Disclosure Risk Measurement with Entropy in Sample Based Frequency Tables

L. Antal  N. Shlomo  M. Elliot

lazslo.antal@postgrad.manchester.ac.uk

University of Manchester

New Techniques and Technologies for Statistics
10 March 2015
1. Idea and Notation

2. Disclosure Risk Measures

3. Results
Outline

1. Idea and Notation
2. Disclosure Risk Measures
3. Results
Idea and Notation

We would like to measure the disclosure risk of a population based frequency table.

Information theoretical expressions (e.g. entropy) can reflect the properties of attribute disclosure.

Notation

- Population based frequency table: $F = (F_1, F_2, \ldots, F_K)$
- Population size: $N = \sum_{i=1}^{K} F_i$
- Sample based frequency table: $f = (f_1, f_2, \ldots, f_K)$
- Sample size: $n = \sum_{i=1}^{K} f_i$
Outline

1. Idea and Notation
2. Disclosure Risk Measures
3. Results
Properties of a desired disclosure risk measure

Properties:

- If only one cell is populated in the table, then the disclosure risk should be high.
- Uniformly distributed frequencies imply low risk.
- The smaller the cells, the higher the disclosure risk.
- The more number of zeroes, the higher the disclosure risk.
- The disclosure risk bounded by 0 and 1.
The Disclosure Risk Measure

- We developed the disclosure risk measure for population based frequency tables first.
- Now we extend it for sample based frequency tables.

The disclosure risk measure for population based frequency tables:

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

where \( D \) is the set of zeroes in \( F \) and \( w = (w_1, w_2, w_3) \) is a vector of weights.
Disclosure Risk Measure for Sample Based Tables

The disclosure risk of a sample based table should be lower than that of the original population based table.

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

where \(E\) is the set of zeroes in the sample based table and \(H(X|Y)\) is the conditional entropy of the original table with respect to the sample based table.
1. Idea and Notation

2. Disclosure Risk Measures

3. Results
Data: 2001 UK census tables
10 selected output areas
\( N = 2449 \)
Weights: \( \mathbf{w} = (0.1, 0.8, 0.1) \)
Initial population based table: output area (10 output areas) × religion
1,000 sample based tables, 1,000 estimated population based frequency tables for each sample based table
Estimation of population based frequency tables:
- Drawing samples from a population based table
- Applying a log-linear model to the sample based tables to estimate population parameters
- Drawing $N - n$ 'individuals’ from a multinominal distribution
- Adding the individuals to the sample based table
## Results

<table>
<thead>
<tr>
<th></th>
<th>Sampling fraction</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>$R_1(F, w)$</td>
<td>From true population frequencies</td>
<td>0.2315</td>
<td>0.2315</td>
</tr>
<tr>
<td></td>
<td>From estimated population frequencies</td>
<td>0.2173</td>
<td>0.2169</td>
</tr>
<tr>
<td>$R_2(F, f, w)$</td>
<td>From true population frequencies</td>
<td>0.1697</td>
<td>0.1533</td>
</tr>
<tr>
<td></td>
<td>From estimated population frequencies</td>
<td>0.1543</td>
<td>0.1400</td>
</tr>
</tbody>
</table>

**Table:** Table: output area (10 output areas) $\times$ religion. 1,000 samples, 1,000 estimated population based table for each sample.
A disclosure risk measure has been extended to sample based tables.

The disclosure risk measure is based on information theory.

Initial results show good estimates for a two-dimensional table. The model needs to be explored for higher dimensional tables.
Thank you for your attention!