SDSS
SYMPOSIUM ON
DATA SCIENCE & STATISTICS
BEYOND BIG DATA: INFLUENCING
SCIENCE, TECHNOLOGY, AND SOCIETY
PITTSBURGH, PA • JUNE 7–10, 2022

RAND
CORPORATION

Variable Suppression and Recoding

SDSS 2022

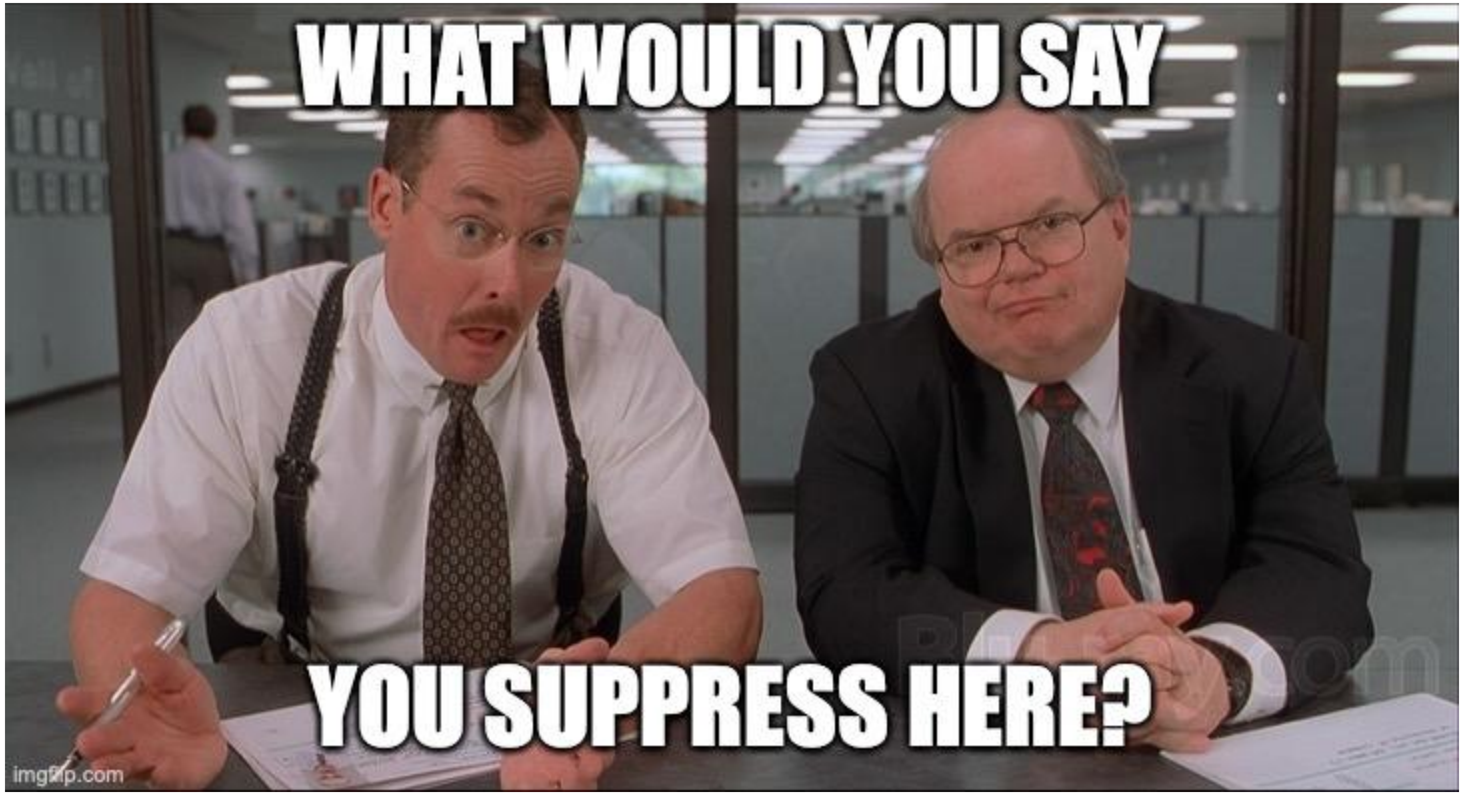
Statistical Data Privacy Techniques for Sharing Sensitive Data

Part 2

Assumptions about the Data for This Lesson

- Assume we have microdata on individuals
- Primarily concerned with categorical data
  - Assume continuous variables will be binned
- For example a survey of teachers:
  - Individual and school demographics
  - Experiences with curriculum/teaching methods
  - Attitudes towards school leadership

What Do We Mean by Suppression?



There are Multiple Types of Suppression for Microdata

- Variable suppression:

ID	Age	Sex	Race
1			
2			
3			

ID	[redacted]	Sex	Race
1	[redacted]		
2	[redacted]		
3	[redacted]		

- Local (value/observation) suppression:

ID	Age	Sex	Race
1			
2			
3			

ID	Age	Sex	Race
1	[redacted]		
2			
[redacted]	[redacted]	[redacted]	[redacted]

Recoding as an Alternative to Local Suppression

- Local suppression induces missingness not at random (MNAR)
- Instead of removing values, we recode categories to collapsed levels
  - E.g., rather than drop race/ethnicity for small minority groups → collapse 7 reported Census race/ethnicity categories to 3 levels
- Recoding removes granularity but allows unbiased estimates*

Suppression for Tabular Data Will Not be Covered

- Cell suppression:

Sex/Race	W	Non-W
M		
F		

Sex/Race	W	Non-W
M		
F		[redacted]

- Table suppression:

Sex/Race	W	Non-W
M		
F		

Subset to Age > 80

Sex/Race	W	Non-W
M	[redacted]	[redacted]
F	[redacted]	[redacted]

Why use Suppression to Protect Data?

- The released data values are unaltered
- Easy to communicate how the data were protected
- Does not require complex statistical models or computational capabilities*

What are Drawbacks to Data Suppression?

- Difficult to determine whether data have been sufficiently protected
- Can require high amount of suppression
- More likely to remove information about minority groups

Returning to the Example Survey of Teachers

- Variables collected include:
  - Individual and school demographics
  - Experiences with teaching approaches
  - Attitudes towards school leadership
- Question: what might we want to suppress?

Image: https://www.rand.org/education-and-labor/projects/aep/run-a-survey.html

Returning to the Example Survey of Teachers

- Variables collected include:
  - Individual and school demographics
  - Experiences with teaching approaches
  - Attitudes towards school leadership
- Question: what might we want to suppress?
  - *Free text responses*
  - *Sensitive opinions (potentially)*
  - *Variables that uniquely identify teachers*

Image: https://www.rand.org/education-and-labor/projects/aep/run-a-survey.html

We Need to Define Our Risk Scenario

- Before suppressing data, we need to know what constitutes risk
- Assumed attack scenario: data linkage

Released File

ID	Age	Sex	Race	ZIP
1	41	F	W	15218
2	77	M	AA	15213
…				

→

Attacker File

ID	Age	Sex	Race	ZIP	Name	
1	41	F	W	15218	XX	✓
2	72	F	AA	15213	XX	✗
…					…	

- *Define: risk is the probability of being matched based on quasi-identifiers*

We Can Define Two Scenarios for This Risk

- Scenario 1: participation in the survey is known
    - Risk based on frequency of quasi-identifiers *among* those in the data
- Scenario 2: participation in the survey is not know
    - Risk based on frequency of quasi-identifiers *among* the population
- Additional necessary assumptions:
    - What data are available to the attacker
        - Knowledge about participation
        - Set of quasi-identifiers
    - Information about the population frequencies (for scenario 2)

Risk When Survey Participation is Known

- E.g., 3 individuals with the same age, sex, and race → risk is ⅓
- Risk can be calculated easily:
  - Compute frequencies for all quasi-identifier combinations
  - Compute individual risk as (1/survey combination$_i$ freq.)
- Common risk metrics:
  - Max risk
  - Mean risk
  - Total (sum) risk

Risk When Survey Participation is Not Known

- Previous steps are replicated
- Additionally, compute frequencies *in the population*
    - E.g., how many in the population with the same age, sex, and race
- Risk is now (1/pop. combination$_i$ freq.)
- Assume complete population frequencies are known
    - More complex methods exist if using only marginal totals or weights

Takeaways From This Lesson

- Knowledge about participation is a crucial assumption
    - Much easier to protect individuals if they are sampled from a population
- Sometimes it is necessary to suppress a lot of data
- Small groups get suppressed more often

Further Reading

- Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. Vol. 2. New York: Wiley, 2012.
- Willenborg, Leon, and Ton De Waal. *Statistical disclosure control in practice*. Vol. 111. Springer Science & Business Media, 1996.
- Skinner, C. J., and David J. Holmes. "Estimating the re-identification risk per record in microdata." *Journal of Official Statistics* 14, no. 4 (1998): 361.
- Templ, Matthias, Alexander Kowarik, and Bernhard Meindl. "Statistical disclosure control for micro-data using the R package sdcMicro." *Journal of Statistical Software* 67 (2015): 1-36.
- Benedetti, Roberto, A. Capobianchi, and L. Franconi. "Individual risk of disclosure using sampling design information." *Contributi Istat* 1412003 (1998).