PRESERVING CONFIDENTIALITY AND QUALITY OF TABULAR DATA:

ARE SAFE DATA NECESSARILY INFERIOR DATA?

Lawrence H. Cox, Associate Director
National Center for Health Statistics
LCOX@CDC.GOV

Bureau of Transportation Statistics Confidentiality Seminar
Washington, DC

September 17, 2003

PRESENTATION HANDOUT–DO NOT QUOTE OR CITE

Statistical Disclosure Limitation (SDL) for Tabular Data

Tabular data

* frequency (*count*) data organized in *contingency tables*
* *magnitude* data (income, sales, tonnage, # employees, ..) organized in sets of tables

Tables

* there can be *many*, many, many tables (national censuses)
* tables can be 1-, 2-, 3-, ...up to many *dimensions*
* tables can be *linked*
* table entries: *cells* (industry = retail shoe stores & location = Washington DC)
* data to be published: *cell values* (first quarter sales for shoe stores in Washington DC = $17M)

What is disclosure?

Count data: disclosure = small counts (1, 2, ...)
Magnitude data: disclosure = dominated cell value

Example:	Shoe company # 1:	$10M
	Shoe company # 2:	$ 6M
	<u>Other companies (total):</u>	<u>$ 1M</u>
	Cell value:	$17M

2 can subtract its contribution from cell value and infer contribution of #1 to within 10% of its true value = *DISCLOSURE*

Cells containing disclosure are called *sensitive cells*

How is disclosure in tabular data *limited* by statistical agencies?
- * identify cell values representing disclosure
- * determine *safe values* for these cells

Example: If estimation of any contribution to within 20% is safe (policy decision), then a safe value above would be $18M

* traditional methods for statistical disclosure limitation
  - Count data:
    - rounding
    - data perturbation
    - swapping/switching
    - cell suppression
  - Magnitude data:
    - cell suppression

What is *cell suppression*?
- * replace each disclosure-cell value by a symbol (*variable*)
- * replace selected other cell values by a symbol (*variable*) to prevent narrow estimates of disclosure-cell values
- * process is complete when resulting system of equations divulges no *unsafe estimates* of disclosure-cell values

Some properties of cell suppression:
- * based on mathematical programming
- * very complex theoretically, computationally, practically
- * destroys useful information
- * thwarts many analyses; favors sophisticated users

How does cell suppression addresses *data quality?*

Cell suppression employs a linear objective function to control *oversuppression*

Namely, the mathematical program is instructed to minimize:

* total value suppressed
* total percent value suppressed
* number of cells suppressed
* logarithmic function related to cell values
* etc.

These are overall (*global*) measures of data distortion

Further, individual cell *costs* or *capacities* can be set to control individual (*local*) distortion

These are all sensible criteria and worth doing

However, they do not preserve statistical properties (*moments*)

Moreover, *suppression destroys data and thwarts analysis*

Controlled Tabular Adjustment (CTA)

* new method for SDL in tabular data
* perturbative method–changes, does not eliminate, data
* alternative to complementary cell suppression
* attractive for *magnitude data* & applicable to count data

<u>Original CTA Method (Dandekar and Cox 2002)</u>

* identify sensitive tabulation cells
* replace each disclosure cell by a *safe value*–namely, move the cell value *down* or *up* until safety is reached
* use linear programming to adjust nonsensitive values in order to restore additivity (*rebalancing*)
* if second and third steps are performed simultaneously, a *mixed integer linear program* (MILP) results. MILP is extremely computationally demanding
* otherwise (most often), the down/up decision is made heuristically, followed by rebalancing via linear programming (LP). LP computes efficiently even for large problems

(Nearly) Actual Example of Magnitude Table with Disclosures

167	317	1284	587	4490	3981	2442	1150	70 (21)	**14488**
57(1)	1487	172	667	1006	327	1683	1138	46 (7)	**6583**
616	202	1899	1098	2172	3825	4372	300(40)	787	**15271**
0	36(10)	0	16(4)	0	0	65	0	140(40)	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1: 4x9 Table of Magnitude Data & Protection Limits for the 7 Disclosure Cells (red)

D	317	1284	D	4490	3981	2442	1150	D	**14488**
D	1487	172	667	1006	327	1679	D	D	**6583**
616	D	1899	1098	2172	3825	4371	D	787	**15271**
0	D	0	D	0	0	70	0	D	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1a: After Optimal Suppression: 11 Cells (*30%*) & 2759 Units (*7.5%*) Suppressed

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1683	1138	39	**6571**
617	196	1899	1095	2172	3825	4372	260	797	**15232**
0	26	0	12	0	0	65	0	180	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1b: After Controlled Tabular Adjustment

167	317	1284	587	4490	3981	2442	1150	70 (21)	**14488**
57(1)	1487	172	667	1006	327	1683	1138	46 (7)	**6583**
616	202	1899	1098	2172	3825	4372	300(40)	787	**15271**
0	36(10)	0	16(4)	0	0	65	0	140(40)	**257**
840	**2042**	**3355**	**2368**	**7668**	**8133**	**8562**	**2588**	**1043**	**36599**

Example 1: 4x9 Table of Magnitude Data & Protection Limits for the 7 Disclosure Cells (red)

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1679	1138	39	**6571**
617	196	1899	1095	2172	3825	4371	260	797	**15232**
0	26	0	12	0	0	70	0	180	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1b: Table After Controlled Tabular Adjustment

167	317	1276	587	4490	3981	2442	1150	91	**14501**
56	1487	172	667	1006	327	1683	1138	35	**6571**
617	202	1899	1098	2172	3825	4372	260	787	**15232**
0	20	0	9	0	0	65	0	194	**288**
840	**2026**	**3347**	**2361**	**7668**	**8133**	**8562**	**2548**	**1107**	**36592**

Example 1c: Table After Optimal Controlled Tabular Adjustment (Regression)

MILP for Controlled Tabular Adjustment (Cox 2000)

Original data: nx1 vector **a**
Adjusted data: nx1 vector $\boldsymbol{a} + \boldsymbol{y}^{+} - \boldsymbol{y}^{-}$
T denotes the coefficient matrix for the tabulation equations
Denote $\boldsymbol{y} = \boldsymbol{y}^{+} - \boldsymbol{y}^{-}$
Cells i = 1, ..., s are the *sensitive cells*
Upper (lower) *protection* for sensitive cell i denoted p_i $(-p_i)$

<u>MILP for case of minimizing sum of absolute adjustments</u>

$$\min \sum_{i=1}^{n} (y_i^{-} + y_i^{+})$$

Subject to:

$$\boldsymbol{T}\ (\boldsymbol{y}) = \boldsymbol{0}$$

$$y_i^{-} = p_i(1 - I_i) \qquad y_i^{+} = p_i I_i \qquad i = 1, \ldots, s \text{ (sensitive cells)}$$

$$0 \le y_i^{-},\ y_i^{+} \le e_i, \qquad i = s+1, \ldots, n \text{ (nonsensitive cells)}$$

$$I_i \text{ binary}, \qquad i = 1, \ldots, s$$

Capacities e_i on adjustments to nonsensitive cells typically small, e.g., based on measurement error

Data Quality Issues

Based on mathematical programming, just like cell suppression CTA can minimize:

* total value suppressed
* total percent value suppressed
* number of cells suppressed
* logarithmic function related to cell values
* etc.

In addition, adjustments to nonsensitive cells can be restricted to lie within *measurement error*

Still, this may not ensure good statistical outcomes, namely,

analyses on original vs adjusted data yield comparable results

Towards Ensuring Comparable Statistical Analyses

Verification of "comparable results" is mostly empirical
Many, many analyses are possible: Which analysis to choose?

Instead, we focus on preserving key statistics and linear models

* mean values
* variance
* correlation
* regression slope

between original and adjusted data

Can do this using direct (*Tabu*) search

I will describe ***how to do so well in most cases using LP***

For simplicity, assume that the down/up decisions for sensitive cells have already been made (by *heuristic*)

Preserving Mean Values

When the LP holds a total fixed, it *preserves the mean* of the
cell values contributing to the total
e.g., fixing the grand total preserves the overall mean

In general, to preserve a mean, introduce (new) constraint:
$\sum$ (adjustments to cells contributing to the mean) = 0

A criticism of CTA is that it introduces too much distortion into
the values of the sensitive cells

In general the intruder does not necessarily know which cells
are
sensitive nor cares to analyze only sensitive data, so
focusing on distortions to sensitive values may be a bit of
a red herring

Still, it is useful to demonstrate how to preserve the mean
of the sensitive cell values, as the method applies to
preserving the mean of any subset of cells

Preserving the mean of the sensitive cell values is equivalent to constraining net adjustment to zero:

$$\sum_{i=1}^{s} (y_i^{+} - y_i^{-}) = \sum_{i=1}^{s} y_i = 0$$

If, as in the original Dandekar-Cox implementation, we allow only two choices for y_i, this is unlikely to be feasible

However, satisfying this constraint is not a problem if we simply expand the set of possible y-values
viz., if we permit slightly larger down/up adjustments

The MILP is:

$$\min \ c(\boldsymbol{y})$$

Subject to:

$$\boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{0}$$

$$\sum_{i=1}^{s} (y_i^{+} - y_i^{-}) = 0$$

$$\left.\begin{array}{c} p_i(1 - I_i) \le y_i^{-} \le q_i(1 - I_i) \\ p_i I_i \le y_i^{+} \le q_i I_i \end{array}\right. \qquad i = 1, \ldots, s$$

$$0 \le y_i^{-}, \ y_i^{+} \le e_i \qquad i = s+1, \ldots, n$$

I_i binary, i = 1, ..., s

q_i are appropriate upper bounds on changes to sensitive cells
$c(\boldsymbol{y})$ is a linear cost function, typically involving sum of absolute adjustments

If the down/up directions are pre-selected, this is an LP

Preserving Variances

Seek: $Var(\boldsymbol{a} + \boldsymbol{y}) \doteq Var(\boldsymbol{a})$, assuming $\bar{y} = 0$

$$Var(\boldsymbol{a} + \boldsymbol{y}) = Var(\boldsymbol{a}) + 2Cov(\boldsymbol{a}, \boldsymbol{y}) + Var(\boldsymbol{y})$$

Define $L(\boldsymbol{y}) = Cov(\boldsymbol{a}, \boldsymbol{y})/Var(\boldsymbol{a}) = (1/(sVar(\boldsymbol{a})))\sum_{i=1}^{s}(a_i - \bar{a})y_i$

L(**y**) is a *linear function* of the adjustments **y**

$$Var(\boldsymbol{a} + \boldsymbol{y})/Var(\boldsymbol{a}) = 2L(\boldsymbol{y}) + (1 + Var(\boldsymbol{y})/Var(\boldsymbol{a}))$$

$$| Var(\boldsymbol{a} + \boldsymbol{y})/Var(\boldsymbol{a}) - 1 | = | 2L(\boldsymbol{y}) + (Var(\boldsymbol{y})/Var(\boldsymbol{a})) |$$

Var(**y**) is nonlinear, but can be linearly approximated

Alternatively: typically *Var(**y**)/Var(**a**) is small*
Thus, variance is approximately preserved by minimizing
$|L(\boldsymbol{y})|$

The absolute value is minimized as follows:

* incorporate two new linear constraints in the system:

$$w \geq L(\boldsymbol{y})$$
$$w \geq -L(\boldsymbol{y})$$

* minimize w

Assuring High Positive Correlation

Seek: $Corr(\boldsymbol{a},\ \boldsymbol{a} + \boldsymbol{y}) \doteq 1$

Corr (**a**, **a + y**) = $Cov(\boldsymbol{a},\ \boldsymbol{a} + \boldsymbol{y}) \div \sqrt{Var(\boldsymbol{a})\ Var(\boldsymbol{a} + \boldsymbol{y})}$

After some algebra,

Corr (**a**, **a + y**) = $(1 + L(\boldsymbol{y})) \div \sqrt{\ Var(\boldsymbol{a} + \boldsymbol{y})\ /\ Var(\boldsymbol{a})}$

Again: min $|L(\boldsymbol{y})|$ yields a good approximation because
it drives both numerator and denominator to one

Assuring Slope of Regression Line(s)

Seek: under ordinary least squares regression

$Y = \beta_1 X + \beta_0$

of adjusted data Y = **a + y** on original data X = **a**,

we want: $\beta_1 \doteq 1$ and $\beta_0 \doteq 0$

$$\beta_1 = Cov(\boldsymbol{a} + \boldsymbol{y}, \boldsymbol{a}) \;/\; Var(\boldsymbol{a}) = 1 + L(\boldsymbol{y}),$$

$$\beta_0 = (\bar{a} + \bar{y}) - \beta_1 \bar{a}$$

As $\bar{y} = 0$, then $\beta_0 \doteq 0$ if $\beta_1 \doteq 1$

This corresponds to $L(\boldsymbol{y}) \doteq 0$ (if feasible)

Note again: this is achieved via min $|L(\boldsymbol{y})|$

The Compromise Solution

Variance is preserved by minimizing L(**y**)
Correlation is preserved by minimizing L(**y**)
Regression slope preserved by $L(\boldsymbol{y}) \doteq 0$ (if feasible)
All subject to $\bar{y} = 0$

If Var(**y**)/Var(**a**) is small (typical case), imposing objective function $\min \, |L(\boldsymbol{y})|$ assures good results **simultaneously**

- for variance
- for correlation
- for regression slope

Shortcut is to incorporate the constraint L(**y**) = 0 (if feasible)

Choosing $L(\boldsymbol{y}) \doteq 0$ is motivated statistically because it implies (near) zero correlation between values **a** and adjustments **y** viz., as solutions **y** and **-y** are interchangeable, this correlation should be zero

Examples

4x9 Table									
Original	*Table*								
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	70000	**14490006**
56250	1487000	172500	667503	1006253	327500	1683000	1138250	46000	**6584256**
616752	202750	1899502	1098751	2172251	3825251	4372753	300000	787500	**15275510**
0	35000	0	16250	0	0	65000	0	140000	**256250**
840502	**2042251**	**3355753**	**2370005**	**7669255**	**8133752**	**8562754**	**2588250**	**1043500**	**36606022**
Protection	*(+/-)*								
0	0	0	0	0	0	0	0	21000	
625	0	0	0	0	0	0	0	7800	
0	0	0	0	0	0	0	40000	0	
0	10500	0	4875	0	0	0	0	42000	

Table 1: 4x9 Table of Magnitude Data and Protection Limits for Its Seven Sensitive Cells (in red)

min $\sum \lvert y_i \rvert$									
166875	307001	1283751	587501	4490751	3981001	2442001	1150000	91000	**14499881**
56875	1487000	172500	667503	1006253	327500	1683000	1141875	38200	**6580706**
616752	202750	1899502	1103626	2172251	3825251	4372753	260000	816300	**15269185**
0	45500	0	11375	0	0	65000	36375	98000	**256250**
840502	**2042251**	**3355753**	**2370005**	**7669255**	**8133752**	**8562754**	**2588250**	**1043500**	**36606022**
min |L-Bnd| (Variance)									
167500	317501	1283751	587501	4490751	3981001	2442001	1150000	91003	**14511009**
55625	1487000	172500	667503	1006253	327500	1683000	1146675	38200	**6584256**
616752	202750	1899502	1098751	2172251	3825251	4372753	260000	787498	**15235508**
0	18791	0	8125	0	0	65000	0	191756	**283672**
839877	**2026042**	**3355753**	**2361880**	**7669255**	**8133752**	**8562754**	**2556675**	**1108457**	**36614445**
max L (Corr.)									
167500	317501	1283751	587501	4490751	3981001	2442001	1129000	91000	**14490006**
55313	1499637	172500	667503	1006253	327500	1683000	1138250	34300	**6584256**
616752	202750	1899502	1098751	2172251	3825251	4372753	359884	787500	**15335394**
937	19250	0	8938	0	0	65000	0	94815	**188940**
840502	**2039138**	**3355753**	**2362693**	**7669255**	**8133752**	**8562754**	**2627134**	**1007615**	**36598596**
min |L| (Regress.)									
167500	317501	1276439	587501	4490751	3981001	2442001	1150000	91000	**14503694**
55625	1487000	172500	667503	1006253	327500	1683000	1138250	34420	**6572051**
616752	202750	1899502	1106063	2172251	3825251	4372753	260000	787500	**15242822**
0	19250	0	8938	0	0	65000	0	194267	**287455**
839877	**2026501**	**3348441**	**2370005**	**7669255**	**8133752**	**8562754**	**2548250**	**1107187**	**36606022**

Table 2: Original Table After Various Controlled Tabular Adjustments Using Linear Programming To Preserve Statistical Properties of Sensitive Cells Only

Results for 4x9 Table

Summary: 4x9 Table		**Linear**	**Programming**
Sensitive Cells	Corr.	Regress. Slope	New Var. / Original Var.
min $\lvert y_i \rvert$	0.98	0.82	0.70
min |L-Bound| (Var.)	0.95	0.93	0.94
max L (Cor.)	0.97	1.20	1.52
min |L| (Reg.)*	0.95	0.93	0.95
All Cells	Corr.	Regress. Slope	New Var. / Original Var.
All 4 Functions	1.00	1.00	1.00

Table 3: Summary of Results of Numeric Simulations on 4x9 Table Using Linear Programming

* = compromise solution

Results for 13x13x13 (Dandekar) Table

Summary: 13x13x13 Table	**Linear**		**Programming**
Sensitive Cells	Corr.	Regress. Slope	New Var. / Original Var.
min $\|y_i\|$	0.995	0.96	0.94
min |L-Bound| (Var.)	0.995	1.00	1.00
max L (Cor.)	0.995	1.00	1.21
min |L| (Reg.)*	0.995	1.00	1.01
All Cells			
All 4 Functions	1.00	1.00	1.00

Table 4: Summary of Results of Numeric Simulations on 13x13x13 Table Using Linear Programming

* = *compromise solution*

Concluding Comments

* statistical agencies have responsibilities
    - to respondents (to maintain confidentiality)
    - to data users (to deliver high-quality data products)

* these responsibilities
    - are often in opposition
    - nevertheless, are not mutually exclusive
    - have, in the past, been approached separately

* research indicates these responsibilities can be addressed
    - simultaneously
    - using systematic, computationally efficient methods