Idea and Notation
Disclosure Risk Measure for Population Based Tables
Disclosure Risk Measure for Sample Based Tables
Models to Estimate Population Frequencies
Numerical Results
Summary

# Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables

L. Antal N. Shlomo M. Elliot laszlo.antal@postgrad.manchester.ac.uk

University of Manchester

Work Session on Statistical Data Confidentiality
Helsinki, 5 October 2015



- Idea and Notation
- 2 Disclosure Risk Measure for Population Based Tables
- 3 Disclosure Risk Measure for Sample Based Tables
- 4 Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model
- 5 Numerical Results
- 6 Summary



#### Idea and Notation

Disclosure Risk Measure for Population Based Tables Disclosure Risk Measure for Sample Based Tables Models to Estimate Population Frequencies Numerical Results Summary

- Idea and Notation
- Disclosure Risk Measure for Population Based Tables
- Disclosure Risk Measure for Sample Based Tables
- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model
- 5 Numerical Results
- Summary



## **Idea** and Notation

- We would like to measure the disclosure risk of sample based frequency tables
- A disclosure risk measure will be developed on the basis of information theoretical expressions

#### Notation

- Frequency table:  $F = (F_1, F_2, \dots, F_K)$
- Population size:  $N = \sum_{i=1}^{K} F_i$
- Sample based table:  $f = (f_1, f_2, ..., f_K)$
- Sample size:  $n = \sum_{i=1}^{K} f_i$
- Set of individuals: I
- Set of sampled individuals: I<sub>S</sub>
- Set of table cells (categories):  $C = \{c_1, c_2, \ldots, c_K\}$

Idea and Notation

Disclosure Risk Measure for Population Based Tables
Disclosure Risk Measure for Sample Based Tables

Models to Estimate Population Frequencies Numerical Results Summary

- Idea and Notation
- 2 Disclosure Risk Measure for Population Based Tables
- Disclosure Risk Measure for Sample Based Tables
- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model
- 5 Numerical Results
- 6 Summary



#### Disclosure Risk Measure for Population Based Tables

Disclosure Risk Measure for Sample Based Tables Models to Estimate Population Frequencies Numerical Results Summary

#### Two Random Variables

#### Categorization of individuals into table cells

Categorization of all individuals

$$X:I\to C$$

Categorization of sampled individuals

$$Y: I_S \rightarrow C$$

#### Disclosure Risk Measure for Population Based Tables

Disclosure Risk Measure for Sample Based Tables Models to Estimate Population Frequencies Numerical Results Summary

## **Entropy** and Conditional Entropy

#### Entropy

$$H(X) = -\sum_{i=1}^{K} Pr(X = c_i) \cdot \log Pr(X = c_i)$$

#### Conditional Entropy

$$H(X|Y) = -\sum_{i=1}^{K} Pr(Y = c_i) \cdot \sum_{i=1}^{K} Pr(X = c_i|Y = c_j) \cdot \log Pr(X = c_i|Y = c_j)$$

$$0 \leq H(X|Y) \leq H(X)$$



#### Disclosure Risk Measure for Population Based Tables

Disclosure Risk Measure for Sample Based Tables Models to Estimate Population Frequencies Numerical Results Summary

# Disclosure Risk Measure for Population Based Tables

Disclosure risk measure:

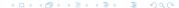
$$R_1(F, \textbf{\textit{w}}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}$$

where

 $\mathbf{w} = (w_1, w_2, w_3)$  is a vector of weights, D is the set of zeroes in the population based table, e is the base of the natural logarithm



- Idea and Notation
- Disclosure Risk Measure for Population Based Tables
- 3 Disclosure Risk Measure for Sample Based Tables
- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model
- 5 Numerical Results
- 6 Summary



# Disclosure Risk Measure for Sample Based Tables

Disclosure risk measure:

$$R_{2}(F, f, \mathbf{w}) = w_{1} \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + w_{2} \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) - w_{3} \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}$$

where

*E* is the set of zeroes in the sample based table

$$R_2(F,f,\boldsymbol{w}) \leq R_1(F,\boldsymbol{w})$$

Log-linear Model

- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model



- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model



# Log-linear Model

- There might be sample zeroes that are not zeroes in the population based table
- Sample based tables might not reflect cell probabilities well
- Log-linear models, applied to samples based tables, provide better estimates of cell probabilities
- In two-dimension: only one model that is not saturated

$$\frac{n_{i\bullet}\cdot n_{\bullet j}}{n}$$

- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model



# Pólya Urn Model

- Balls in an urn
- f<sub>1</sub> balls of colour 1, f<sub>2</sub> balls of colour 2, etc.
- $\theta$  black balls, where  $\theta$  is a parameter
- In each step we draw a ball from the urn
- If the ball is coloured, then we replace it and add a new ball of the same colour to the urn
- If the ball is black, then we replace it and add a ball of a new colour to the urn
- New colours compensate for sample zeroes



### Estimation of $\theta$

Number of cells that are zeroes in the sample based table but positive in the population based table:

$$|E| - |D|$$

Introduce

$$W_z = \left\{ egin{array}{ll} 1 & ext{if the } z ext{th draw is a black ball} \\ 0 & ext{if the } z ext{th draw is a coloured ball} \end{array} \right.$$

• We obtain  $\theta$  by solving the following equation (numerically):

$$|E| - |D| = \sum_{z=1}^{N-n} E(W_z) = \sum_{z=1}^{N-n} \frac{\theta}{n + \theta + z - 1}$$

- Idea and Notation
- Disclosure Risk Measure for Population Based Tables
- Disclosure Risk Measure for Sample Based Tables
- Models to Estimate Population Frequencies
  - Log-linear Model
  - Pólya Urn Model
- 5 Numerical Results
- Summary



#### Data

- Data: extract from 2001 UK census data
- 10 selected output areas
- Output area  $\times$  religion table: K = 90 cells. N = 2449
- Generated and real data
- 1000 samples, 1000 estimated population based tables for each sample
- Original disclosure risk: average of 1000 values
- Estimated disclosure risk: average of 1000 · 1000 = 10<sup>6</sup> values

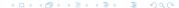
Idea and Notation Disclosure Risk Measure for Population Based Tables Disclosure Risk Measure for Sample Based Tables Models to Estimate Population Frequencies Numerical Results Summary

## **Numerical Results**

Generated		Original disc. risk		Log-linear model		Pólya urn model	
and real data		$R_2(F, f, (0.1, 0.8, 0.1))$		$R_2(\hat{F}, f, (0.1, 0.8, 0.1))$		$R_2(\hat{F}, f, (0.1, 0.8, 0.1))$	
•	Sampling fr.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Generated table	0.1	0.1538	0.0043	0.1568	0.0039	-	-
(log-linear m.)	0.05	0.1427	0.0059	0.1416	0.0054	-	-
Generated table	0.1	0.1694	0.0049	-	-	0.1758	0.0053
(Pólya urn m.)	0.05	0.1535	0.0061	-	-	0.1640	0.0057
Real	0.1	0.1697	0.0048	0.1715	0.0173	0.1764	0.0186
table	0.05	0.1535	0.0061	0.1731	0.0254	0.1821	0.0283

Table: Results of disclosure risk measures on generated and real population based tables

- - Log-linear Model
  - Pólya Urn Model
- Summary



## Summary

- A disclosure risk measure for population based tables has been extended to measure the disclosure risk of sample based tables
- Two models have been used to estimate population frequencies
- The results show relatively good estimates of the disclosure risk
- Further research should be done to measure the disclosure risk of higher dimensional frequency tables

# Thank you for your attention!