Discussion of "Statistical Disclosure Limitation: Releasing Useful Data for Statistical Analysis"

Nancy J. Kirkendall
Energy Information Administration
April 28, 2003

BTS Confidentiality Seminar Series, April 2003

A Subtle Difference

- Steve says "Statistical disclosure limitation needs to assess tradeoff between preserving confidentiality and usefulness of released data."
- I would phrase it differently. Statistical agencies are required to preserve confidentiality, and within that constraint must make released data as useful as possible.

Basic Agreement

- We need better approaches to providing more useful information while protecting confidentiality.

Count vs. Magnitude Data

- Steve stresses the importance of using methods based on likelihood function.
  - He uses count data.
  - Distributional theory for count data in tables well established
- Most EIA data are magnitude data.
  - Data may follow any distribution, with skew distributions most common.
  - Not obvious how to base general methods on the likelihood function

Count vs. Magnitude Data (continued)

- Steve claims that using LP and IP approaches to finding bounds is NP hard -- for count data.
- For magnitude data finding optimal set of complementary suppressions in 3 or more dimensions is NP hard. Finding bounds is possible, and software is available.

Software to Compute Bounds

- Up to 3D has been available for decades. (CONFID, Census)
- More than 3D since ’95 (ACS), since ’01 (DAS)
- If table adds, bounds are computed.
- If table does not add, two approaches
  - Make minor adjustments to make the table add. Then compute bounds (ACS, CONFID, Census)
  - If the table does not add because of rounding, explicitly account for constraints due to the rounding process (DAS)

Teaching Survey Staff to Use Confidentiality Software

- Difficult for people to understand table dimensionality
- We need
  - A tutorial to teach people how to translate tables in pubs into the mathematical structure of SDL for input into software
  - User friendly interface to do it automatically

Releasing Useful Data

- I will use Steve's example 2 to compare information released via
  - Steve's method
  - Suppression
  - Controlled tabular adjustment
- Example based on theory that low cell count = sensitive

Example 2, with 6 variables (ABCDEF)

- Steve determines that he can release the margins ADE, ABCE, and BF. (And nothing else.) Bounds indicate no confidentiality concern.

- However, he is releasing only 15% of all possible cells.

Comparison of Amounts of Data Released

- Of the 2^6 = 64 interior cells, there are a total of 3^6 =729 cells (including all marginal totals).
- Steve releases 105 ($3^2+3^3+3^4-3^2-3^1$) cells. So 105/729=14.4% of data are released.
- Cell suppression, thanks to Ramesh Dandekar
  - 9 sensitive cells (6 interior and 3 marginal totals using n =3 or less as sensitive)
  - 103 complementary suppressions
  - **"Swiss cheese"** approach releases (729-103-9)/729=84.6% of data

Comparison of Amounts of Data Released (continued)

- Ramesh also applied his controlled tabular adjustment.
  - Adds or subtracts something from sensitive cells to protect
  - Adjusts other cells to balance the table
  - Result is release of counts for 100% of the cells
- The challenge is to make sure inferences are preserved.

How to Assure Inferences are Preserved?

- Ramesh regularly provides a histogram showing the distribution of percentage changes made to cells
  - This documents changes made.
- Research needed to define an appropriate set of statistical tests
  - To document the impact of changes on statistical analysis

Changing Data to Protect Confidentiality

- Not everyone thinks it is a good idea.
- Some users do not trust the result.
- When Ruben proposed simulating microdata in 1993 the users were aghast – they wanted the data.
- How to convince users the adjusted data are as good for inferences as the original?
- How to convince respondents that SDL has been applied?

However

- The sensitive cells in establishment data are frequently the small ones.
  - High percent change to sensitive cells – is this worse than "W"?
  - Small changes to big cells might be viewed as using different bases for rounding. Might be able to sell this.
- In some situations market dominated by giants – e.g., Large civil US Airliner Manufacturers.
  - Not sure there is much that can be done if there is one giant in a cell

Tables versus Query System

- Challenges in confidentiality not the same
  - Comparisons not really fair
- Current approaches
  - Protect microdata. Then any tabulations are OK.
  - Apply confidentiality protection to tables. Any data not suppressed can be released.
  - NISS is trying to do something different.

In Addition to Research on Methods, We Need

- Comparisons of SDL methods on the same data sets, to facilitate real comparisons
  - Ramesh has provided 8 simulated data sets.
- Agreement on standard measures for comparison
- Research to define a standard set of statistical tests to determine whether two tables provide same (multivariate) inferences
- Development of documentation for the public describing changes without allowing "intruder" to undo protection

Now for A Different Spin
"What is Sensitive?"

(thanks to Gordon Sande for this example)

	Total	Tax Cheat	Does Own Taxes	Uses Tax Service
Total	4500	1561	1719	1220
Head waiter	1000	960	20	20
Tinker	2000	500	1400	100
Tailor	1000	100	100	800
Lawyer	500	1	199	300

Sources

- Ramesh Dandekar, EIA – work using Example 2. Research on controlled adjustment or synthetic tabular adjustment, simulated data
- Gordon Sande, Sande and Associates, Inc– general insights, use of rounding to protect data, software, last example
- Tore Delanius and Ivan Fellegi – work in the 1970's – did the initial work on the danger of "association" in tables.