^{pq}(X) = x_1 - \frac{q}{p} \sum_{i=3}^{N} x_i .$$

B.1.c. The (n,k) Rule

The **(n,k) rule**, or dominance rule was described as follows in Working Paper 2. "Regardless of the number of respondents in a cell, if a small number (n or fewer) of these respondents contribute a large percentage (k percent or more) of the total cell value, then the so-called **n respondent, k percent rule** of cell dominance defines this cell as sensitive." Many people consider this to be an intuitively appealing rule, because, for example, if a cell is dominated by one respondent then the published total alone is a natural upper estimate for the largest respondent's value. Although coalitions are not specifically discussed in the derivation of the (n,k) rule, agencies select the value of n to be larger than the number of any suspected coalitions. Many agencies use an (n,k) rule with n = 1 or 2.

The linear sensitivity measure for the (n,k) rule is given by

$$S^{(n,k)}(X) = \sum_{i=1}^{n} x_i - \frac{k}{100 - k} \sum_{i=n+1}^{N} x_i .$$

If $N \leq n$, $S^{(n,k)} = \sum_{i=1}^{N} x_i > 0$ and cell X is sensitive for all values of k.

B.1.d. The Relationship Between (n,k) and p-Percent or pq Rules

Table 1 is designed to assist users in selecting a value of the parameter p for use with the p-percent rule with coalitions of size 1 (or for the value of the ratio, q/p, for the pq rule with coalitions of size 1) when they are used to thinking in terms of the (n,k) rule. For various values of p% (q/p), the table shows the value of k_1 and k_2 such that if the linear sensitivity rule for $(1,k_1)$ or $(2,k_2)$ is positive then the linear sensitivity rule for the p-percent (p/q) rule will be positive. With this formulation, the p-percent (pq) rule is more conservative. It will suppress more cells than will either of the two (n,k) rules individually, and also more than the combination rule based on the two (n,k) rules.

The derivation of the inequalities used in Table 1 are presented in the Technical Notes at the end of this Chapter. Additionally, the sensitivity regions for (n,k), p-percent, and pq rules are illustrated graphically in the Technical Notes.

To illustrate the use of Table 1, if the analyst wants to make sure that a cell where the largest respondent contributes more than 75 percent of the total is suppressed, and that a cell where the largest two respondents exceed 85 percent of the total is suppressed, he/she could approximately accomplish this by using the p-percent rule with p equal to 33.3 percent, or the pq rule with information gain, q/p=3.

The p-percent, pq and (n,k) rules as well as the combination rule,

$$S^{comb} = \max(S^a(X), S^b(X))$$

are subadditive linear sensitivity rules. (Here $S^a(X)$ and $S^b(X)$ denote any two subadditive linear sensitivity measures.) Any of these rules is acceptable from a mathematical point of view.

However, the p-percent or pq rule is preferred for two major reasons. First, the tolerance interval concept directly parallels methods currently used for complementary suppression, (see section B.2.a.iii). Second, as illustrated in the table above and the example in the Technical Notes, the p-percent (pq) rule provides more consistent protection areas than a single version of the (n,k) rule.

TABLE 1
Relationship Between Suppression Regions for
p-Percent or (pq) Rule and (1,k), (2,k) Rules

		$S^{p\%}(X) > 0$ and	Sensitive when cell
p	q/p	x_1/T exceeds:	$(x_1+x_2)/T$ exceeds:
50.0%	2	66.7%	80.0%
33.3%	3	75.0%	85.7%
16.7%	6	85.7%	92.3%
11.1%	9	90.0%	94.7%

NOTE: $T = \sum_{i=1}^{N} x_i$ is the cell total.

B.2. Complementary Suppression

Once sensitive cells are identified by a primary suppression rule, other nonsensitive cells must be selected for suppression to assure that the respondent level data in sensitive cells cannot be estimated too accurately. This is the only requirement for a proposed set of complementary cells for tables of magnitude data and is generally considered to mean that a respondent's data cannot be estimated more closely than plus or minus some percentage.

There are two ways respondent level data can be compromised. First, implicitly published unions of suppressed cells may be sensitive according to the linear sensitivity measure. This depends on the characteristics of the respondent level data in the cell union, and tends to be a problem only where the same respondents contribute to both cells. Second, the row and column equations represented by the published table may be solved, and the value for a suppressed cell estimated too accurately. Automated methods of **auditing** a proposed suppression pattern may be needed to assure that the primary suppressions are sufficiently well protected (see Section B.2.a).

Any set of cells proposed for complementary suppression is acceptable as long as the sensitive cells are protected. For small tables this means that selection of complementary cells may be done manually. Typically the data analyst knows which cells are of greatest interest to users (and should not be used for complementary suppression if possible), and which are of less interest to users (and therefore likely candidates for complementary suppression.) Manual selection of complementary cells is acceptable as long as the resultant table provides sufficient protection to the sensitive cells. An automated audit should be applied to assure this is true.

For large systems of tables, for example, those based on an Economic Census, the selection of complementary cells is a major effort. Manual selection of cells may mean that a sensitive cell is inadvertently left unprotected or that consistency is not achieved from one table to another in a publication. Inconsistency in the suppression patterns in a publication increases the likelihood of inadvertent disclosure. For this reason linear programming techniques have been applied to the selection of cells for complementary suppression by statistical agencies for many years. As an additional benefit, agencies expect automated selection of the complementary cells will result in less information lost through suppression. Examples of the theory and methods for automatic selection of cells for complementary suppression are discussed in Section B.2.b.

B.2.a. Audits of Proposed Complementary Suppressions

B.2.a.i. Implicitly Published Unions of Suppressed Cells Are Sensitive

If sensitive cells are protected by suppressing other internal table cells but publishing the marginal totals, the implicit result is that the unions of the suppressed cells in rows, columns and layers are published. Thus, one way to audit the protection supplied by the suppression pattern is to apply the linear sensitivity rule to those unions to assure that they are not sensitive. While this type of audit is a simple matter for small tables, Cox (1980) points out that for large tables it may be computationally intractable unless a systematic approach is used. This type of audit is not included in standard audit software because of its dependence on respondent level data.

Clearly a table for which suppression patterns have been selected manually requires an audit to assure that the pattern is acceptable. Early versions of complementary suppression software used approximation arguments to select cells for complementary suppression (individual respondent data were not used.) These methods did guarantee that unions of suppressed cells were not sensitive as long as different respondents contributed to each cell. However, if the same respondents contributed to multiple cells in a cell union, an audit was needed.

B.2.a.ii. Row, Column and/or Layer Equations Can Be Solved for Suppressed Cells

A two-dimensional table with row and column subtotals and a three-dimensional table with row, column and layer subtotals can be viewed as a large system of linear equations. The suppressed cells represent unknown values in the equations. It is possible that the equations can be manipulated and the suppressed values estimated too accurately. Audits for this type of disclosure require the use of linear programming techniques. The output of this type of audit is the maximum and the minimum value each suppressed cell can take given the other information in the table. When the maximum and the minimum are equal, the value of the cell is disclosed exactly. To assure that cells cannot be estimated too accurately the analyst makes sure the maximum and the minimum value for the suppressed cell are no closer to the true value than some specified percentage protection.

It is well known that a minimal suppression pattern where marginal totals are presented will have at least two suppressed cells in every row, column and layer requiring a suppression. This is not sufficient, however, as was illustrated in II.C.2.a.

B.2.a.iii. Software

Automated methods of auditing a suppression pattern for the second problem have been available since the mid 1970's at the U.S. Census Bureau, (Cox, 1980) and at Statistics Canada, (Sande, 1984). Modern versions of audit software set up the linear programming problem as described in Zayatz (1992a) and use commercially available linear programming packages. All audit systems produce upper and lower estimates for the value of each suppressed cell based on linear combinations of the published cells. The data analyst uses the output from the audit to determine whether the protection provided to the sensitive cells by the proposed complementary cells is sufficient. These audit methods are applicable to tables of both magnitude and frequency.

In more recent formulations of the complementary suppression problem at the U. S. Census Bureau both types of audits are subsumed into the algorithm that selects cells for complementary suppression. The company level contributions for a cell are used in selecting a protection level or tolerance interval for each cell which will provide protection to all respondents in the cell, and the algorithm which selects cells for complementary suppression now assures that the primary cells cannot be estimated more accurately than that specified tolerance interval. The complementary suppressions selected by such computer systems do not require additional audits.

B.2.b. Automatic Selection of Cells for Complementary Suppression

Automatic methods,of selecting cells for complementary suppression have also been available since the late 1970's at Statistics Canada, (Sande, 1984), and at the U. S. Census Bureau, (Cox, 1980). These programs typically rely on linear programming methods, either using standard approaches or approaches which make use of special structures in the data, such as network theory. The Statistics Canada software, CONFID, has been made available to U. S. Government agencies, where it is currently being implemented and evaluated. Complementary suppression software is applicable to tables of both frequency and magnitude.

In the straightforward implementation of linear programming, sensitive cells are treated sequentially beginning with the most sensitive. At each step (i.e. for each sensitive cell) the set of complementary cells which minimizes a cost function (usually the sum of the suppressed values) is identified. Zayatz (1992b) describes the formulation for two-dimensional tables. Zayatz (1992a) gives the parallel formulation for three-dimensional tables. As above, these are implemented by using a commercially available linear programming package. The disadvantage of the straightforward linear programming approach is the computer time it requires. For large problems, it is essentially impossible to use.

Another linear programming approach is based on describing the table structure as a mathematical network, and using that framework and the required tolerance intervals for each cell to balance the table (Cox, 1992). The network methods are favored because they give the same result as the straightforward linear programming methods, but the solution requires much less computer time.

The network method is directly applicable to two-dimensional tables (Cox and Ernst, 1982; Cox, 1987b) and to two-dimension}l tables with subtotal constraints in one dimension (Cox and George, 1989). Subtotal constraints occur when data in one dimension have a hierarchical additive structure. One common example of this structure occurs when one variable is the Standard Industrial Classification (SIC) code. An interior table cell might relate to a specific 4 digit SIC code, with subtotals given by 3-digit SIC codes, and the marginal total given by the appropriate 2-digit code. Sullivan (1992b) describes how to represent tables with this hierarchical structure in a network.

Complementary suppression and controlled rounding can both be solved using network theory. The parallelism between the two problems was demonstrated in Cox, Fagan, Greenberg and Hemmig (1986). Ernst (1989) demonstrated the impossibility of representing a general three or higher dimension table as a network. For this reason, complementary suppression for three-dimensional tables currently uses one of two approaches, (Zayatz, 1992a). The straightforward linear programming methods can be used for small three-dimensional tables. However, for large three-dimensional tables, an iterative approximate approach based on a sequence of two-dimensional networks is used. The complementary suppression pattern identified by this approximate approach must still be audited to assure that an individual sensitive cell cannot be estimated too accurately.

There is continuing research in developing faster and more efficient procedures for both two-dimensional and three-dimensional tables, (see Greenberg, 1986; Kelly, 1990; Kelly, Golden and Assad, 1990a and 1990b; Kumar, Golden and Assad, 1992; Desilets, Golden, Kumar and Wang, 1992; Lougee-Heimer, 1989; and Wang, Sun and Golden, 1991). Mathematical approaches mentioned in current research include methods based on integer programming, network flow theory, and neural networks.

As mentioned above, one possible objective function for automated procedures is to minimize the sum of the suppressed values. With this objective function, automated procedures tend to suppress many small cells, a result not generally considered "optimal" by the analyst. As observed by Cox (1992) "what data analysts want to see coming out of the complementary suppression process isn't always minimum number of suppressions and isn't always minimum value suppressed, but rather sometimes one and sometimes the other and, in general, a suppression pattern that somehow balances these two objectives to avoid worst-case scenarios."

Further research is needed into the identification of cost functions for use in selecting the "optimal" complementary suppressions. Possibilities here include both research into a cost function to be used for a single run of the software, as well as cost functions for use in multiple runs of the software. An example is development of a cost function to be used during a second pass through the software to remove superfluous suppressions. Rowe (1991) and Zayatz (1992b) provide examples of current research into cost functions.

Another reason the complementary cells selected by automated methods do not provide the "optimal" set for the table as a whole is that all current implementations protect sensitive cells sequentially. For any given sensitive cell, the complementary cells selected to protect it will be

optimal according to the objective function, conditional on all suppressions selected for previously considered sensitive cells. The sequential nature of the approach leads to over-suppression.

In spite of the lack of "optimality" of the result, the automated complementary cell suppression procedures identify useful sets of complementary suppressions. However, work is often needed to fine tune, reduce over-suppression, and assure that the analysts' nonmathematical definition of an "optimal" solution is more closely realized.

B.3. Information in Parameter Values

Agencies may publish their suppression rules, however, they should keep the parameter values they use confidential. Knowledge of both the rule and the parameter values enables the user to make better inferences concerning the value of suppressed cells, and may defeat the purpose of suppression.

For example, assume that an agency uses the p-percent rule with p=20 percent, and that the same value of p is used to determine the protection regions for complementary suppression. We assume that a cell total is 100 and that the cell is sensitive according to the p-percent rule. That cell will be suppressed along with other suitable complementary cells. For this cell (as with any suppressed cell), any user can use a linear programming package to calculate upper and lower bounds for the cell total based on the published row and column equations. Assume that this leads to the following inequality:

$$80 = \text{lower bound} \leq \text{cell total} \leq \text{upper bound} = 120.$$

In this case, the protection region used in selecting cells for complementary suppression assures that the cell total cannot be estimated more closely than plus or minus 20 percent of the cell value, or plus or minus 20 in this case. A knowledgable user has thus uniquely determined that the value of the suppressed cell total must be 100. Once the total for one suppressed cell has been uniquely determined, it is likely that other cell values can easily be derived by subtraction from published marginal totals.

C. Technical Notes: Relationships Between Common Linear Sensitivity Measures

This section illustrates the relationship between the p-percent, pq and (n,k) rules described in the text by using plots of regions of cell sensitivity. To simplify this presentation we make a few assumptions. First, for the p-percent rule we assume there are no coalitions (c=1) and for the (n,k) rules we consider only n=1 and n=2. Second, replace $\sum_{i=3}^{N} x_i$ by $(T - x_1 - x_2)$. Third, divide each sensitivity rule through by the cell total, T, and multiply by 100. Finally, set $z_i = 100x_i/T$, the percent contributed to the cell total by company i. The sensitivity rules can be written

$$S^{p\%}(X) = (1 + \frac{100}{p}) z_1 + \frac{100}{p} z_2 - \frac{100}{p} 100,$$

$$S^{pq}(X) = (1 + \frac{q}{p}) z_1 + \frac{q}{p} z_2 - \frac{q}{p} 100,$$

$$S^{(1,k_1)}(X) = (1 + \frac{k_1}{100 - k_1}) z_1 - \frac{k_1}{100 - k_1} 100$$

$$S^{(2,k_2)} = (1 + \frac{k_2}{100 - k_2}) z_1 + (1 + \frac{k_2}{100 - k_2}) z_2 - \frac{k_2}{100 - k_2} 100$$

The regions where these sensitivity rules are positive (i.e. where the cells are sensitive) are shown in Figure 1. The horizontal axis represents the percent contributed by the largest unit, z_1 and the vertical axis represents the percent contributed by the second largest unit, z_2. Since $z_1 \geq z_2$ and $z_1 + z_2 \leq 1$ (the sum of the two largest is less than or equal to the cell total), the only possible values in a table cell will be in the lower triangular region bounded from below by the line $z_2 = 0$, from above by the line $z_1 = z_2$ and to the right by the line $z_1 + z_2 = 1$.

The $(1,k_1)$ and $(2,k_2)$ rules are particularly simple to illustrate graphically. The inequality $(1, k_1)$ rule simplifies, and a cell is classified as sensitive if $z_1 > k_1$. The dividing line between sensitive and nonsensitive region is given by a vertical line through the point $(0,k_1)$ Similarly, the inequality for the $(2,k_2)$ rule simplifies and a cell is classified as sensitive if $(z_1 + z_2) > k_2$. The dividing line between the sensitive and nonsensitive regions is the line through the points $(0,k_2)$ and $(k_2,0)$. This line intersects $z_1=z_2$ at the point $(k_2/2,k_2/2)$. In all cases the sensitive region is the area to the right of the dividing line. The sensitivity regions for the (1,75) and (2,85) rules are illustrated in Figure 1A.

For the p-percent rule the inequality above yields the boundary line for sensitive cells as the line joining the points (0,100) and $(\frac{100}{\frac{p}{100} + 1}, 0)$. This line intersects $z_1=z_2$ at the point $(\frac{100}{\frac{p}{100} + 2}, \frac{100}{\frac{p}{100} + 2})$ The pq rule is the same, with q/p = 100/p.

FIGURE 1A
EXAMPLES OF SUPPRESSION REGIONS
THE (N,K) RULE WITH N=1 AND K=75, N=2 AND K=85

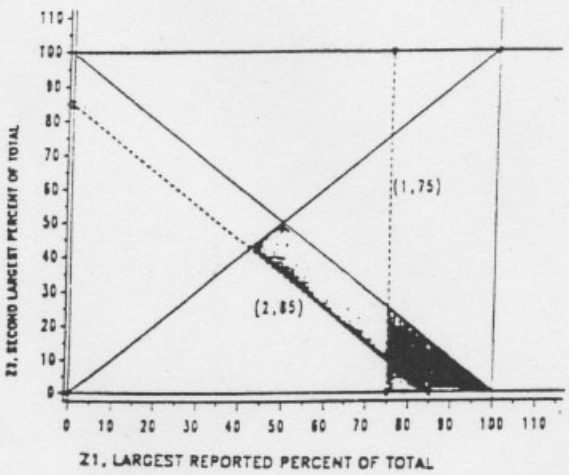


FIGURE 1B
EXAMPLES OF SUPPRESSION REGIONS
THE P-PERCENT RULE WITH P=17.65 PERCENT, AND P=35.3 PERCENT

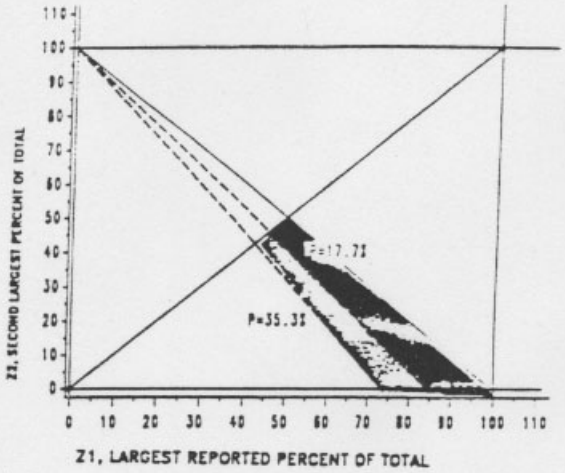


FIGURE 1C
P-PERCENT LESS CONSERVATIVE THAN (2,85)
P = 17.7 PERCENT

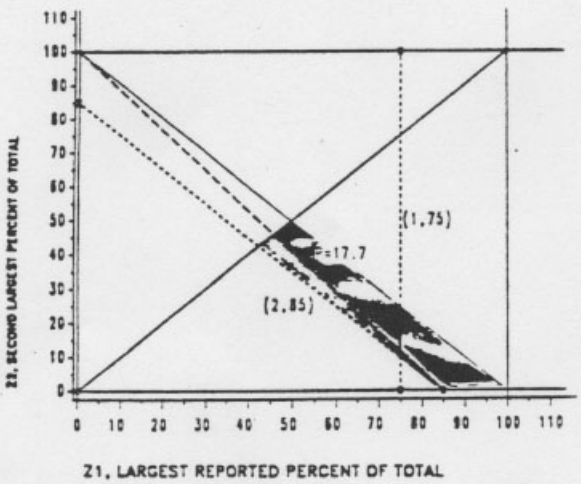
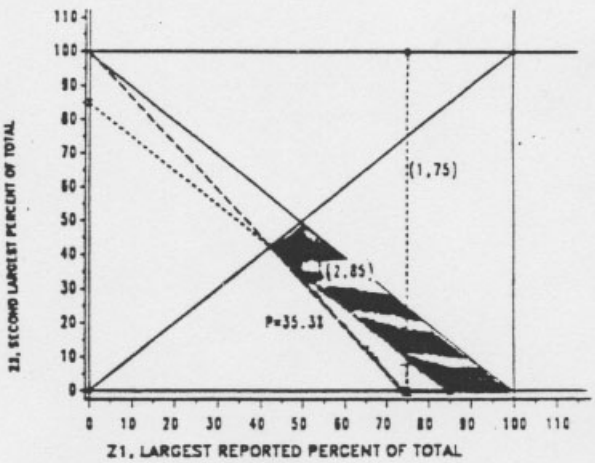


FIGURE 1D
P-PERCENT MORE CONSERVATIVE THAN (2,85) AND (1,75)
P = 35.3 PERCENT



Note: Values corresponding to cells in a table are in the triangle bounded by the lines z1 = z2, z2 = 0, and z1 + z2 = 1. Values which correspond to sensitive cells are shaded.

Figure 1B shows the sensitivity regions for the p-percent rule with p=17.65 and p=35.29. The selection of these values of p will be discussed below. Note that if p=0, the sensitivity line falls on top of the line $z_1 + z_2 = 1$. At that point there are no sensitive cells. Similarly if p is negative, there are no sensitive cells.

To Find p so that $S^{p\%}(X) \leq S^{(n,k)}(X)$ for all cells, X.

Consider the case where the (n,k) rule is being used and there is also a requirement that no respondent's contribution be estimable to within p-percent of its value. We would like to find the value of p so that the p-percent rule is closest to the (n,k) rule with $S^{(n,k)}(X) \geq S^{p\%}(X)$. Thus, there may be cells classified as sensitive by the (n,k) rule which would not be sensitive by the p-percent rule, but all cells classified as sensitive by the p-percent rule would be classified as sensitive by the (n,k) rule. Consider the (2,85) rule illustrated in Figure 1A. The p-percent rule, closest to the (2,85) rule, which would satisfy this requirement would be the one which intersects the line $z_2=0$ at the same point as the (2,85) rule. Thus, for a given value of k_2 we must have

$$\frac{p}{100} = \frac{100}{k_2} - 1.$$

Similarly, if we were first given the value of p for the p-percent rule, we must have

$$k_2 = \frac{100}{\frac{p}{100} + 1}.$$

For the (2, 85) rule, p/100 = 15/85 = .1765, so that p=17.65 percent. Figure 1C shows the (2,85) sensitivity region along with the less conservative p=17.65 percent region.

For the (1, k_1) rule, the p-percent rule closest to the (1,75) rule satisfying this requirement would be the one intersecting the line $z_1=z_2$ at the point (75,75). For a given value of k_1 we must have

$$\frac{p}{100} = \frac{100}{k_1} - 2.$$

Similarly, if we were first given the value of p,

$$k_1 = \frac{100}{\frac{p}{100} + 2}.$$

With $k_1 = 75$, the less conservative p-percent rule would have p=-66.7, which would result in no cell suppression. For p=17.65%, we would need $k_1 = 45.94$, a very restrictive rule.

To find parameter p so that $S^{p\%}(X) \geq S^{(n,k)}(X)$ for all X.

We would like to find the value of p so that the p-percent rule is closest to the (n,k) rule with $S^{(n,k)}(X) \leq S^{p\%}(X)$. Thus, there may be cells classified as sensitive by the p-percent rule which would not be sensitive by the (n,k) rule, but all cells classified as sensitive by the (n,k) rule would be classified as sensitive by the p-percent rule. Again, we consider the (2,85) rule as illustrated in Figure 1A.

In this case the most conservative p-percent rule needed would be the one that intersects the line $z_1=z_2$ at the same point as the (2, 85) rule. Given the value of k_2 this leads to

$$\frac{p}{100} = \frac{200}{k_2} - 2.$$

If we were first given the value of p, we would need

$$k_2 = \frac{200}{\frac{p}{100} + 2}.$$

For $k_2 = 85$, this gives p/100 = 200/85 - 2 = .3529. Figure 1D shows the (2,85) sensitivity region along with the p=35.29 percent region.

To find the most conservative p% rule needed to include the sensitivity region of the $(1,k_1)$ rule, we need the p-percent rule which intersects the line $z_2 = 0$ at the same point as the $(1,k_1)$ rule. Given the value of k_1, this leads to

$$\frac{p}{100} = \frac{100}{k_1} - 1.$$

If we were first given the value of p, we would need

$$k_1 = \frac{100}{\frac{p}{100} + 1}.$$

For the (1,75) rule, this leads to p/100 = 25/75 = .3333.

To find the $(1,k_1)$ rule going through the same point as the (2,85) rule and the p-percent rule with p=35.29%, substitute the desired value of p into the above equation and find $k_1 = 73.91$.

In this case since we started with the (2,85) rule, which lead to p = 35.29, a consistently less conservative $(1, k_1)$ rule is the one that has $k_1 = 73.91$. Thus the p-percent rule with p=35.29 provides slightly more protection than either the (2,85) rule or the (1,73.91) rule.

Table 1 in the text summarizes these results for selected values of p, or equivalently for selected values of q/p.

Example

Consider the three cells below. Let x_1^k represent the largest value reported by a respondent in cell k; x_2^k the second largest value reported by a respondent in cell k; and so on. Here we assume that respondents report in only one of the cells 1, 2 or 3. Cell membership is denoted by the superscript k. Superscript T represents the total.

	Cell 1	Cell 2	Cell 3	Total
	$x_1^1 = 100$	$x_1^2 = 1$	$x_1^3 = 100$	$x_1^T = 100$
		$x_2^2 = 1$		$x_2^T = 100$
		$x_3^2 = 1$		$x_3^T = 1$
		.		.
		.		.
		.		.
		$x_{20}^2 = 1$		$x_{22}^T = 1$
SUM	100	20	100	220

Assume that we are using the (n,k) rule with n=2 and k=85 percent. As described above, the related rules are the p-percent rule with p=17.65 (more conservative), the p-percent rule with p=35.29 (less conservative) and the (1,73.91) rule.

Using any of these rules, Cell 1 and Cell 3 are clearly sensitive (N=1, so S(X) > 0). It is also easy to verify that using any sensible rule Cell 2 is not sensitive. We consider two cells, the union of Cell 1 and Cell 2 and the Total.

The cell sensitivities for these rules are

$S^{(2,85)}$ (Cell 1 $\cup$ Cell 2) = 100 + 1 - 5.667*19 = -6.67
$S^{17.6\%}$ (Cell 1 $\cup$ Cell 2) = 100 - 5.667*19 = -7.67
$S^{(1,73.91)}$(Cell 1 $\cup$ Cell 2) = 100 - 2.833*20 = 43.34
$S^{35.29\%}$ (Cell 1 $\cup$ Cell 2) = 100 - 2.834*19 = 46.16

$S^{(2,85)}$ (Total) = 100 + 100 - 5.667*20 = 86.66
$S^{17.6\%}$ (Total) = 100 - 5.667*20 = -13.34
$S^{(1,73.91)}$ (Total) = 100 - 2.833*120 = -239.96
$S^{35.29\%}$ (Total) = 100 - 2.834*20 = 43.32

The union of Cell 1 and Cell 2 is not sensitive according to the (2,85) rule and the 17.65% rule. However, both the (1,75) and the 33.3% rule classify the cell as sensitive. Looking at the respondent level data, it is intuitively reasonable that the union of Cell 1 and Cell 2 is sensitive, even though the rule of choice for this example was to protect only against dominance by the 2 largest respondents. This cell corresponds to the point (83.3,.008) on Figure 1.

The Total is sensitive for the (2,85) rule and the p-percent rule with p=35.3%. It is not sensitive for the (1,73.9) rule or the p-percent rule with p=17.6%. This point corresponds with the point (45.5,.45.5) on Figure 1.

Consider the inconsistency in using the (2,85) rule alone. In the above example, if the union of cell 1 and cell 2 (not sensitive by the (2,85) rule,) is published, then the largest respondent knows that the other respondents' values sum to 20, and each of other respondents knows that the other respondents' values sum to 119. If the total (sensitive by the (2,85) rule) is published then the largest two respondents each knows that the sum of the remaining respondents' values is 120, and each of the small respondents knows that the sum of the others' values is 219.

Intuitively, it would seem that more information about respondent's data is released by publishing the nonsensitive union of cell 1 and cell 2 than by publishing the sensitive total. The inconsistency can be resolved by using a combination of (n,k) rules, such as the (1,73.91) and (2,85), or by using a single p-percent rule with p = 35.29 or a pq-rule with q/p = 2.83. These changes result in additional, but more consistent suppressions.

Proponents of the simple (2,85) rule claim that more protection is needed when respondents have competitors with values close to their own. Proponents of the simple (1, 75) rule claim that more protection is needed if the cell is dominated by a single respondent. These people argue that the use of a simple (n,k) rule allows them to determine which rules are needed for their special situations without the additional suppressions which would result from a more consistent approach.