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

Laszlo Antal*, Natalie Shlomo*, Mark Elliot*

* University of Manchester, UK, email: laszlo.antal@postgrad.manchester.ac.uk, natalie.shlomo@manchester.ac.uk, mark.elliot@manchester.ac.uk

Abstract. We extend a disclosure risk measure defined for population based frequency tables to sample based frequency tables. The disclosure risk measure is based on information theoretical expressions, such as entropy and conditional entropy, that reflect the properties of attribute disclosure. To estimate the disclosure risk of a sample based frequency table we need to take into account the underlying population and therefore need both the population and sample frequencies. However, population frequencies might not be known and therefore they must be estimated from the sample. We consider two probabilistic models, a log-linear model and a so-called Pólya urn model, to estimate the population frequencies. Numerical results suggest that the Pólya urn model may be a feasible alternative to the log-linear model for estimating population frequencies and the disclosure risk measure.

1 Introduction

Statistical agencies measure the disclosure risk before releasing statistical outputs, such as frequency tables. This work discusses how information theoretical definitions, such as entropy and conditional entropy, can be employed to measure the disclosure risk in two-dimensional sample based frequency tables. A similar approach has been followed and a disclosure risk measure has been introduced in [1] for population based frequency tables. However, there has been no attempt to employ a similar disclosure risk measure to sample based tables. In this paper we show that the disclosure risk measure can be applied to two-dimensional sample based tables. The disclosure risk measure reflects the properties of attribute disclosure properly as set out in [1].

The population from which a sample is drawn may be known or unknown to the statistical agency. The disclosure risk assessment of a sample based table is more straightforward in the former case. If the population is unknown, the population frequencies can be estimated from the sample. We then use the estimated population based table to estimate the disclosure risk of the sample based table.

The outline of the paper is as follows. In Section 2 we introduce the notation we follow throughout the paper. Section 3 describes how the entropy and the conditional entropy can be applied to assess disclosure risk in tabular data. Section 4 presents the disclosure risk measure. Section 5 proposes two models for estimating the population frequencies and the disclosure risk measure when the population is unknown. A simulation study with numerical results can be found in Section 6, followed by a conclusion in Section 7.

2 Notation

The frequency tables we deal with have K cells. Table cells are denoted $C = \{c_1, c_2, \ldots, c_K\}$. The (potentially unknown) population based frequencies are $F = (F_1, F_2, \ldots, F_K)$, and their sample based counterparts are denoted $f = (f_1, f_2, \ldots, f_K)$. The population size and the sample size are $N = \sum_{i=1}^{K} F_i$ and $n = \sum_{i=1}^{K} f_i$, respectively. The set of individuals of the population is $I = \{a_1, a_2, \ldots, a_N\}$. The set of sampled individuals, denoted by $I_S = \{b_1, b_2, \ldots, b_n\}$, is a subset of the population, $I_S \subseteq I$.

In order to present our results, we need to introduce two random variables. The variables, X and Y, provide the classification of individuals into table cells for the whole population (X) and for the sampled individuals (Y).

$$X : I \to C \ ,$$
$$Y : I_S \to C \ .$$

X is an extension of Y in the following sense. If we restrict X to I_S, then we will get Y, since $I_S \subseteq I$ and an individual in I_S is classified in the same table cell by X and Y. Note that X is not always known in practice.

Denote the distribution of X by $P = (p_1, p_2, \ldots, p_K) = \left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N}\right)$, while that of Y by $Q = (q_1, q_2, \ldots, q_K) = \left(\frac{f_1}{n}, \frac{f_2}{n}, \ldots, \frac{f_K}{n}\right)$.

Estimated population frequencies are referred to as $\hat{F} = (\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_K)$.

3 Entropy and conditional entropy

The basis of the proposed disclosure risk measure is the entropy. The entropy of X is given as follows.

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

Note that $H(X)$ is never negative. It takes its maximum value if (and only if) F is uniform. The maximum value is $\log K$.

The entropy of Y may be defined similarly. Since (1) depends only on the F table, we sometimes refer to $H(X)$ as the entropy of F.

The conditional entropy of two variables also has an important role in our disclosure risk measure. Since the domain of X and that of Y are different, the conditional entropy of X and Y cannot be defined directly. In order to calculate the conditional entropy, we modify the variables.

First we define a new set of (imaginary) individuals, denoted by $\tilde{I}$ as follows. If we multiply F by n and f by N, then we get the $n \cdot F = (n \cdot F_1, n \cdot F_2, \ldots, n \cdot F_K)$ and $N \cdot f = (N \cdot f_1, N \cdot f_2, \ldots, N \cdot f_K)$ frequency tables. Note that the entropy of F is equal to that of $n \cdot F$ and the entropy of f is the same as that of $N \cdot f$. It is easy to see that $\sum_{i=1}^{K} n \cdot F_i = \sum_{i=1}^{K} N \cdot f_i = n \cdot N$. Therefore $n \cdot N$ imaginary individuals contribute to each table. We assume that the same imaginary individuals contribute to the two tables. This set of individuals is $\tilde{I}$. The two variables are

$$\tilde{X} : \tilde{I} \to C$$

and

$$\tilde{Y} : \tilde{I} \to C \ .$$

The conditional entropy, defined below, depends on the $Pr(\tilde{X} = c_i | \tilde{Y} = c_j)$ conditional probabilities. We have not defined the probabilities unambigously, since we have to define for each imaginary individual where the individual falls by both $\tilde{X}$ *and* $\tilde{Y}$.

We assume that $\tilde{X}$ and $\tilde{Y}$ are as 'similar' to each other as possible. This assumption means that the maximum possible number of individuals fall into the same category by $\tilde{X}$ and $\tilde{Y}$. For instance, if $n \cdot F_1 \leq N \cdot f_1$, then the $n \cdot F_1$ imaginary individuals that fall in c_1 by $\tilde{X}$ also fall in c_1 by $\tilde{Y}$. This assumption reduces the number of possible $(\tilde{X}, \tilde{Y})$ pairs. Instead of selecting one of the possible pairs, we use the average $Pr(\tilde{X} = c_i | \tilde{Y} = c_j)$ conditional probabilities over the possible pairs in order to define the conditional entropy in (2). More details can be found in [1].

We define the conditional entropy of X and Y as follows.

$$H(X|Y) = H(\tilde{X}|\tilde{Y}) =$$
$$- \sum_{j=1}^{K} Pr(\tilde{Y} = c_j) \cdot \sum_{i=1}^{K} Pr(\tilde{X} = c_i | \tilde{Y} = c_j) \cdot \log Pr(\tilde{X} = c_i | \tilde{Y} = c_j) \qquad (2)$$

The conditional entropy is always smaller or equal to the entropy, $H(\tilde{X}|\tilde{Y}) \leq H(\tilde{X}) = H(X)$.

4 The disclosure risk measure

4.1 The disclosure risk measure for population based frequency tables

The disclosure risk measure, which has been introduced for population based frequency tables, is a weighted average as follows.

$$R_1(F, \boldsymbol{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \tag{3}$$

Here D is the set of zeroes in the population based table, therefore $|D|/K$ is the proportion of zeroes. $\boldsymbol{w} = (w_1, w_2, w_3)$ is a vector of weights, $w_i \geq 0$, $i = 1, 2, 3$, $\sum_{i=1}^{3} w_i = 1$.

4.2 The disclosure risk measure for sample based frequency tables

While population based tables include every individual, only selected individuals contribute to sample based frequency tables. Sampling can be considered as a special statistical disclosure control (SDC) method. The smaller number of individuals in sample based tables ensures protection against attribute disclosure to a certain extent. An intruder faces more uncertainty in a sample based table than in a population based table. Zeroes in a sample based table seemingly increase the chance of attribute disclosure. However, a zero in a sample based table is not necessarily zero in the population based table.

A disclosure risk measure for sample based frequency tables (f) is as follows.

$$\begin{aligned} R_2(F, f, \boldsymbol{w}) &= w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + \\ & w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \end{aligned} \tag{4}$$

Here E is the set of zeroes in the sample based table and e is the base of the natural logarithm. The above disclosure risk measure was developed for perturbed population based frequency tables. Since sampling can be considered as an SDC method, the formula can be applied directly to sample based tables. Note that the power of the first term reduces as follows since in our case $D \subseteq E$.

$$\frac{|D \cup E|}{|D \cap E|} = \frac{|E|}{|D|} .$$

We assume that the population size (N) is known to the statistical agency, therefore the third term of the above formula can be calculated with ease. Our aim is to estimate $H(X)$, $H(X|Y)$ and $|D|$ from the sample based table when the population frequencies are unknown. In this paper, this aim is achieved by estimating population frequencies. From the estimated population frequencies the above mentioned

quantities can be calculated as for known population based tables. The population frequencies may be estimated from the sample by probabilistic models.

5 Models to estimate population frequencies

We present numerical results in Section 6 using two modelling approaches for estimating population frequencies, a log-linear model approach and a so-called Pólya urn model approach. The results are derived from generated and real population based tables in order to assess the estimation error arising from the sampling in the first case and from the sampling and estimation of population parameters in the second case.

5.1 Log-linear model approach

A sample based frequency table may contain zero cells that have positive values in the population based table due to the random sampling. Therefore cell probabilities might not be reflected properly in a sample based table. Log-linear models can compensate for sample-based (random) zero cells and introduce positive cell probabilities by taking the table structure into account. On the other hand, log-linear models can also estimate positive cell probabilities when there should be a true population (structural) zero.

We apply a log-linear model to two-dimensional (sample based) frequency tables. In this situation we can only include main effects in the mode which will have the effect of estimating positive cell values even for those cells that are true (structural) zeroes in the population.

Denote the sum of row i by $n_{i\bullet}$ and that of column j by $n_{\bullet j}$. The expected cell count under the log-linear model is

$$\hat{\mu}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \ .$$

Dividing the above formula by n provides (estimated) cell probabilities $\hat{p}_{ij} = \frac{\hat{\mu}_{ij}}{n}$.

5.2 Pólya urn model approach

The urn model has been employed in [8] to estimate population uniques in a frequency table. Now we use a similar model to estimate all population frequencies.

The model starts with positive sample based frequencies. The frequencies are represented by coloured balls in an urn. The urn contains $f_1 > 0$ balls of colour 1, $f_2 > 0$ balls of colour 2, etc. In addition to the coloured balls, θ black balls are also placed into the urn, where θ is a parameter to be estimated. In each step we draw a ball from the urn. If it is a coloured ball, then we replace it and add a new ball of the same colour to the urn. If the ball we draw is black, then the ball is replaced and another of a new colour is placed into the urn. The balls of new colours account for sample zeroes.

In our case there might be true zeroes in the sample based table, therefore we do not assume that all sample frequencies are positive. However, zeroes do not influence the estimated population frequencies.

The estimation of the θ parameter has an impact on the number of newly introduced frequencies. The number of zeroes in the population based table plays an important role in the estimation of θ and in our disclosure risk measure with respect to the first term in (4). A high θ might result in a large number of new frequencies, therefore the number of zeroes in the population based table might be underestimated. Similarly, a low θ might imply a high number of population zeroes. We determine θ according to the number of zeroes in the population based table.

First assume that $|D|$ is known. The number of cells that are zeroes in the sample based table but positive in the population based table is $|E| - |D|$. Denote W_z, $z = 1, 2, \ldots, N - n$, an indicator variable as follows.

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

The expected number of new colours is $E(\sum_{z=1}^{N-n} W_z) = \sum_{z=1}^{N-n} E(W_z)$. The total number of balls before the zth draw is $n + \theta + z - 1$. Since the number of black balls is constant at θ, therefore $E(W_z) = \dfrac{\theta}{n + \theta + z - 1}$. We obtain θ by solving the following equation (numerically):

$$|E| - |D| = \sum_{z=1}^{N-n} \frac{\theta}{n + \theta + z - 1} \ . \tag{5}$$

Assume now that $|D|$ is unknown. In order to use (5), we need to estimate $|D|$ from the sample based table. Section 9.8 of [2] provides expected frequencies of frequencies. The expected number of zeroes is given by the following formula.

$$\widehat{|D|} = \sum_{i=1}^{K} (1 - p_i)^N \ ,$$

where p_i is the probability of cell c_i. We estimate p_i, $i = 1, 2, \ldots, K$ by applying an independent log-linear model to the sample based table.

Therefore, (5) can be rewritten as follows.

$$|E| - \widehat{|D|} = \sum_{z=1}^{N-n} \frac{\theta}{n + \theta + z - 1} \ . \tag{6}$$

We can solve (6) numerically to obtain the estimate $\hat{\theta}$.

6 Simulation study

In this section we present results of a simulation study to assess the estimation error of the disclosure risk measure in (4). We use a real population based table and a table that is generated according to known model parameters estimated from the real table. The aim is to assess the estimation error arising from sampling alone and the estimation error arising from both sampling and estimated model parameters.

6.1 Data

The dataset we used is an extract from the 2001 UK census data. The dataset consists of $N = 2449$ individuals of 10 selected output areas. The output area (10 output areas) $\times$ religion two-dimensional table has $K = 90$ cells. The frequencies are shown in Table 1.

181	0	0	1	17	1	1	83	18
138	2	4	2	0	0	1	36	16
130	0	0	0	22	4	1	61	40
173	0	0	1	14	4	1	97	22
142	2	5	0	15	6	1	37	21
129	0	0	0	0	0	1	69	20
118	2	0	2	24	9	1	38	20
130	0	0	0	34	1	1	82	32
148	3	0	0	0	2	1	38	21
136	1	2	0	13	0	0	55	16

Table 1: Original frequency table

To obtain the generated population table for assessing the log-linear model approach, we applied the log-linear model with main effects on Table 1. The estimated cell probabilities, denoted by $(\hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$, were then used as the parameters of a multinomial distribution. We drew N individuals from $Multinom(N; \hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$. When assessing the estimation error arising from sampling alone, we use these same parameters $(\hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$ for estimating the disclosure risk measure.

To obtain the generated population table for assessing the Pólya urn model approach, we use θ given in (5) to generate the population based frequencies from the sample based frequencies.

6.2 Simulation method

For the simulation study, we drew 1000 simple random samples from the (original or generated) population using two sample fractions of 0.1 and 0.05. $R_2(F, f, \boldsymbol{w})$ can be calculated on the (original or generated) population based table for each of the sample based tables. The average of the $R_2(F, f, \boldsymbol{w})$ values is considered as the

'original disclosure risk'. For this simulation, we use the following weights in the disclosure risk measure: $\boldsymbol{w} = (0.1, 0.8, 0.1)$.

When population frequencies are assumed unknown, we need to estimate them from the sample based table. In the log-linear approach, for the case of the generated population table with known parameters, we estimate the population frequencies by drawing $N - n$ individuals from $Multinom(N - n; \hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$ and adding these frequencies to the respective sample-based table. For the case of the real population table, we estimate the population frequencies by applying the log-linear model with main effects to the sample-based table (f). The resulting table provides estimated cell probabilities. Denote them by $(\hat{q}_1^S, \hat{q}_2^S, \ldots, \hat{q}_K^S)$, where the superscript S refers to the sample. $N - n$ individuals are drawn from $Multinom(N - n; \hat{q}_1^S, \hat{q}_2^S, \ldots, \hat{q}_K^S)$, and are then added to the sample-based table, thereby estimating the population frequencies.

In the Pólya urn approach, for the case of the generated population table we use θ given by (5) and for the case of the real population table, we estimate θ on each of the sample based tables as defined in (6).

The simulation is carried out as follows for each sample fraction and for each (original or generated) population. On each of the 1000 sample-based tables we estimate the population frequencies 1000 times. For each estimated population-based table ($\hat{F}$) of each sample-based table we obtain an estimated disclosure risk measure ($R_2(\hat{F}, f, \boldsymbol{w})$). Note that the overall number of the $R_2(\hat{F}, f, \boldsymbol{w})$ values is equal to $1000 \cdot 1000$. The average of the $R_2(\hat{F}, f, \boldsymbol{w})$ values is considered as the final 'estimated disclosure risk'.

6.3 Numerical results

Table 2 presents the results of the simulation study using both the generated and real population based tables and two sampling fractions 0.1 and 0.05. We compare the 'original disclosure risk' with the 'estimated disclosure risk'. The weights for the disclosure risk measure are $\boldsymbol{w} = (0.1, 0.8, 0.1)$.

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

For the log-linear model, using the generated population based table with known parameters under the log-linear model, we see that we can obtain close estimates to

the original disclosure risk measures when only sampling error is considered. The estimated disclosure risk measure based on the real population table is slightly higher than the original disclosure risk. The overestimation is worse for the smaller sample fraction.

The Pólya urn modelling approach provides only slightly less accurate estimates than the log-linear modelling approach but there appears to be overestimation both in the generated table and the real population table.

7 Conclusion

In this paper, we present an information theoretical based disclosure risk measure for two-dimensional sample based tables. Under the generated population based table with known parameters, the disclosure risk can be estimated accurately and therefore the estimation error arising from the sampling alone appears to be unbiased. However, the estimated disclosure risk for a real population based table where we need to account for the estimating of the parameters from the sample based table is less accurate. The Pólya urn model approach is a feasible alternative to the log-linear model approach. Further reseach needs to be carried out in order to provide a more accurate approximation of the disclosure risk using different size tables with varying sampling fractions and levels of random and true zero cells in the population. In addition, further research is needed to explore the estimation of disclosure risk in higher dimensional tables.

Acknowledgements

This work was funded by the ONS-ESRC PhD studentship (Ref. ES/J500161/1).

References

[1] Antal, L. and Shlomo, N. and Elliot, M (2014) "Measuring Disclosure Risk with Entropy in Population Based Frequency Tables", *Privacy in Statistical Databases*, 62–78, Springer.

[2] Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (2007) "Discrete Multivariate Analysis: Theory and Practice", Springer.

[3] Bunge, J. and Fitzpatrick, M. (1993) "Estimating the Number of Species: a Review", *Journal of the American Statistical Association*, **88/421**, 364–373.

[4] Cover, T. M. and Thomas, J. A. (2006) "Elements of Information Theory", 2nd. ed., Wiley, Hoboken.

[5] Goodman, L. A. (1949) "On the Estimation of the Number of Classes in a Population", *The Annals of Mathematical Statistics*, **20/4**, 572–579.

[6] Haas, P. J., Naughton, J. F., Seshadri, S. and Stokes, L. (1995) "Sampling-Based Estimation of the Number of Distinct Values of an Attribute", *Proceedings of the 21th International Conference on Very Large Data Bases*, 311–322.

[7] Hoppe, F. M. (1984) "Pólya-Like Urns and the Ewens' Sampling Formula", *Journal of Mathematical Biology*, **20/1**, 91–94.

[8] Samuels, S. M. (1998) "A Bayesian Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment", *Journal of Official Statistics*, **14**, 373–384.

[9] Shlomo, N. (2007) "Statistical Disclosure Control Methods for Census Frequency Tables", *International Statistical Review*, **75**, 199–217.

[10] Skinner, C. J. and Shlomo, N. (2012) "Estimating Frequencies of Frequencies in Finite Populations", *Statistics & Probability Letters*, **82/12**, 2206–2212.

[11] Willenborg, L. and de Waal, T. (2001) "Elements of Statistical Disclosure Control", Lecture Notes in Statistics, Springer