STATISTICAL ANALYSIS OF SUPPRESSED TABULAR DATA

Lawrence H. Cox
National Institute of Statistical Sciences
cox@niss.org

2013 UNECE/Eurostat work session on statistical data confidentiality
Ottawa, Ontario
October 29, 2013

The Problem

- since the 1940s, statistical offices have employed *cell suppression* for statistical disclosure limitation (SDL) of tabular data
- statistical analysis of tables in the presence of suppressions is difficult, esp. for less sophisticated analysts
- statistical imputation of suppressed cells is tricky as missing-ness is deterministic, not probabilistic
- most analytical methods for tables require complete tables

Potential Solution

- *deconstruct* suppression pattern to identify alternative tables suitable as surrogates for analysis of the original table
- perform analysis(es) on surrogate(s)

Alternative Table

- feasible table in respect to suppression pattern

Table Reconstruction

Frequency count data

- invoke iterative proportional fitting
- fit a log-linear model: MLE = surrogate
- new, direct approach
- construct a set of *alternative tables* based on algebraic *moves* from the original table, together with associated probabilities
- perform analysis on a surrogate table, or on a probability sample of tables and combine the analyses

Magnitude data (establishment-based)

- problem far less studied
- many users "analyze" only individual or collections of cell values, e.g., within a specific industry
- analysis compromised if these suppressed

Table Deconstruction

- identify or estimate sets of feasible values for suppressed entries
- identify alternative tables to the original table
- rule out/reject some alternative tables or values based on prior information, deterministic analysis, or probabilistic analysis
- identify alternative tables (expected to be) exchangeable with the original table for inferential purposes

Intruder and reconstruction

- reconstructs suppressed cells to obtain precise estimates of individual contributions
- employs primarily deterministic methods

Analyst and deconstruction

- deconstructs suppression pattern to identify reliable/realistic alternative tables
- accepts those tables exchangeable for analysis
- rejects non-exchangeable tables

Transparency Issues

- transparent SDL would involve revealing
  - # disclosure rule
  - # aspects of suppression rule/algorithm
- which can
  - # reduce number of alternative tables
  - # reduce number of alternative values
  - # thereby erode/threaten data security

Mathematical basis for CCS: Circuits

Example

D_{11} (1)	18	D_{13} (6)	**25**
13	D_{22} (5)	**D_{23} (2)**	**20**
D_{31} (4)	**D_{32} (1)**	10	**15**
18	**24**	**18**	**60**

Table: 4 sensitive (bold), 6 suppressed cells

- true values of suppressions in parentheses
- **disclosure rule**: sensitive cell = 1, 2, 3, or 4
- **suppression rule**:
  - **#** minimize number of cells suppressed (or total value suppressed)
  - # preserve zero-cells (optional)
- 4 sensitive (**bold**), 2 complementary cells
- this pattern optimal wrt. both number of cells (6) and total value (19) suppressed

Circuits

+/-	0	-/+
0	**-/+**	**+/-**
-/+	+/-	0

Interpretation

- 0-4 units can be *moved* in the + direction thru D_{11}
- 1 unit can be moved in the – direction
- 6 alternative values for D_{11}: D_{11} = 0,1,.., 5
- 6 alternative tables (including original)

D_{11} (1)	18	D_{13} (6)	**25**
13	D_{22} (5)	**D_{23} (2)**	**20**
D_{31} (4)	**D_{32} (1)**	10	**15**
18	**24**	**18**	**60**

D_{11} = 1 (Original table and released pattern)

D_{11} (3)	18	D_{13} (4)	**25**
13	D_{22} (3)	**D_{23} (4)**	**20**
D_{31} (2)	**D_{32} (3)**	10	**15**
18	**24**	**18**	**60**

D_{11} = 3 (Alternative table and pattern)

- all suppressions in D_{11} = 3 table are sensitive
- D_{11} = 1 and D_{11} = 3 optimal patterns identical
- analyst cannot rule out D_{11} = 3 as true table

D_{11} (1)	18	D_{13} (6)	**25**
13	D_{22} (5)	**D_{23} (2)**	**20**
D_{31} (4)	**D_{32} (1)**	10	**15**
18	**24**	**18**	**60**

D_{11} = 1 (Original table and released pattern)

0	18	7	**25**
D_{21} (13)	6	**D_{23} (1)**	**20**
D_{31} (5)	0	**D_{33} (10)**	**15**
18	**24**	**18**	**60**

D_{11} = 0 (Alternative table and pattern)

- only 1 sensitive cell in alternative table
- only 4 suppressions in optimal pattern
- D_{11} = 0 pattern differs from released pattern
- if analyst knows suppression rule (transparency issue), can apply rule to this alternative table and *rule out* D_{11} = 0 table
- can evaluate all alternative tables similarly

Table deconstruction and analysis of suppressed tables

Example: multiply previous table by 100

D_{11} (100)	1800	D_{13} (600)	**2500**
1300	D_{22} (500)	**D_{23} (200)**	**2000**
D_{31} (400)	**D_{32} (100)**	100	**1500**
1800	**2400**	**1800**	**6000**

Magnitude table: 4 sensitive, 6 suppressed cells

- conditional chi-square statistic compares alternative $\{c_i\}$ and original $\{a_i\}$ table values
- $\chi^2_{(df)} = \sum_i \frac{(c_i - a_i)^2}{a_i}$ *df* =degrees of freedom
- indices *i* restricted to suppressed entries
- if, as here, suppression pattern consists of a single circuit, then df = 1
- this corresponds to an integer quantity *d* that can be moved around the circuit without violating nonnegativity: $c_i = a_i \pm d$
- here, $-100 \leq d \leq 400$ (501 possible values)

$$\chi^2_{(1)} = \sum_i \frac{d^2}{a_i} = d^2 \sum_i \frac{1}{a_i}$$

In this example

- sum of reciprocals = 0.032
- for $\alpha = 0.05$, chi-square critical value = 3.84
- for $d^2 > 120$, χ^2 -statistic exceeds critical
- alternative tables with $|d| \geq 12$ are **not** reliable surrogates for original table
- 23 (not 501) surrogates: d = -11, ..., 0, ..., 11
- $89 \leq D_{11} \leq 111$

Note

- analyst and NSO can compute *d*
- NSO knows true table on which set of reliable alternative tables is centered
- analyst does not know true table
- analyst must home-in on true table

Analysis

- select a sample surrogate table and analyze it
- draw a sample of (or all) surrogates, analyze, and from distribution of analytical outcomes produce a representative analytical outcome or a combined outcome