

Anonymisation as Disclosure Avoidance

Mark Elliot

**Confidentiality and Privacy Group
Centre for Census and Survey
Research**

University of Manchester



Proposition

- Data is completely anonymised iff no disclosure about any specific population unit is possible from that data.
- Such a state is theoretical and not guarantees.

Overview

- Statistical Disclosure – some basics
- Statistical Disclosure and Microdata
 - General Concepts
 - Our Approach
- Statistical Disclosure and Aggregate Data
 - General Concepts
 - Our Approach

Statistical Disclosure: Some Basics

Statistical Disclosure: a definition

“The revealing of information about a population unit through the statistical matching of information already known to the revealing agent (often known as a *data intruder*) with other anonymised information (or *target dataset*) to which the intruder has access.”

Key Concepts

- *Identification*: the association of a particular data unit with particular population unit.
- *Attribution*: the association of particular attribute with a particular population unit.

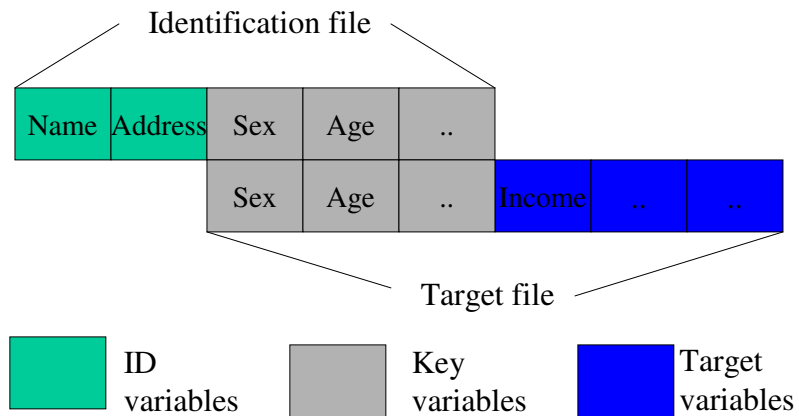
Statistical Disclosure Research

- **Disclosure risk assessment.**
- Disclosure control methodology.
- Analytical validity.

- Microdata and Aggregate data.
- Business and Personal data.
- Intentional and Consequential data

Disclosure Risk: Microdata

The Disclosure Risk Problem: Identification



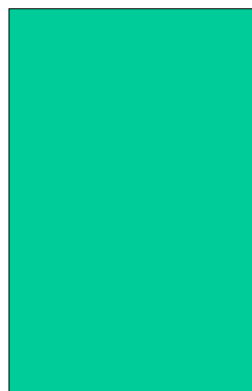
Risk Assessment methods

- File Level
 - Population Uniqueness e.g Bethlehem(1990), Samuels(1998)
 - **DIS; Skinner and Elliot(2002)**
- Record level
 - Statistical modelling (Fienberg and Makov 1998, Skinner and Holmes 1998)
 - **Computational Search Elliot et al (2002); Haglin et al (2009)**

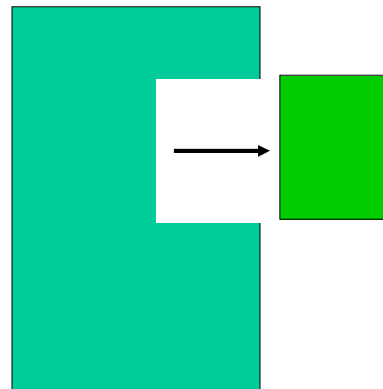
Data Intrusion Simulation

- Uses microdata set (or table) itself to estimate risk - no population data.
- An estimate of the probability of a correct match (given a unique match).
- Special method: sub-sampling and re-sampling.
- General method: derivation from the equivalence class structure.

The DIS Method

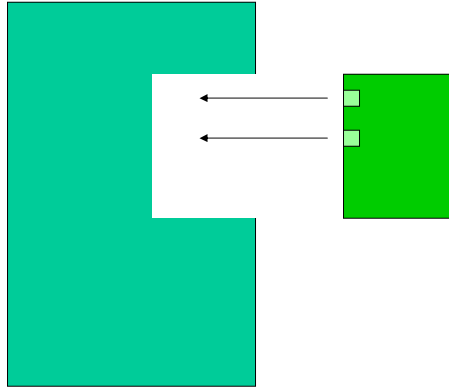


Microdata sample



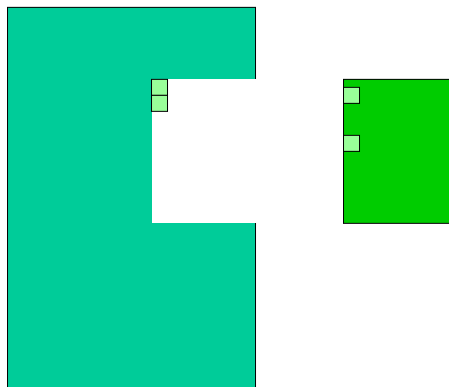
Remove a small number of records

The DIS Method II



Copy back a random number of the removed records (at a probability equivalent to the original sampling fraction)

The DIS Method III



Match the removed fragment against the truncated microdata file

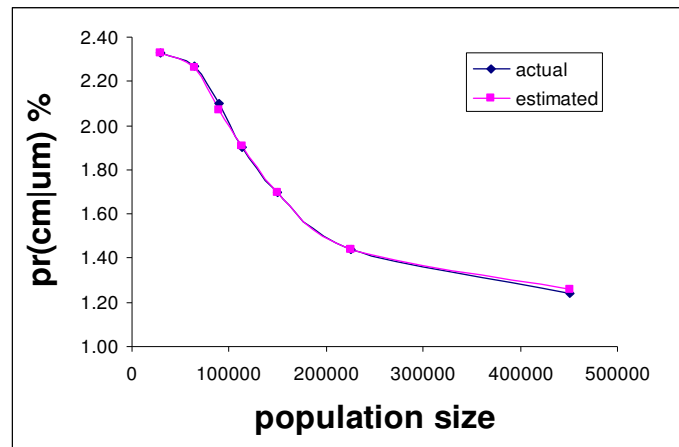
The General DIS Method I

<i>Eq Class</i>	<i>Copied Back</i>	
	<i>Yes</i>	<i>No</i>
<i>Unique</i>	Correct Unique Match	Non Match
<i>Pair</i>	Multiple match including correct	False Unique Match
<i>Triple+</i>	Multiple match including correct	False Multiple Match

Validation

- *Empirical validation studies comparing with the results obtained using population data: Empirical results: No bias and small error. Elliot (2001)*
- *Mathematical proof: Skinner and Elliot (2002).*

Pr(cm|um) for 2% sample with basic key (age sex marital status)



Levels of Risk Analysis

- DIS
 - Works at the file level
 - Very good for comparative analyses
 - e.g. SAMs

Levels of Risk Analysis

- Record level risk is important
 - Variations in risk topography
 - Risky records

Special Uniques

- Original concept
 - Counterintuitive geographical effect, indicated two types of sample uniques.
 - Random and Special
 - Special
 - Demographically unusual (e.g. 16 y-o widow)
 - Random
 - Effect of sampling and variable definition

Special Uniques

- Changing definition:
 1. Sample uniques which remain unique despite geographical aggregation
 2. Sample uniques which remain unique through any variable aggregation
 3. Sample uniques on subset of key variables
 4. Dichotomy to Dimension

SUDA software

- Evaluation version available free under licence.
- Used at ONS, ABS, Stats Singapore, Stats NZ.



Aggregate Data



The Issues

- Aggregate data is often full population data, so measures based on identification disclosure and sampling are meaningless
- A better approach is to evaluate what can be inferred through attribute disclosure

The Disclosure Risk Problem: Attribution

Income levels for two occupations				
	High	Medium	Low	Total
Accademic	0	100	50	150
Pop Stars	100	50	5	155
Total	100	150	55	305

The Disclosure Risk Problem: Attribution

Income levels for two occupations				
	High	Medium	Low	Total
Accademics	1	100	50	150
Pop Stars	100	50	5	155
Total	100	150	55	305

The Disclosure Risk Problem: Attribution

Income levels for two occupations				
	High	Medium	Low	Total
Accademic	0	100	50	150
Pop Stars	100	50	5	155
Total	100	150	55	305

Our Approach:SAP

- Rather than assess the risk of actual attribute disclosure we estimate the probability of producing a **potentially disclosive table**, which we define as any table containing at least one zero.
- The method/measure we propose can be applied to:
 - Single tables
 - Groups of tables
 - Unperturbed and perturbed tables
 - **Unpublished tables**

The Bounds Problem

- In a general sense any set of tables can be viewed as a set of bounds on the full table. For example if we release two one way frequency tables:

Var A	
X	1
Y	2
Z	10
Total	13

Var B	
P	6
Q	3
R	4
total	13

The Bounds Problem

We are effectively releasing the marginals to a two-way frequency table where the entire joint distribution has been suppressed

Var A	Var B			Total
	P	Q	R	
X	?	?	?	1
Y	?	?	?	2
Z	?	?	?	10
Total	6	3	4	13

The cells in the joint distribution can be expressed as a set of bounds (or ranges of feasible values)

Var A	Var B			Total
	P	Q	R	
X	0-1	0-1	0-1	1
Y	0-2	0-2	0-2	2
Z	3-6	0-3	1-4	10
Total	6	3	4	13

The Subtraction – Attribution Probability (SAP) Method

- The risk associated with a table release depends on the set of tables jointly, rather than on the individual tables.
- SAP can be used on single tables, groups of tables, perturbed or unperturbed tables.
- Bounds are calculated and then the probability of an intruder producing one or more upper bounds of zero by *subtracting k* random individuals from the table is calculated
- The output can be set for user defined levels of k

Example output

Mean probability of recovering a zero in a table given a subtraction level (k) for Income Support broken down by Gender and Age for exact and rounded output area tables.

k	Gender			Age		
	Exact	Rounded	Proportion	Exact	Rounded	Proportion
1	0.157	0.000	0.000	0.814	0.000	0.000
2	0.203	0.000	0.000	0.822	0.000	0.000
3	0.248	0.000	0.001	0.830	0.001	0.001
4	0.290	0.012	0.043	0.838	0.011	0.013
5	0.330	0.017	0.051	0.845	0.013	0.016
6	0.367	0.021	0.056	0.853	0.016	0.019
7	0.401	0.024	0.059	0.859	0.019	0.022
8	0.433	0.026	0.061	0.866	0.022	0.025
9	0.461	0.035	0.075	0.872	0.031	0.035
10	0.488	0.037	0.076	0.878	0.034	0.038
20	0.685	0.061	0.090	0.929	0.069	0.074
50	0.951	0.094	0.098	0.992	0.123	0.124
100	0.999	0.100	0.100	1.000	0.132	0.132
Mean	0.463	0.033	0.071	0.877	0.036	0.041

What I have not talked about:

- The intruder's:
 - Motivation
 - Data
- Spontaneous recognition
- Consequences of disclosure
- Future Internet and other aspects of the socio-technical Zeitgeist

Conclusions

- Complete anonymisation is
 - best defined as the absence of disclosure risk
 - a theoretical state
- Pragmatic anonymisation
 - Low levels of disclosure risk
 - Needs to refer to