# Statistical disclosure control and micro data

## Problems and criteria

# Contents

- Stating the problem(s)
- Possible criteria
  - uniqueness
  - rare combinations
- sensitive variables
- household variables
- regional variables
- sampling weights
- Miscellaneous topics

# Leading to the problem (1)

Release of records on individual respondents:

1. Data-set leaving the institute
    - persons/social
2. Data-set analysed at the institute (on-site)
    - demographic
    - economic
3. Data-set analysed remotely
    - remote execution
    - remote access

# Leading to the problem (2)

| ***Soc.Sec. Nr.*** | ***Gender*** | ***Age class*** | ***Region*** | ***Education*** | ***Profession*** | ***Income*** |
|---|---|---|---|---|---|---|
| 1927384123 | Female | 40-55 | The Hague (large) | Higher | Civil Servant | 40,000 |
| 1927384124 | Male | 30-40 | Urk (small) | Middle | Fisherman | 20,000 |
| 1927384125 | Male | 55+ | Amsterdam (large) | Unknown | Mayor | 100,000 |
| 1927384126 | Male | 20-30 | Dordrecht (medium) | Lower | Plumber | 30,000 |
| 1927384127 | Female | 55+ | Staphorst (small) | Higher | Surgeon | 100,000 |
| 1927384128 | Male | 30-40 | Woensdrecht (small) | Higher | IT consultant | 45,000 |
| 1927384129 | Male | 55+ | Rotterdam (large) | Unknown | Surgeon | 100,000 |
| 1927384130 | Female | 20-30 | Borger (tiny) | Middle | Violin maker | 35,000 |
| 1927384131 | Female | 30-40 | Utrecht (large) | Lower | House cleaner | 15,000 |
| … | … | … | … | … | … | … |
| … | … | … | … | … | … | … |

## Leading to the problem (3)

(Re-) identification of respondents

- Remove **direct**/formal identifiers
  - (name & address,
    social security number,
    bank account number,
    registration number at Chamber of Commerce, …)

Not enough:

- Rare combinations of **indirect** identifiers

# Leading to the problem (4)

Examples:

Unique combination of indirect identifiers

- Place of residence: Budapest  
  Occupation: Mayor
- Profession: Employee at SN  
  Place of residence: Dordrecht, The Netherlands  
  Education: PhD  
  Date of birth: 10/10/1967

That’s PP!

# Leading to the problem (5)

Examples:

Rare combinations of indirect identifiers

- Gender: Female
  Profession: Neurologist
  Place of work: Utrecht, The Netherlands
  Age: 55+

## The real problem

(Re-) identification could disclose sensitive information:

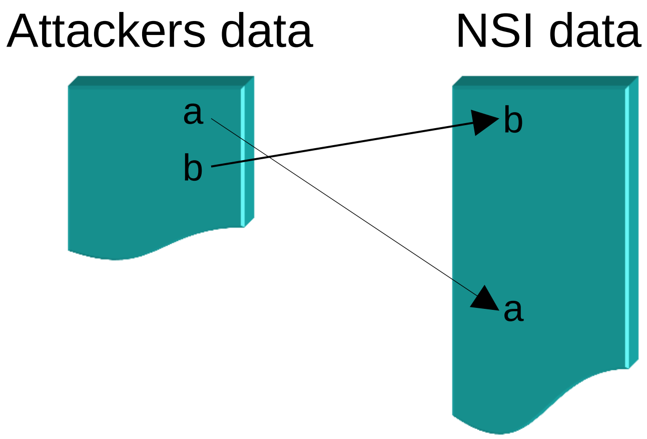
1. Mayor of Budapest is identified
2. Additional info in that record: criminal past

## Goal(s) of SDC with micro data:

Prevent (re-) identification

Prevent occurrence of rare combinations
(define ‘rare’)

# Disclosure scenarios



Matching

- direct search
- fishing

Knowledge about response

Spontaneous recognition

# Disclosure risk

- Number of (direct/indirect) identifiers
- Number of categories per identifier
- (Population) frequency of each category
- Relations between identifiers
- Quality of attacker’s a priori knowledge
- Statistical twins in population
- Costs of identification

# Criteria

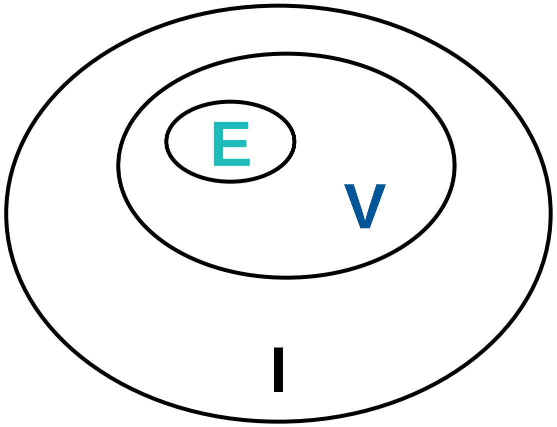
- Identifying variable:
  - value may, possibly in combination with other values, lead to (re-) identification
  - value is easily determined (by acquaintances)

- Sensitive variable:
  - value discloses not easily determined information about respondent
    (e.g.: sexual behaviour, criminal past, physical and mental health, income, …)

# Identifying variables

- Direct (formal) identifiers
  - Name, address, social security number, …

- Indirect identifiers, differentiated into e.g.,
  - Extremely identifying (**E**)
  - Very identifying (**V**)
  - Identifying (**I**)

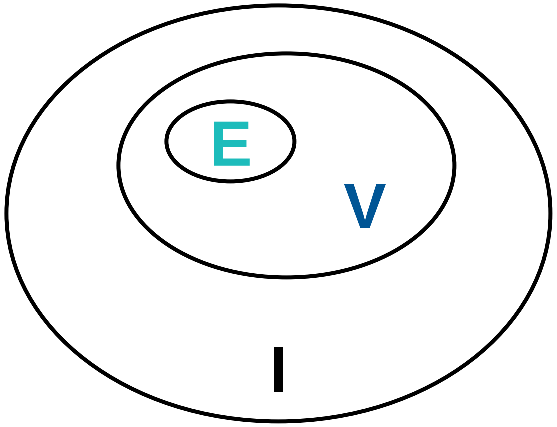


# Examples

- Extremely identifying:
  - Regional variables (residence, work, …)
- Very identifying:
  - Gender, nationality

    + Extremely identifying variables
- Identifying:
  - Age, occupation, education

    + Very identifying variables



# Criteria

Check certain combinations of identifying variables

(Population) frequency > certain threshold

common combination 👍 Safe!

(Population) frequency <= certain threshold

rare combination 👎 Not Safe!

↓

To be protected

# Combinations

Check the (population) frequencies in all combinations consisting of:

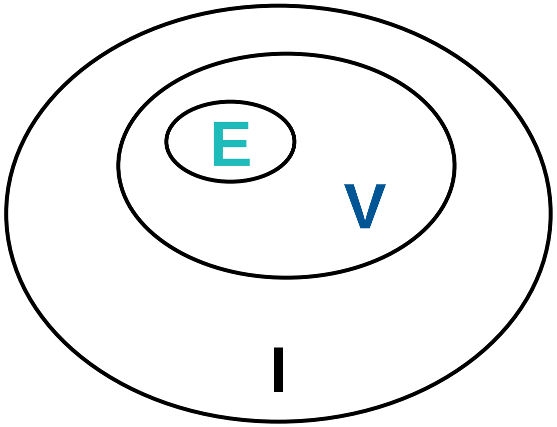
identifying

×

very identifying

×

extremely identifying



# Per record risk

Model attacker’s behaviour (scenario)

Use that model to estimate the probability that a specific record is disclosed
(re-identification risk)

All risk above a certain threshold is then considered sensitive and would require SDC-methods

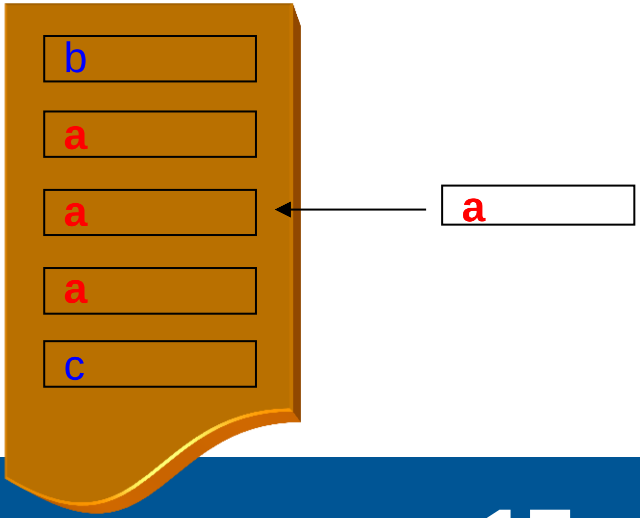
# Per record risk

Frequency count approach: crude per record risk

Frequency count $F_k$ in population

Assume that record is randomly linked with one of the $F_k$ possible matches

Probability that is correct: $1/F_k$



# Disclosure risk

More elaborate models

e.g., attacker has database with

identifiers + key variables *k*

attacker tries to link using key variables

disclosure risk for target record *i* given by

$$r_k = P(\mathit{correct}\ \text{link}\ i \leftrightarrow i^* \mid \text{observed sample})$$

estimate for $r_k$ implemented, using a.o.

sampling design

# Disclosure risk

Basic part of risk for individual $i$:

1/(number of individuals in the population with the same combinations of key variables as $i$) = $1/F_k$

But $F_k$ is unknown!

Modelling needed, e.g., (Benedetti and Franconi, 1998)

$$F_k \mid \pi_k \sim \text{Poisson}(N\pi_k)$$
$$f_k \mid F_k, \pi_k, p_k \sim \text{Binom}(F_k, p_k)$$

# Disclosure risk

$p_k$ is probability that member of population group $C_k$ falls in sample

To estimate $p_k$ for each key $k$ we use the sampling weights $w_i$ available for each record:

$$\hat{p}_k = \frac{f_k}{\sum_{i \in C_k} w_i}$$

in μ-ARGUS

NB: $w_i$ must make 'sense'

# Disclosure risk

Other possibilities:

Use log-linear models to estimate $F_k$

(Elamir and Skinner, 2006)

# Disclosure risk

Other possibilities:

Principle of k-anonymity: each distinct pattern of key variables is possessed by at least k records in the microdata file
(need to choose the number of key variables)

A popular choice is k=3, implying that the same pattern of key variables is possessed by at least 3 records in the microdata file

# Disclosure risk

Other possibilities:

Principle of l-diversity: a group of observations with the same pattern of key variables that contains at least l represented values for the sensitive variable
(need to choose the number of key variables)

For 2-diversity 2 distinct values for the sensitive variable appear in the group of observations with the same pattern of key variables

# Special variables (1)

***Household variables:***

set of records usually have same score on this kind of variables

households are often unique

referring to household

(e.g., household income)

referring to individuals

(e.g., religion)

# Special variables (2)

Possible solution:

prevent regrouping household records

Criterion:

provide sufficient number of households
with same score on household variables

**NB: Changes of scores on these variables should be done consistently over the set of records!**

# Special variables (3)

***Regional variables:***

differentiation

- direct, e.g., place of residence
- indirect, e.g., degree of urbanisation

# Special variables (4)

***Sampling weights:***

Can be (helpful in) identifying!

Examples:

Excluding age in records, but including weights based on oversampling certain ages in certain regions

Weighting scheme depending on region

# Miscellaneous topics (1)

Consider to

- Limit the number of identifying variables
- Outdate the micro data set
- Randomise the order of the records
- Provide only one set per survey

# Miscellaneous topics (2)

Pay special attention to

- Matching with other files

- Panel surveys
  - (mutations are very identifying)

# Miscellaneous topics (3)

Legal measures:

Contract

- only for statistical research
- no attempt to disclose data
- data disclosed by accident may not be misused
- no matching allowed
- results of research must be screened
- data must be destroyed after use

# Types of released micro data

## Micro data for public use

- general public
- educational aspect

## Micro data for research

- established research institutes
- DANS

## Micro data for remote analyses

# Micro data for Research (1)

Contract (DANS)

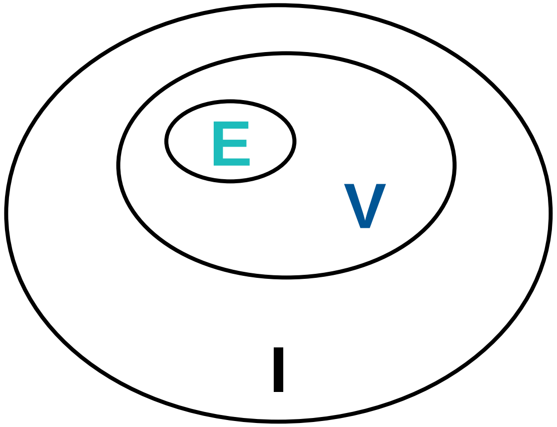
No direct/formal identifiers

Each combination

**E** × **V** × **I**

should occur at least 100 times

in the *(target) population*



# Micro data for Research (2)

**E** : Extremely identifying variables

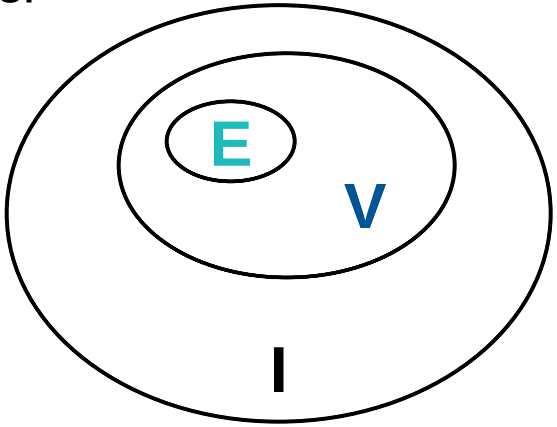
regional variables (residence, place of work)

**V** : Very identifying variables

sex, ethnicity, nationality, extremely identifying vars.

**I** : Identifying variables

occupation, education, age, very identifying vars.



# Micro data for Research (3)

Relation population ⟺ survey/sample

| $f = n/N$ | *threshold in sample* |
|---|---|
| < 1/200 | 1 |
| 1/200 - 1/100 | 2 |
| 1/100 - 1/50 | 3 |
| 1/50 - 1/2 | $2 + 114\,f$ |
| 1/2 - 1 | $19 + 80\,f$ |

# Micro data for Research (4)

Trading off the level of detail on

*business, occupation and education*

versus

*regional variables*

At least *m* inhabitants per region

# Micro data for Public Use (1)

- No direct/formal identifiers
- Micro data set at least one year old
- At most 15 indirect identifiers
- No direct regional variables
  - only 1 kind of indirect regional variables
  - values of indirect regional variable sufficiently spread

# Micro data for Public Use (2)

Sufficiently spread:

- Geographically:
  Each area should spread over at least 6 provinces (The Netherlands = 12 provinces)

- Demographically:

  No municipality in each area may account for more than 50% of total number of inhabitants in that area

# Micro data for Public Use (3)

Check following combinations:

- at least 200 000 individuals in population for each category of identifying variable
- at least 1000 individuals in population for each category in crossing of two identifying variables

# Micro data for Public Use (4)

At least 5 households per combination of categories of household variables

Sampling weights should not provide additional identifying information

Records should be in random order

No sensitive variables…

# Micro data for Remote Analyses

- **Remote execution:**

  Scripts are sent (on-line) to NSI that applies them to micro data. SDC is applied before returning the results.

  (Compare with on-site micro data)

- **Remote access:**

  On-line access to confidentialized micro data sets.

  (Compare with DANS micro data under contract or on-site)