OVERVIEW OF STATISTICAL DISCLOSURE LIMITATION

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

DIMACS Working Group on Privacy/Confidentiality
of Health Data
DIMACS
Rutgers University, Piscataway NJ
December 10-12, 2003

WHAT IS
STATISTICAL DISCLOSURE?

WHY IS IT A PROBLEM?

* Qualitatively
* Quantitatively

WHAT CAN BE DONE
TO LIMIT STATISTICAL DISCLOSURE?

QUALITATIVE/POLICY ISSUES

What is confidentiality preservation?

* holding close information of a personal or proprietary nature pertaining to a respondent, and not revealing it (directly or indirectly) to an unauthorized third party

What is statistical confidentiality protection?

* preserving confidentiality in statistical data products

What is statistical disclosure?

* statistical disclosure occurs when the release of a data product enables a third party to learn more about a respondent than originally known (T. Dalenius)

Note: "*Respondent*" refers to direct providers of data (person, organization, business) and to "units of analysis" they represent (families, corporations, groups)

Is confidentiality important?
Why should the data provider preserve confidentiality?

* required by law, regulation or policy
* ethical obligation: the *social contract*
* practical considerations
    - data accuracy
    - data completeness
    - developing trust

How is confidentiality threatened by release of statistical data?

* overt or derived identification and disclosure of individual respondent data

* identification thru matching attributes to another data file, leading to disclosure of individual attributes

* associate large percentage of an identifiable group with a characteristic (*group disclosure*)

Must confidentiality preservation be absolute?
What is its relative importance?

* the balance issue: *right to privacy vs. need to know*

* absolute confidentiality preservation is impossible: releasing any data divulges something about each respondent

* technology limits what can be done
  - technology to limit disclosure
  - technology to cause disclosure

* in principle:
  - minimum disclosure protection and data quality and completeness standards are not incompatible
  - a joint optimum can be reached

* in practice:
  - the balancing process is iterative
  - incompatibilities are resolved in favor of preserving confidentiality

What factors affect statistical disclosure?

* factors affecting *likelihood* of disclosure

- number of variables
- level(s) of data aggregation or presentation
- accuracy/quality of data
- sampling rate(s)
- knowledge about survey participation
- distribution of characteristics
- time
- insider knowledge

* factors affecting the *risk* of disclosure

- likelihood of disclosure
- number of confidential variables
- sensitivity of confidential data
- time
- target of disclosure
    # targeted respondent
    # arbitrary respondent: *fishing expedition*
    # group disclosure
- existence/quality of matching files
- motivation/abilities of intruder
- cost to achieve disclosure
- ease to access/manipulate data

QUANTITATIVE/STATISTICAL ISSUES

Statistical Disclosure in Tabular Data: An Illustration

AGE CATEGORY	RACE CATEGORY						
	1	6	**4**	7	6	7	31
	6	7	6	5	7	**1**	32
	3	6	5	7	6	7	34
	6	7	6	6	7	6	38
	2	6	7	**2**	6	5	28
	18	32	28	27	32	26	163

Incidence of Death Related to a Specific Disease in a State

Releaser determines: *disclosure* occurs whenever a cell count is (or can be reliably inferred to be) between 1 - 4

This results in 6 primary disclosure cells (in **bold**)

Traditional *disclosure limitation methods*:

Rounding (base B = 5), perturbation, cell suppression

ROUNDING

Conventional Rounding

(round to nearest multiple of B = 5)

0	5	**5**	5	5	5	30 (25)
5	5	5	5	5	**0**	30 (25)
5	5	5	5	5	5	35 (30)
5	5	5	5	5	5	40 (30)
0	5	5	**0**	5	5	30 (20)
20 (15)	30 (25)	30 (25)	25 (20)	30 (25)	25 (20)	165 (130)

( ) = sum of rounded entries

Rounded table is **NOT** additive!!!

165 - 130 = **35** individuals are not accounted for!!!

Controlled Rounding

- round to an *adjacent multiple* of B = 5
- preserve additivity within the table
- multiples of B = 5 remain fixed

0	5	**5**	5	5	10	30
5	10	5	5	10	**0**	35
5	5	5	10	5	5	35
5	10	5	5	5	5	35
0	5	10	**0**	5	5	25
15	35	30	25	30	25	160

Many different Controlled Roundings are possible
This CR is *optimal* as it is close as possible to the original table
CR methodology for 2-D tables based on network optimization

Random (Unbiased) Controlled Rounding also possible

(Controlled) (Random) Perturbation is analogous

COMPLEMENTARY CELL SUPPRESSION

Suppressing only the disclosure cells

D	6	***D***	7	6	7	31
6	7	6	5	7	***D***	32
D	6	5	7	6	7	34
6	7	6	6	7	6	38
D	6	7	***D***	6	5	28
18	32	28	27	32	26	163

Suppression pattern is *inadequate* due to ability of attacker to reconstruct/estimate one or more suppressions using the row and column equations

Need *complementary cell suppression*, viz., suppress additional nondisclosure cells to thwart reconstruction or narrow estimation of *primary disclosure cells*

Heuristic complementary cell suppression

D_{11}	6	D_{13}	7	6	7	31
6	7	6	D_{24}	7	D_{26}	32
D_{31}	6	D_{33}	7	6	7	34
6	7	6	6	7	6	38
D_{51}	6	7	D_{54}	6	D_{56}	28
18	32	28	27	32	26	163

This does better and appears to adequately limit disclosure
However, $D_{51} = 2$:
Row 2 + Row 5 - Col 4 - Col 6 = 32 + 28 - 27 - 26 = 7:

$$7 = (D_{24} + D_{26} + 26) + (D_{51} + D_{54} + 19) - (D_{24} + D_{54} + 20) - (D_{26} + D_{56} + 20) = D_{51} + 5$$

Detecting such *structural insufficiency* usually requires mathematical programming, viz., subject to the row and column constraints, compute min $\{D_{51}\}$ and max $\{D_{51}\}$

A better suppression pattern

D	6	***D***	7	6	7	31
6	7	***D***	5	7	***D***	32
D	6	5	***D***	6	7	34
6	7	6	6	7	6	38
D	6	7	***D***	6	***D***	28
18	32	28	27	32	26	163

Mathematically, this pattern is equivalent to

D_{11}	D_{13}	0	0	5
0	D_{23}	0	D_{26}	7
D_{31}	0	D_{34}	0	10
D_{51}	0	D_{54}	D_{56}	9
6	10	9	6	31

This pattern has some desirable features:

- not structurally insufficient
- minimum possible number of cells suppressed
- minimum possible total value suppressed

This pattern does not appear inadequate:

- at least two suppressions in each row/column
- reduced row/col equations add to at least 5

However, appearances can be deceiving

Suppression Audit

Linear analysis reveals *exact bounds* for suppressed entries:

[0,2]	6	**[3,5]**	7	6	7	31
6	7	[5,7]	5	7	**[0,2]**	32
[1,5]	6	5	[5,9]	6	7	34
6	7	6	6	7	6	38
[0,5]	6	7	**[0,4]**	6	[4,6]	28
18	32	28	27	32	26	163

A suppression pattern is *adequate* (passes audit), if the interval for each disclosure cell contains the open interval **(0,5)**

This suppression pattern *fails the audit* for **3** cells

Detecting such *numerical insufficiency* requires mathematical programming or other algorithms and software, implemented knowledgeably

Could publish audit bounds in lieu of "**D**"

An adequate suppression pattern

[0,5]	6	**[0,5]**	7	6	7	31
6	7	6	[0,6]	7	**[0,6]**	32
[0,6]	6	[2,8]	7	6	7	34
6	7	6	6	7	6	38
[0,6]	6	[4,10]	**[1,7]**	6	[0,6]	28
18	32	28	27	32	26	163

Mathematically, this pattern is equivalent to

D_{11}	D_{13}	0	0	5
0	0	D_{24}	D_{26}	6
D_{31}	D_{34}	0	0	8
D_{51}	D_{53}	D_{54}	D_{56}	16
6	16	7	6	35

CONTROLLED TABULAR ADJUSTMENT

Complementary cell suppression:

- an *NP hard problem*: difficult theoretically and practically
- produces "tables with holes"
- thwarts statistical analysis

An alternative method (to be discussed Friday) called
controlled tabular adjustment

- produces a full and fully analyzable table(s)
- is close to the original table(s)
    * locally (cell by cell)
    * globally (minimizes a measure of overall distortion)
- preserves important statistical properties of the table(s)

Controlled Tabular Adjustment: Example

Original table:

RACE CATEGORY

AGE CATEGORY							
	1	6	**4**	7	6	7	31
	6	7	6	5	7	**1**	32
	3	6	5	7	6	7	34
	6	7	6	6	7	6	38
	2	6	7	**2**	6	5	28
	18	32	28	27	32	26	163

Incidence of Death Related to a Specific Disease in a State

Adjusted table:

AGE CATEGORY	RACE CATEGORY						
	0	6	5	6	6	8	31
	7	7	6	5	7	0	32
	5	6	5	5	6	7	34
	6	7	6	6	7	6	38
	0	6	6	5	6	5	28
	18	32	28	27	32	26	163

Incidence of Death Related to a Specific Disease in a State

This solution minimizes sum of absolute adjustments subject to preserving marginal totals

Various other optimization criteria are available, leading to other solutions

For example:

If in addition adjustments to the 24 nondisclosure cells are limited to a maximum of 1 unit, then an optimal adjusted table is:

RACE CATEGORY

AGE CATEGORY							
	0	6	5	6	6	8	31
	7	7	6	5	7	0	32
	5	6	5	6	5	7	34
	6	7	6	5	8	6	38
	0	6	6	5	6	5	28
	18	32	28	27	32	26	163

Incidence of Death Related to a Specific Disease in a State

Statistical Disclosure in Microdata: An Illustration

Public Use Microdata (PUM) File from a Survey of Schools
All students grades 8-12 from sampled schools are interviewed

Age	Sex	Edu.	Alcohol Use	Drug Use	Sexually Active
14	*F*	8	*Y*	*N*	*Y*
14	*F*	9	*Y*	*N*	*N*
14	*M*	9	*Y*	*Y*	*N*
14	*M*	9	*Y*	*N*	*N*
15	*F*	10	*N*	*N*	*Y*
15	*M*	10	*Y*	*N*	*Y*
15	*M*	10	*Y*	*Y*	*Y*
16	*F*	10	*N*	*N*	*Y*
16	*F*	11	*Y*	*N*	*N*
16	*F*	11	*N*	*Y*	*Y*

Q: What can an outsider (PUM user) infer about individuals?
A: Nothing.

Q: What can the school or a parent infer about individuals?
A: 14F8 alc + sex; 14F9 alc; 15F10 sex; 16F10 sex

Q: What more can a student infer about another student?
A: 14M9, 15M10, 16F11 know all about counterpart

What techniques are available to limit statistical disclosure in microdata?

* *restrict* data dissemination

* *sample* the data
  - population file is drawn from a sample survey
  - *subsample* the population file

* abbreviate the data
  - remove direct identifiers
  - reduce the number of variables
  - remove *salient* records and/or
    records from salient respondents
  - *suppress* item detail
  - *topcode* sensitive items

* aggregate the data
  - *collapse* geographic identifiers
  - collapse data categories

* *switch* data: 1990 U.S. Decennial Census

* multiple methods: 2000 U.S. Decennial Census

What administrative procedures are available?

* remove the problem: respondent *waivers*

* anticipate: microdata *checklists*

* limit data dissemination
    - restricted access
    - restricted use
    - *encrypted* microdata
    - statistical data base query systems

* data abbreviation
    - eliminate variables from the released data file
    - eliminate respondents from the released data file
        # eliminate high risk records
        # release a sample
    - suppress selected item detail
    - *truncate* distributions: top (or bottom) code items l
    - release different file *extracts* to different data users

Disclosure limitation techniques (cont.)

* data aggregation or grouping
    - coarsen data
        # collapse data categories/detail
        # replace continuous data by categories
    - *microaverage* responses
    - release data summaries
        # tabulations
        # regression equations
        # variance/covariance matrices

* data modification
    - *round* item data (random or controlled)
    - *perturb* item data (random or controlled)
    - replace item data by *imputations*

* data fabrication
    - statistical matching
    - data *swapping*
    - data switching

New approaches to disclosure limitation in microdata

* *supersample* the data file
    - sample the (population) data file with replacement
    - reweight the new file
    - release or subsample the new file

* data fabrication / *synthetic data*

* statistical *data base query* systems
    - static
    - dynamic

* use of *contextual data*

* alternative forms of data release
    - *interval* data
    - maps and graphics

* combined use of respondent waivers and
    data user non-disclosure agreements

* probability based measures of disclosure risk combined with information based measures of data utility

EMERGING AREAS

Statistical data base query systems

Spatial data/models

Statistical maps

Releasing models in lieu of data