Balancing respondent confidentiality and data user needs

Consumer Expenditure Survey

Arcenis Rojas
What is the crux?

- Conflicting goals
  - Maximize data access
  - Protect respondents identity
Why is confidentiality important?

- Ensure future cooperation by respondents
- It’s the law
Title 13?

Federal law to protect identities of survey respondents
Who determines threats?

- Disclosure Review Board (DRB) by the U.S. Census
How could microdata reveal respondents’ identity?

- High income
- High expenditures
- High age
- Small PSUs
How to protect respondents’ confidentiality?

Conceal information that *could* reveal respondents
How to protect respondents’ confidentiality?

Two stages:

- Census removes *obvious* identifiers
- BLS suppresses *data related* identifiers
How to protect respondents’ confidentiality?

- **Top-code**: Provide average of expenditures above threshold
- **Re-code**: Change metadata but provide numerical data
- **Suppress**: Delete numerical data or entire record
How to protect respondents’ confidentiality?

- **Top-code**: Provide average of expenditures above a threshold
- **Re-code**: Change metadata but provide numerical data
- **Suppress**: Delete numerical data or entire record
How do we topcode?

- Determine critical value
- Find values exceeding critical value
- Average values exceeding critical value
- Replace values with top-coded values
Topcoding example
Topcoding example

Critical value
Topcoding example

Extreme values
Topcoding example

Topcoded values
How to determine critical values?

- **Percentiles:**
  - Population & expenditure: 99.5 %
  - Sample: 97 %

- **Outside sources:**
  If sample differs from population
Distribution in Sample

Outlier?
Distribution in Population

probably not
How to protect respondents’ confidentiality?

- **Top-code**: Provide average of expenditures above a threshold
- **Re-code**: Change metadata but provide numerical data
- **Suppress**: Delete numerical data or entire record
How do we recode?

- Find values that meet criteria
- Determine method:
  - Generalize info
  - Change info
- Replace original metadata with recoded metadata
Re-code: Generalize information

- Broaden production year of cars
  - From Toyota Corolla 1999
  - To Toyota Corolla 1990s
Re-code: Change information

- Change data to comparable data
- Change respondents’ age over 82 to 87
How to protect respondents’ confidentiality?

- **Top-code**: Provide average of expenditures above a threshold
- **Re-code**: Change metadata but provide numerical data
- **Suppress**: Delete numerical data or entire record
Suppress

Delete the reported data or delete the entire record
How to suppress?

- Blank out numerical value but maintain metadata
- Erase entire record
Suppression

- Blanking numerical data
  - Blank values of normal but infrequent purchases
  - Example: Specialized mortgages
Suppression

- Complete eradication
  - Erase entire record
  - Example: Airplane purchase
Reverse engineering

What’s X?

5 = 3 + X
Reverse engineering

Prevent the use of available information to deduce protected information
How to prevent reverse engineering?

- Find protected values
- Protect them in all locations
- Protect related values
Reverse engineering

- Scenarios
  - Within file
  - Across files
Reverse Engineering:
Within File

- Income = Wage + taxes
- 1000 = 800 + 200
- 1000 = 750 + 200
- 950 = 750 + 200

- Critical value: 700
- Topcode value: 750
Reverse Engineering: Within File

- Income = Wage + taxes
- 1000 = 800 + 200
- 1000 = 750 + 200
- 950 = 750 + 200

- Critical value: 700
- Topcode value: 750
Reverse Engineering: Within File

- Income = Wage + taxes
- 1000 = 800 + 200
- 1000 = 750 + 200
- 950 = 750 + 200

- Critical value: 700
- Topcode value: 750
Reverse Engineering: Within File

- $\text{Income} = \text{Wage} + \text{taxes}$
- $1000 = 800 + 200$
- $1000 = 750 + 200$
- $950 = 750 + 200$

- Critical value: 700
- Topcode value: 750
Reverse Engineering: Across Files

- **Income**
  Topcoded income in FMLI
  =>$ \text{Topcode associated UCC in ITBI}$

- **Expenditure**
  Topcoded expenditures in EXPN/FMLI
  =>$ \text{topcode associated UCC in MTBI}$
How do we document?

- Flag the values
  - **T**: Topcoded value
  - **D**: Valid value
What percentage of data points changed?

- Un-weighted impact:
- Weighted impact:
Impact on trends?

- No: ???????
- Small: ???????
- Large: Area and income extremes
Arcenis Rojas

(202)-691-6884
Rojas.Arcenis@bls.gov
Next presentation...