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

Outline

Outline

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: $f = (f_1, f_2, \dots, f_K)$
- Sample size: $n = \sum_{i=1}^{K} f_i$
- Set of individuals: I
- Set of sampled individuals: I_S
- Set of table cells (categories): $C = \{c_1, c_2, \dots, c_K\}$

Outline

Two Random Variables

Categorization of individuals into table cells

- Categorization of all individuals

$$X : I \to C$$

- Categorization of sampled individuals

$$Y : I_S \to C$$

Entropy and Conditional Entropy

Entropy

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

Conditional Entropy

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

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

Disclosure Risk Measure for Population Based Tables

Disclosure risk measure:

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

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

Outline

Disclosure Risk Measure for Sample Based Tables

Disclosure risk measure:

$$R_2(F, f, \boldsymbol{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} +$$
$$w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}$$

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

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

Outline

1. 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

Outline

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

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

Outline

Pólya Urn Model

- Balls in an urn
- f_1 balls of colour 1, f_2 balls of colour 2, etc.
- θ black balls, where θ 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 θ

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

$$|E| - |D|$$

- Introduce

$$W_z = \begin{cases} 1 & \text{if the } z\text{th draw is a black ball} \\ 0 & \text{if the } z\text{th draw is a coloured ball} \end{cases}$$

- We obtain θ by solving the following equation (numerically):

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

Outline

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 $1000 \cdot 1000 = 10^6$ values

Numerical Results

Generated and real data		Original disc. risk $R_2(F, f, (0.1, 0.8, 0.1))$		Log-linear model $R_2(\hat{F}, f, (0.1, 0.8, 0.1))$		Pólya urn model $R_2(\hat{F}, f, (0.1, 0.8, 0.1))$	
	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

Outline

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!