Disclosure Risks in Microdata

Jingchen (Monika) Hu

Vassar College

Data Confidentiality

Outline

Outline

Microdata and the ACS sample

- Microdata: also called record-level or respondent-level data, is a collection of a number of variables / attributes for a survey of individuals or business establishments.
- The American Community Survey (ACS) sample:

```
ACSdata <- read.csv(file = "ACSdata.csv")
head(ACSdata)
```

```
##   SEX RACE MAR LANX WAOB DIS HICOV MIG SCH HISP
## 1   2    1   1    2    1   2     1   3   1    1
## 2   1    1   1    2    1   2     1   3   1    1
## 3   2    6   5    1    4   2     2   2   1    1
## 4   1    1   3    2    1   1     1   1   1    1
## 5   2    1   1    2    1   2     1   1   1    1
## 6   1    1   1    2    1   2     1   1   1    1
```

Microdata and the ACS sample cont'd

Variable	Information
SEX	1 = male, 2 = female
RACE	1 = White alone, 2 = Black or African American alone, 3 = American Indian alone, 4 = other, 5 = two or more races, 6 = Asian alone
MAR	1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married
LANX	1 = speaks another language, 2 = speaks only English
WAOB	born in: 1 = US state, 2 = Puerto Rico and US island areas, oceania and at sea, 3 = Latin America, 4 = Asia, 5 = Europe, 6 = Africa, 7 = Northern America
DIS	1 = has a disability, 2 = no disability
HICOV	1 = has health insurance coverage, 2 = no coverage
MIG	1 = live in the same house (non movers), 2 = move to outside US and Puerto Rico, 3 = move to different house in US or Puerto Rico
SCH	1 = has not attended school in the last 3 months, 2 = in public school or college, 3 = in private school or college or home school
HISP	1 = not Spanish, Hispanic, or Latino, 2 = Spanish, Hispanic, or Latino

Microdata and the ACS sample cont'd

When this sample is released to the public, can you think of potential disclosure risks for all the individuals in this survey sample?

Outline

Terminology

- Disclosure: disclosing sensitive information of an individual.
- What can be sensitive information?

Terminology

- Disclosure: disclosing sensitive information of an individual.
- What can be sensitive information?

- Disclosure risks: the risks that an intruder or attacker, who uses publicly available database to derive confidential information about individuals who are in the database.

Terminology

- Disclosure: disclosing sensitive information of an individual.
- What can be sensitive information?

- Disclosure risks: the risks that an intruder or attacker, who uses publicly available database to derive confidential information about individuals who are in the database.

- Example: your neighbor with $\{\text{SEX} = 1, \text{RACE} = 1, \text{MAR} = 1\}$ in the publicly available ACS sample. What additional information about about your neighbor can you learn?

Type 1: Identification disclosure

- Find out about the identify of the person an intruder is looking for.
- Select all individuals sharing the same combination $\{\text{SEX} = 1, \text{RACE} = 1, \text{MAR} = 1\}$:

```
NeighborSet <- ACSdata %>% filter(SEX == 1 & RACE == 1 & MAR == 1)
dim(NeighborSet)
```

```
## [1] 2192   10
```

Type 1: Identification disclosure

- Find out about the identify of the person an intruder is looking for.
- Select all individuals sharing the same combination $\{\text{SEX} = 1, \text{RACE} = 1, \text{MAR} = 1\}$:

```
NeighborSet <- ACSdata %>% filter(SEX == 1 & RACE == 1 & MAR == 1)
dim(NeighborSet)
```

```
## [1] 2192   10
```

- Random guess which one is your neighbor - risk as a probability:

```
1/dim(NeighborSet)[1]
```

```
## [1] 0.0004562044
```

Type 1: Identification disclosure

- Find out about the identify of the person an intruder is looking for.
- Select all individuals sharing the same combination $\{\text{SEX} = 1, \text{RACE} = 1, \text{MAR} = 1\}$:

```
NeighborSet <- ACSdata %>% filter(SEX == 1 & RACE == 1 & MAR == 1)
dim(NeighborSet)
```

```
## [1] 2192   10
```

- Random guess which one is your neighbor - risk as a probability:

```
1/dim(NeighborSet)[1]
```

```
## [1] 0.0004562044
```

- Important to note: we do not actually possess the true identity, so only the data holder (e.g. U.S. Census Bureau) can declare an identification disclosure if it happens.

Exercises

- #1 Additional information: $\{\text{WAOB} = 1\}$. How many records? How likely for you to find out the identity?

Exercises

- #1 Additional information: $\{\text{WAOB} = 1\}$. How many records? How likely for you to find out the identity?

- #2 If you are allowed to know about one more variable / attribute in the survey, which one will increase your chance the most to identify your neighbor?

Exercises

- #1 Additional information: $\{\text{WAOB} = 1\}$. How many records? How likely for you to find out the identity?
- #2 If you are allowed to know about one more variable / attribute in the survey, which one will increase your chance the most to identify your neighbor?
- #3 A unique observation with $\{\text{SEX} = 1, \text{RACE} = 5, \text{MAR} = 3, \text{LANX} = 1, \text{WAOB} = 3\}$. What can you learn about her?

Exercises

- #1 Additional information: $\{\text{WAOB} = 1\}$. How many records? How likely for you to find out the identity?
- #2 If you are allowed to know about one more variable / attribute in the survey, which one will increase your chance the most to identify your neighbor?
- #3 A unique observation with $\{\text{SEX} = 1, \text{RACE} = 5, \text{MAR} = 3, \text{LANX} = 1, \text{WAOB} = 3\}$. What can you learn about her?

```
ACSdata %>%
  filter(SEX == 1 & RACE == 5 & MAR == 3 & LANX == 1 & WAOB == 3)
```

```
##   SEX RACE MAR LANX WAOB DIS HICOV MIG SCH HISP
## 1   1    5   3    1    3   2     1   1   1    2
```

Type 2: Attribute disclosure

- An intruder correctly infers the true value of one (or a set of) unknown variable(s) / attribute(s) of an individual.
- For a uniquely identified person:

```
ACSdata %>%
  filter(SEX == 1 & RACE == 5 & MAR == 3 & LANX == 1 & WAOB == 3)
```

```
##   SEX RACE MAR LANX WAOB DIS HICOV MIG SCH HISP
## 1   1    5   3    1    3   2     1   1   1    2
```

- What about for a non-uniquely identified person, say your neighbor with $\{\text{SEX} = 1, \text{RACE} = 1, \text{MAR} = 1\}$, what if we want to find out about his sensitive DIS status?

Type 2: Attribute disclosure cont'd

- One simple strategy is to check the DIS breakdown among the 2191 individuals:

```
NeighborSet %>% count(DIS) %>% group_by(DIS)
```

```
## # A tibble: 2 x 2
## # Groups:   DIS [2]
##     DIS     n
##   <int> <int>
## 1     1   339
## 2     2  1853
```

- We could randomly guess the neighbor's DIS value given this proportion (our best guess).
- Important to note: we do not actually possess the true DIS value, so only the data holder (e.g. U.S. Census Bureau) can declare an attribute disclosure if it happens.

Summary

- Later in the semester, we will learn how to evaluate disclosure risks more formally.
- The methods we cover make assumptions of the intruder's knowledge and behavior.
- There are methods which do not make these assumptions (formal privacy).

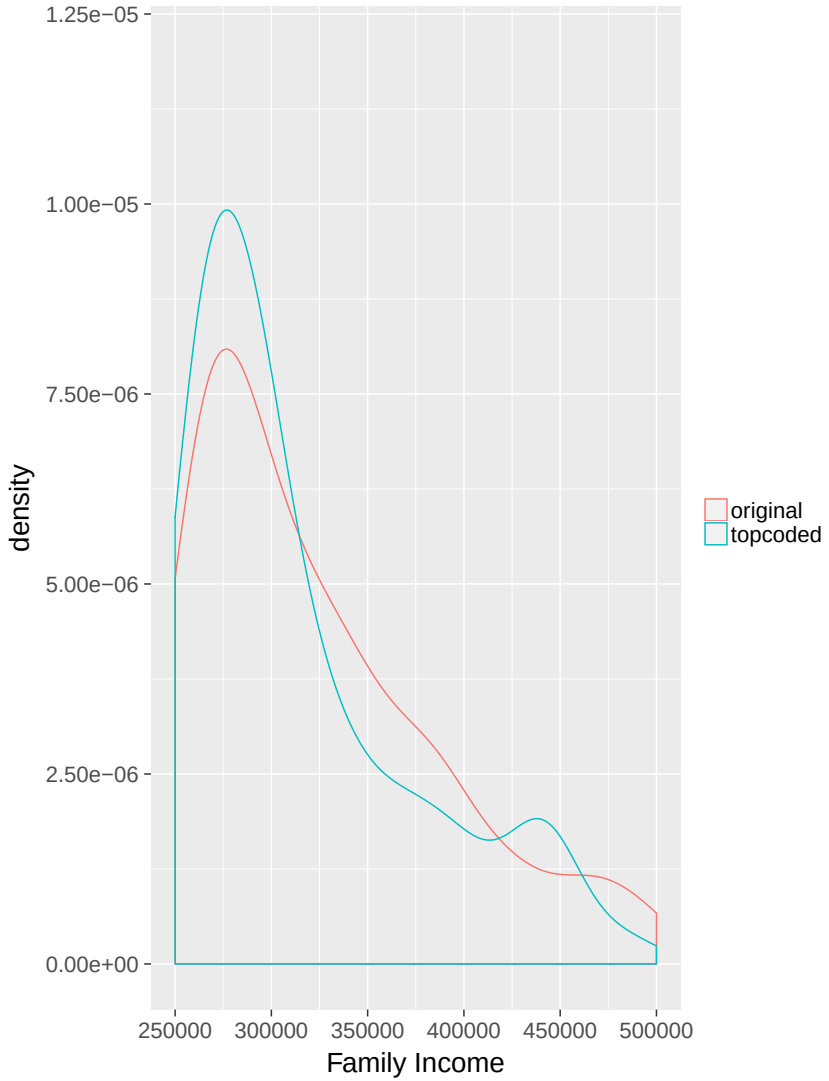
Outline

Traditional methods

Hundepool et al. (2012)

- Adding noise
- Recoding, topcoding, bottom coding
- Resampling
- Data swapping

Examples of topcoding



Synthetic data approach

Rubin (1993) and Little (1993) proposed the synthetic data.

- Simulate records from Bayesian statistical models that are estimated from the original confidential data.
- Balance of data utility and disclosure risks:
  - preserve relationships of variables.
  - low disclosure risks.
- Allow data analysts to make valid inference for a wide class of analyses.

Examples of synthetic data results - utility

Hu and Savitsky (2019+)

	estimate	95% C.I.
Data	72090.26	[70127.02, 74053.50]
Synthesizer	72377.12	[70412.90, 74415.81]

Table 2: Table of C.I. of mean family income.

	estimate	95% C.I.
Data	50225.15	[48995.01, 52000.00]
Synthesizer	50538.50	[49043.63, 52115.76]

Table 3: Table of C.I. of median family income.

Examples of synthetic data results - utility

Hu and Savitsky (2019+)

	estimate	95% C.I.
Data	153916.30	[147582.40, 159603.80]
Synthesizer	152597.10	[147647.40, 157953.80]

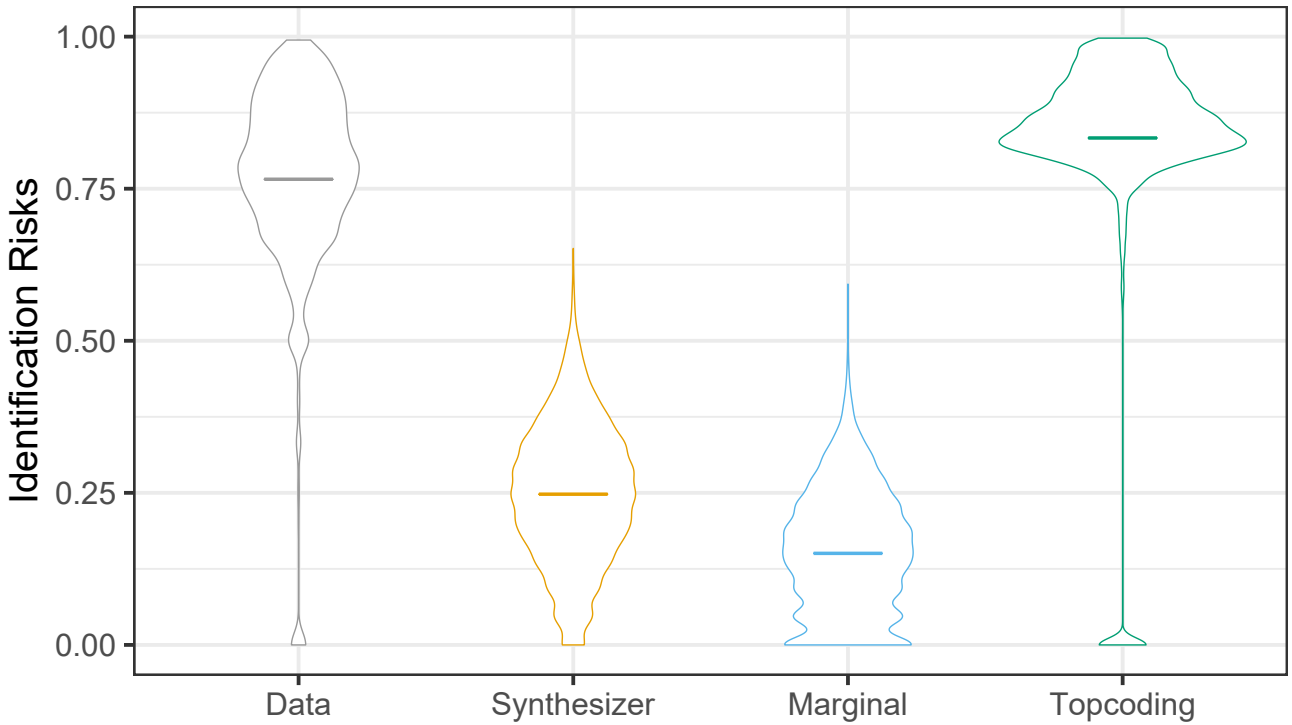
Table 4: Table of C.I. of 90% quantile.

	estimate	95% C.I.
Data	-45826.20	[-49816.29, -41836.11]
Synthesizer	-46017.29	[-50239.20, -41795.37]

Table 5: Table of C.I. of predictor Earner 2 of family income

Examples of synthetic data results - disclosure risks

Hu and Savitsky (2019+)



Bayesian synthesizers

- To generate synthetic data, we first develop Bayesian statistical models on the original confidential data.
- We can then generate synthetic data from the estimated Bayesian synthesizer, i.e. simulate variables / attributes from the posterior predictive distribution.

Bayesian synthesizers

- To generate synthetic data, we first develop Bayesian statistical models on the original confidential data.
- We can then generate synthetic data from the estimated Bayesian synthesizer, i.e. simulate variables / attributes from the posterior predictive distribution.

- Toy example:
  - A Bayesian synthesizer:
    - ★ the sampling model: $y_i \overset{i.i.d.}{\sim} \text{Normal}(\mu, \sigma),\ i = 1, \cdots, n.$
    - ★ a prior for μ: $\mu \sim \pi(\mu \mid \theta)$.
  - Simulate synthetic data:
    - ★ simulate posterior draws: $\mu* \sim \pi(\mu \mid y_1, \cdots, y_n)$.
    - ★ simulate posterior predictive draws: $y_i* \sim \text{Normal}(\mu*, \sigma),\ i = 1, \cdots, n.$

Brainstorm potential Bayesian synthesizers

For sensitive

- continuous variable(s).
- binary variable(s).
- categorical variable(s).

Synthetic data production process

- Develop suitable Bayesian synthesizer(s).
- Generate synthetic data: partial vs full.
- Evaluate disclosure risks and utility.
- Further tuning if either disclosure risks or utility or both are not satisfactory.
- Determine the release of synthetic microdata.

References

- Hu, J. and Savitsky, T. D. (2019+). Risk-efficient Bayesian pseudo posterior data synthesis for privacy protection. arXiv: 1908.07639.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P. P. (2012). Statistical Disclosure Control, Wiley Series in Survey Methodology, Wiley.
- Little, R. J. A. (1993). Statistical analysis of masked data. Journal of Official Statistics 9, 407-426.
- Rubin, D. B. (1993). Discussion statistical disclosure limitation. Journal of Official Statistics 9, 461-468.