Refining disclosure controls for the Census of Fatal Occupational Injuries (CFOI)

Danny Friel

Alyssa Gillen, Julie Krautter, Yvan Saastamoinen

Government Advances in Statistics Computing (GASP) conference

June 14, 2023

Disclosure control

- The direct or indirect release of sensitive or private information about a survey or census unit
- Data users have access to other information
- Goal: minimize practical risk

Census of Fatal Occupational Injuries (CFOI)

- Publishes a complete count of fatal injuries each year
- Protecting CFOI data is challenging
  - No sampling
  - Fatal injuries are rare events
  - Exact counts are important
  - Sixteen grouping variables for cells

Primary vs. secondary suppression

Primary suppression only	
The count for occupation 3 doesn’t meet publishability criteria	
Occupation	Number of fatal injuries
All occupations	**100**
Occupation 1	80
Occupation 2	18
Occupation 3	--

Even though this cell is suppressed, we have enough information to compute its value: 100 – 80 – 18 = 2

Primary and secondary suppressions	
The count for occupation 2 is suppressed as well	
Occupation	Number of fatal injuries
All occupations	**100**
Occupation 1	80
Occupation 2	--
Occupation 3	--

With two cells suppressed, we don’t have enough information to compute either value.
Possible values include 20 and 0, 19 and 1, 10 and 10, 15 and 5...

The CFOI Hypercube

- Scans for:
  - Primary suppressions
  - Secondary suppressions (within-tables)
  - Secondary suppressions (across tables) for up to four CFOI variables

CFOI variables	Number of variables	Number of cells screened by hypercube
Fatal injuries by industry	1	**706**
Fatal injuries by industry, event/exposure	2	706 x 8 = **5,648**
Fatal injuries by industry, event/exposure, state	3	706 x 8 x 56 = **316,288**
Fatal injuries by industry, event/exposure, state, age	4	706 x 8 x 56 x 9 = **2,846,592**

Proposal 1: publishing zeroes

- Zero counts mean that no cases met the criteria for a cell
- Zero counts are especially meaningful from a program & policy perspective
- Two questions:
  - How do zero cells impact the effectiveness of secondary suppressions?
  - How can the hypercube be trained to only suppress zeroes when they pose a substantial confidentiality risk?

Counting zeroes for industry-event cells (Table A-1)

	Table A-1 cells
Published	3,734 (4.76%)
Not published	74,764
Total	**78,498**

Zero	65,750 (87.9%)
Non-zero	9,014

Zero-count cells make up the majority of Table A-1 cells

Adding zeroes to a data table (simulated data)

	All Events	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6
Industry A	**22**	**4**	**--**	**8**	**--**	**4**	**--**
Industry A-1	**8**	--	--	--	--	1	--
Industry A-2	**12**	--	--	3	--	3	--
Industry A-3	**2**	--	--	--	--	--	--

Adding zeroes to a data table (simulated data)

	All Events	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6
Industry A	**22**	**4**	**--**	**8**	**0**	**4**	**--**
Industry A-1	**8**	--	--	--	0	1	--
Industry A-2	**12**	--	--	3	0	3	--
Industry A-3	**2**	0	0	--	0	0	0

Adding zeroes to a data table (simulated data)

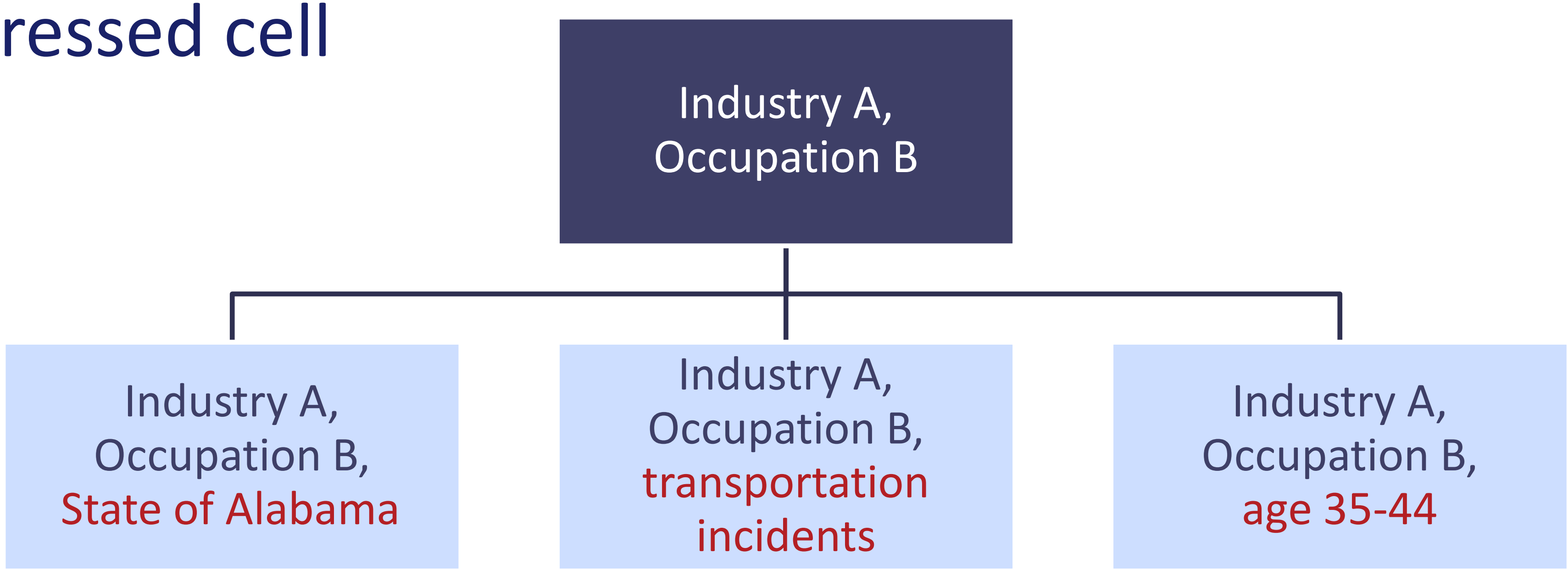
	All Events	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6
Industry A	**22**	**4**	**--**	**8**	**0**	**4**	**--**
Industry A-1	**8**	--	--	--	0	1	--
Industry A-2	**12**	--	--	3	0	3	--
Industry A-3	**2**	--	0	--	0	0	--

Preliminary results

- More than half of zeroes may be publishable.
  - Zeroes must not be published if:
    - They can be used to derive the value of a suppressed nonzero cell
  - Zeroes could be published if:
    - They are the child of a non-zero cell or a suppressed zero cell
  - Zeroes can also be suppressed at random, to infuse more uncertainty
    - E.g., if there are > 3 zeroes in a table row, up to 2 may be randomly suppressed

Proposal 2: Parent-child suppressions

- How does the hypercube identify cross-table suppressions?
- One option: any cell that is a mathematical subset of a suppressed cell



Industry A, Occupation B

Industry A, Occupation B, State of Alabama

Industry A, Occupation B, transportation incidents

Industry A, Occupation B, age 35-44

Proposal 3: use of categorical ranges

- Suppressions make it more difficult to back out information about individual cases
  - Confidentiality vs usability tradeoff
- Is it possible to safely provide partial information about sensitive cells?
  - Use ranges like 1-5 and 6-10 instead of fully suppressing sensitive cells

Summary

- Zero-count cells and child cells must be displayed selectively
  - Zeroes and child cells can be used to derive sensitive information
  - Cells that are non-sensitive in one table may be sensitive in another
  - Additional uncertainty can be infused as needed
- Partial information could be provided for some cells
- Post-processing time and resources may limit options

Contact Information

Thank you to my collaborators:

Alyssa Gillen
Julie Krautter
Yvan Saastamoinen

Danny Friel
Office of Compensation and Working Conditions
Friel.Daniel@bls.gov