Methods of Confidentiality Protection

Arnold P. Reznek
U.S. Census Bureau
CES Room 2K128F
Washington, DC 20233
301-763-1856
Fax 301-763-5935
Arnold.phillip.reznek@census.gov

April 19, 2007

Plan of talk

- Define Confidentiality, Disclosure, and set the context.
- Very General principles
- Tables
- Microdata
- Models

Confidentiality

- Confidentiality
  - Def. 1: Release (“disclose”) data so that
    - Dissemination of data about respondent is not harmful to respondent
    - Immune from legal process
  - Def. 2: Treat info a respondent provides
    - In a “relationship of trust”
    - So it will only be used in ways consistent with original purpose.

Confidentiality and Disclosure

- Statistical disclosure – when released statistical data (tabular or individual records) reveal confidential information about an individual respondent.
- Types of Disclosure
  - Identity – respondent is identified from released file
  - Attribute - sensitive information about a respondent is revealed
  - Inferential - released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise

Legal Requirements (U.S.)

- Census Bureau
  - Title 13, U.S. Code
- Other U.S. Statistical Agencies –
  - See Tech. Paper 22 (FCSM 2005) Chap. 3.
- Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002
  - Information collected or acquired for exclusively statistical purposes under a pledge of confidentiality

The Balancing Act

- Publish as much valuable statistical information as possible without violating the confidentiality of respondents
- Preserve data utility while avoiding disclosure
- Conflicting forces – driven by computer revolution
  - Increasing disclosure risks from agencies' traditional data outputs (Sweeney 2001).
  - Increased value of research with data.

General Principles of Disclosure Review and Control

- Traditional methods
  - Suppression
  - Coarsening
  - Adding noise (explicit or via swapping)
- Types of data
  - Business
  - Household
- New methods (Abowd lecture)
  - Partially-synthetic data
  - Fully-synthetic data

Tables of Magnitude Data

- Disclosure occurs if user can estimate a respondent's value "too closely."
- Risk is at firm (enterprise) level
- Linear sensitivity measures (Cox 2001)
  - p%, (n,k), other rules
- Complementary Suppressions

Complementary Suppressions

- Other cells to "protect" primary suppressions
- This is Tricky
  - Example - Jewett (1992) p. I-D-4.

	Total	Col 1	Col 2	Col 3	Col 4
Total	510	100	100	160	150
Row 1	155	25	A	40	B
Row 2	125	E	20	F	30
Row 3	150	30	C	K	D
Row 4	80	G	10	H	20

- Cell K is unprotected – can "back out" exact value!

Complementary suppressions (cont'd)

- Linear Programming-based methods
- Fischetti-Salazar (2005) advances
- But – too much info lost!
  - E.g. ~ 1.5 million cells suppressed in County Business Patterns (Isserman and Westervelt 2006)
- Community is exploring alternatives

Alternatives to Cell Suppression: Adding Noise

Adding Noise

- Each unit's data perturbed up or down randomly by number "close to" a (confidential) percentage.
- Estabs of a single firm are perturbed in the same direction.
- Overall distribution of perturbations symmetric about 1.
- Sensitive cells perturbed most.
- All cells can be shown.

Adding Noise (Cont'd)

U.S. Census Bureau using noise in Quarterly Workforce Indicators (QWI)

- Longitudinal Employer Household Dynamics (LEHD) program
- Local Employment Dynamics (LED) partnership with states
- Employment, layoffs, hires, separations, job creations & destructions, earnings, etc.

Adding noise (cont'd)

SDC for QWI

- Permanent multiplicative noise factor for each record
- All statistics distorted
- Analytically valid statistics – unbiased, time series properties preserved
- Some cells still suppressed
- Now working on replacing suppressions with synthetic data

Adding Noise – (cont'd)

Census Bureau considering noise in County Business Patterns, Economic Census

Noise addition could affect Census RDCs

- Question: should research output use "real" or "noisy" microdata?
- Noise addition should decrease inquiries about "tabular" projects.

Microdata (in U.S., household)

- In U.S., almost always household

A few typical records:

Age	Race	Sex	Marital S.	Occup.	State
26	B	M	Married	Doctor	DE
54	W	F	Divorced	Teacher	MD
40	A	M	Widowed	Lawyer	FL

Practices for Microdata

- Removal of variables
- Subsampling
- Geographic thresholds
- Rounding
- Combining categories
- Changing continuous variables into categorical variables
- Noise addition

Practices for Microdata

- Thresholds on categories
- Data swapping
- Top and bottom coding
- Synthetic data

Typical Demographic Table

	Married	Divorced	Other	Total
0-18	2	1	0	3
19-65	45	39	20	104
66+	15	10	6	31
Total	62	50	26	138

Practices for Frequency Count Data

Data swapping

Aggregates, means, medians, other quantiles

Universe thresholds

Traditional rounding

Group quarters data

Profiles (300 people and check for slivers)

SDC for Regressions, and Model Servers

SDC for statistical model output becoming more important

- RDCs
- Model servers

SDC in Regressions

Some models produce (parts of) tables

- Regressions, logits, probits with only fully-interacted right-hand side dummy variables
- Correlation and covariance matrices that include dummies
- Regressions run on “nested” samples
  - Can reveal means of dependent variable via “subtraction”

Linear Regression

Models with Only Discrete Regressors: Regressions Can Produce Tables

- Example: Model with two dummy vars, with and without interactions, and adding a continuous variable: (CPS data from Berndt (1991) Chap. 5
  - Lwage = natural log of wage
  - Colplus = 1 for persons with a college degree or above, 0 otherwise
  - fe = 1 if female
  - colfe = colplus*fe, - interaction, i.e., females with a college education.
  - Ex = experience (age – education –6);
  - Exsq = square of experience

Linear Regression (cont'd)

Table 2: Regressions with Dummies			
A. Mean of lnwage by education and sex			
Education	Sex	N	Mean
No College	Male	220	2.0829
College	Male	69	2.4279
No College	Female	188	1.8298
College	Female	57	2.2777
College	All	126	2.3600
No College	All	408	1.9663
All	Male	245	2.1653
Alle	Female	289	1.9340
All	All	534	2.0592

B. Regression Model Estimates			
Coefficient	Model (1)	Model (2)	Model (3)
Constant (a)	2.0717	2.0829	1.7168
Colplus (b)	0.3918	0.3450	0.3524
Fe (c)	-0.2289	-0.2531	-0.2784
Colfe (d)		0.1029	0.1362
Ex			0.0403
Exsq			-0.0007

Results:

- (1) Without interaction, dummies not = means
- (2) With interaction they do:
  - a = mean (log) income of males without college
  - a+b = mean income of males with college
  - a+c = mean income of females without college
  - a+b+c+d =income of females with college.
- (3) With interaction and dummies, not so

SDC in Regressions (cont'd)

Other Risks in Regression Models (NISS project researchers)

- Indicator for single enterprise
- Use of transformations to generate outliers that pull regression line close to single observation
- Regression diagnostics – synthetic diagnostics

Regressions produced by data from several agencies while maintaining confidentiality

Model Servers

How they work

- Users do not have access to microdata
- Users specify models remotely
- System returns disclosure-free output

SDC

- All the considerations discussed above under models
- System needs to prevent “attacks”

Several countries and agencies developing model servers

References

Cox, L. (2001). "Disclosure Risk for Tabular Economic Data." Chapter 8 in Doyle et al. (2001).

Cox, L. (2002). "Confidentiality Issues For Statistical Database Query Systems." Invited Paper for Joint UNECE/Eurostat Seminar on Integrated Statistical Information Systems and Related Matters (ISIS 2002). (17-19 April 2002, Geneva, Switzerland). Available at http://www.unece.org/stats/documents/ces/sem.47/15.s.e.pdf.

Doyle, P., J.I. Lane, J.J.M. Theeuwes, and L.V. Zayatz (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: Elsevier Science B.V.

Federal Committee on Statistical Methodology (FCSM, 2005). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, DC, U.S. Office of Management and Budget. Revision of 1994 edition; available at http://www.fcsm.gov/working-papers/spwp22.html.

Fischetti, M. and J.J. Salazar (2005). "A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods." *Operations Research*, Vol. 53, No. 5, pp. 819-829.

Giessing, S. (2001). "Nonperturbative Disclosure Control Methods for Tabular Data." Chapter 9 in Doyle et al. (2001).

Isserman, A.M. and Westervelt, J. (2006). "1.5 Million Missing Numbers: Overcoming Employment Suppression in *County Busness Patterns* Data." *International Regional Science Review* 20, 2: pp. 311-335.

Isserman, A.M. and Westervelt, J. (2006). "1.5 Million Missing Numbers: Overcoming Employment Suppression in *County Busness Patterns* Data." *International Regional Science Review* 20, 2: pp. 311-335.

Reiter, J. (2003). "Model Diagnostics for Remote Access Regression Servers." *Statistics and Computing,* 13, pp. 371-380.

References (cont'd)

Sweeney, L. (2001). "Information Explosion." In Doyle et al. (2001), Chapter 3.

Willenborg, L. and T. de Waal (2001). *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics, volume 155. New York: Springer-Verlag.

Willenborg, L. and T. de Waal (1996). *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, volume 111. New York: Springer-Verlag.