Microdata protection

A method that combines subsampling and calibration

Maxime Bergeat

French National Institute for Statistics and Economic Studies
Statistical Methods Division

UNECE Work Session on Statistical Data Confidentiality
5th October 2015

Outline

Outline

Anonymization process - "Traditional" approach

- Distinguish in the dataset between
  - Direct identifiers
  - Quasi-identifiers
  - Non-identifying variables
- Objectives for disclosure risk reduction
  - k-anonymity
  - Limit re-identification risk

Some methods for disclosure risk reduction

- Non-perturbative methods
  - Global recoding
  - Local suppression
  - Classical approach to reach k-anonymity
- Perturbative methods
  - Difficulty to know when it is enough perturbed
- Generation of synthetic data
  - We won't talk more about it in this presentation

Outline

Subsampling and calibration

- Two main steps
- Step 1 (disclosure risk reduction): suppression of all records with a too high risk of re-identification
  - No requirement about how is re-identification risk is estimated
- Step 2 (restore data utility in resulting file): calibration in order to maintain some margins
  - Sociodemographic variables
  - Possible to use variables of interest for calibration!

Computation of calibrated weights

- Initial sample denoted S, subsample after deletion of risky records denoted S'
- Goal of calibration for a variable which takes the value x_k for a record $k \in S$: computation of weights $w_k \ \forall \, k \in S'$ verifying:

$$\sum_{k \in S'} w_k x_k = \sum_{k \in S} d_k x_k$$

- d_k: "final" weight (present in original dataset) after correction for non-response and potential consideration of auxiliary information with previous calibration
  - Not necessarily the weight used to initialize the calibration procedure

Original microdata

Full name	Gender	Age category	Favourite meal	Sampling weight
Ambre Marval	Female	-24	raclette	1 000
Cathy Pradel	Female	-24	fish soup	1 500
France Jabot	Female	25-49	sauerkraut	2 000
Ghislaine Metayer	Female	+50	calf's head	1 100
Mireille Henri	Female	+50	calf's head	1 400
Robert Briton	Male	-24	fish soup	800
Louis Brandt	Male	25-49	raclette	1 100
Jean Achard	Male	25-49	beef stew	1 900
Jacques Crillou	Male	+50	sauerkraut	1 200

Subsampling

Gender	Age category	Favourite meal	Sampling weight
Female	-24	raclette	1 000
Female	-24	fish soup	1 500
Female	+50	calf's head	1 100
Female	+50	calf's head	1 400
Male	25-49	raclette	1 100
Male	25-49	beef stew	1 900

2-anonymous dataset

Calibration

Gender	Age category	Favourite meal	**Calibrated weight**
Female	-24	raclette	**1 400**
Female	-24	fish soup	**1 900**
Female	+50	calf's head	**1 700**
Female	+50	calf's head	**2 000**
Male	25-49	raclette	**2 100**
Male	25-49	beef stew	**2 900**

Calibration variables: "Gender" and "Age category"

Outline

"Thefts, violence and safety" survey

- French household survey with use of Internet and paper questionnaire for data collection
- Main goal: compare results of this survey with another french survey with personal interviews
- 12 901 respondents
- Considered delinquency acts: theft in the housing, vehicle theft, other theft with violence, other theft without violence, physical violence, threat

Data after global recoding

- Quasi-identifiers:
  - Gender
  - Income (-1000€ / 1000-2000€ / 2000-3000€ / 3000-6000€ / +6000€)
  - Age category (-24, 25-34, 35-44, 45-54, 55,64, +65)
  - Size of urban unit (rural area / urban unit with less than 100 000 inhabitants, urban unit with more than 100 000 inhabitants, Paris)
  - Highest qualification achieved (5 categories)
  - Lives in a couple
  - Household size (1 person / 2 / 3-4 / 5+)
- Goal for disclosure risk reduction: get a 3-anonymous dataset

First 3-anonymous dataset: local suppression

Quasi-identifier	Cost of suppression	Number of deletions
Gender	70	0
Income	60	4
Age	50	9
Size of urban unit	40	25
Qualification	30	493
Lives in a couple	20	351
Household size	10	2 296

3178 deletions

Second 3-anonymous dataset: global suppression and weight calibration

- 3033 suppressed records (23.5%)
- Calibration variables:
  - Cross-variable gender $\times$ synthetic indicator of victimisation (3 categories)
  - Cross-variable Age category $\times$ synthetic indicator of victimisation
  - Cross-variable Size of urban unit $\times$ synthetic indicator of victimisation
  - Cross-variable Highest qualification achieved $\times$ synthetic indicator of victimisation
  - Cross-variable Household size $\times$ synthetic indicator of victimisation
- Method to compute calibrated weights: *raking ratio*

Descriptive statistics - Victimisation rates (1)

Original microdata			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	87.3%	9.7%	3.0%
2	84.4%	12.1%	3.5%
3-4	82.5%	12.3%	5.2%
5+	77.9%	16.0%	6.1%

3-anonymous dataset after local suppression			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	88.3%	9.0%	2.7%
2	85.7%	11.1%	3.2%
3-4	82.4%	12.2%	5.4%
5+	77.6%	14.6%	7.8%

Descriptive statistics - Victimisation rates (2)

Original microdata			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	87.3%	9.7%	3.0%
2	84.4%	12.1%	3.5%
3-4	82.5%	12.3%	5.2%
5+	77.9%	16.0%	6.1%

3-anonymous dataset after subsampling and calibration			
Household size	Non-victim	Victim of one delinquency act	"Multi-victim"
1	87.3%	9.7%	3.0%
2	84.4%	12.1%	3.5%
3-4	82.5%	12.3%	5.2%
5+	77.9%	16.0%	6.1%

Descriptive statistics - Victimisation rates (3)

Original microdata			
Lives in a couple	Non-victim	Victim of one delinquency act	"Multi-victim"
Yes	85.6%	11.1%	3.3%
No	79.5%	14.2%	6.3%

3-anonymous dataset after local suppression			
Lives in a couple	Non-victim	Victim of one delinquency act	"Multi-victim"
Yes	85.6%	11.2%	3.2%
No	79.3%	14.3%	6.4%

3-anonymous dataset after subsampling and calibration			
Lives in a couple	Non-victim	Victim of one delinquency act	"Multi-victim"
Yes	84.9%	11.6%	3.5%
No	79.2%	14.5%	6.3%

Conclusion

- Microdata protection: a trade-off between reduction of disclosure risk and loss of data utility
- Presented method based on subsampling and weight calibration
  - Total control of disclosure risk
  - Encouraging utility results in perturbed file
  - A method designed for public use files
- Further questions
  - How to measure loss of data utility when looking at sophisticated statistics?
  - Suppression of records: a radical solution!
    - Replace them with a non-risky "close" record?
    - Or synthetic data?

 A. Hundepool *et al.*
Statistical disclosure control,
Wiley Series in Survey Methodology, 2012.

 J.-C. Deville and C.-E. Särndal.
Calibration estimators in survey sampling,
Journal of the American Statistical Association, 87:376–382, 1992.

Thank you for your attention ☺