Smoothing Contingency Tables To Estimate a Global Risk Measure

Daniela Ichim

Istituto Nazionale di Statistica, via C. Balbo 16, Rome, Italy.
(ichim@istat.it)

Abstract. In this paper a global measure of the re-identification risk in microdata files is analyzed. A penalized maximum likelihood approach is described. Mainly the problem of smoothing two-way contingency tables will be addressed; further possible developments are indicated. The methodology was applied to data stemming from the Italian 2001 Census and the Labour Force survey.

1 Introduction

To face the increasing demand from users, the National Statistical Institutes (NSI) disseminate more often microdata files. Such dissemination should be constrained to the confidentiality pledge under which a statistical agency collects survey data. To protect the confidentiality of respondents, a statistical disclosure control (SDC) methodology is generally applied. This methodology may be divided in two main parts. In a first stage, with respect to an adopted disclosure scenario, the risk of disclosure of each unit is assessed/estimated. Then, a masking method is applied to guarantee that no confidential information about respondents could be retrieved from the disseminated microdata file. This paper addresses only the first problem: the disclosure risk assessment. Moreover, the risk of disclosure is here defined as the risk of re-identification.

After the removal of direct identifiers, e.g. name and address, other indirect identifiers, called key variables, could still allow the re-identification of a unit. Usually, most of the key variables registered in social microdata files are categorical. Particular values taken by variables like place of residence, gender, age, citizenship, and marital status could correspond to a unique person in the population. Therefore, the risk of re-identification for such data is estimated by means of rareness concepts, see, for example, [2] and [4]. In this work it is assumed that the key variables are all categorical.

This paper is divided in three parts. In section 2 the framework used for the re-identification risk estimation and its link to the log-linear models is introduced. In section 3 a smoothing strategy, the penalised likelihood approach, is discussed as a technique to estimate a disclosure risk measure in contingency tables. The penalised likelihood methodology was applied to simulated and real data. In section 4, several results are illustrated. In section 5 conclusions are drawn and some further possible developments are indicated.

2 Measures of Disclosure Risk

Before microdata file dissemination, the NSI generally make assumptions on the tools an intruder might use to breach the confidentiality of respondents. It is usually assumed that the intruder may access some external database containing direct identifiers. It is further assumed that the intruder would use the shared variables as comparison variables in a matching experiment. There are many implicit assumptions in this disclosure scenario. Many facets of this scenario were previously discussed in literature, see, for example, [7] and [10]. The NSIs commonly quantify the disclosure risk by means of the re-identification risk, that is, the probability of a correct match, see [10].

As the units sharing the same values for all the categorical variables have the same re-identification risk, see [4] and [7], the key variables are cross-classified; a contingency table with K cells is then derived. Obviously, the re-identification risk depends on both the population and sample frequencies of these cells. Let F_k denote the population frequency and let f_k denote the sample frequency of the k-th cell, $k = 1, \ldots, K$. The usage of only sample frequencies is not sufficient because the risk could be overestimated. The global measure of risk discussed in this paper is the number of sample uniques that are also population uniques. Following the approach in [10], this risk measure may be written as:

$$\tau_1 = \sum_{k=1}^{K} \mathbb{I}(F_k = 1, f_k = 1)$$

τ_1 cannot be directly computed because it depends on the unknown population frequencies F_k. Some modelling assumptions are needed in order to derive an estimable expression of the global risk measure. It is generally assumed that the population frequencies are independently Poisson distributed with means λ_k. In each cell, a Bernoulli sampling scheme is assumed, with selection probability equal to π_k. It follows that the sample frequencies f_k are also independent following Poisson distributions, see [10].

Then an estimation of the global risk measure τ_1 may be expressed as in (1).

$$\hat{\tau_1} = \sum_{k=1}^{K} \exp(-\mu_k(1 - \pi_k)/\pi_k), \quad \mu_k = \pi_k \lambda_k \tag{1}$$

$\hat{\tau}_1$ depends on both the sampling fractions, π_k and the expected cells frequencies. It should be observed that the summation should be done on the sample uniques only. Moreover, for simplicity, it is assumed that the sampling fractions are equal across the cells of the contingency table, i.e. $\pi_k = \pi, k = 1, \ldots, K$.

To estimate τ_1 the relationships between the expected cell frequencies are generally modelled by means of a log-linear model including the desired main effects and interactions, see equation (2).

$$\log(\mu_k) = \boldsymbol{x}_k^{'}\boldsymbol{\beta} \tag{2}$$

The estimates are then computed by maximizing the relevant part of the log-likelihood function $\mathcal{L}(\boldsymbol{\beta}) = \sum(f_k \log(\mu_k) - \mu_k)$. Iterative algorithms like iterative proportional fitting (IPF) or Newton-Raphson may be used to maximise the likelihood $\mathcal{L}(\boldsymbol{\beta})$.

Table 1: Example of a 2 x 2 table.

a	b
c	x

3 Smoothing Contingency Tables

For large sparse tables, the likelihood could get maximized on the boundary of the parameter space and too many cells estimates might be zero. Two possible solutions are the table redesign or the addition of a flattening constant. Both solutions have their drawbacks either because they do not solve the given dissemination problem or because the sample size is artificially increased. In [1] and [3] more details on these methods are given.

A valid alternative could be the usage of parsimonious models. Anyway, in the risk estimation framework, see [8], it was observed that when a simple (independence) log-linear model is used, the estimation of μ_k would be based on information from all the cells having in common even a single characteristic.

3.1 Local Neighbourhoods

In [8], it was proposed to find a compromise between the model complexity and the quantity of information used: complicate a little bit the model, but use only the information from the neighbouring cells. Of course, the neighbourhoods may be defined only for ordinal variables. Consequently, it was supposed that a distance between cells may be defined, namely $d(k^{'}, k)$.

This approach is based on the assumption that in a certain neighbourhood, $\log(\boldsymbol{\mu})$ may be approximated by a polynomial, i.e. $\log(\boldsymbol{\mu}) = \left[\beta_0 + \ldots + \beta_t d(k^{'}, k)^t\right]$. Then, it was proposed to maximize, for each cell, the local likelihood function:

$$\mathcal{LL}(\boldsymbol{\beta}) = \sum_{k^{'} \in N^k} \left[f_{k^{'}} \left[\beta_0 + \ldots + \beta_t d(k^{'}, k)^t\right] - \exp\left(\beta_0 + \ldots + \beta_t d(k^{'}, k)^t\right)\right] \qquad (3)$$

where N^k denotes the a-priori selected neighborhood of the k-th cell.

In [8] several choices of N^k and d are presented, taking into account their possible multi-dimensionality, too. Different aspects of the local neighborhood approach were discussed in [6].

3.2 Smoothness and Independence

The main idea in the previous proposal is that the sample uniques with small values neighbouring cells are more likely population uniques. This idea could be further generalized. If smoothness is assumed, the neighbouring cells should have similar values.

Let's see what smoothness means in practice. Consider a simple 2 by 2 table, see table 1. If a, b and c have similar values, small or large, it doesn't matter, and if smoothness is assumed, the fourth value, x, should take more or less the same value. This means that the cross-ratio $\theta = \frac{ax}{bc}$ is approximately 1. Values of θ close

to 1 represent the independence of the two categorical variables, while values of θ farther from 1 represent stronger levels of association, i.e., no independence.

In the SDC framework, the cross-ratio θ could be a possible way to quantify the smoothness.

To distinguish between the sample uniques that are also population uniques and those who aren't, smoothness in the contingency table should be assumed. The maximisation of the log-likelihood function $\mathcal{L}(\boldsymbol{\beta})$ could be constrained to a smooth solution. As usual in optimisation problems, to take into account a constraint, one could penalise for missed smoothness or, equivalently, one could penalise for the missed independence. Maximizing the penalized likelihood given in equation (4) would guarantee a smooth solution.

$$\mathcal{PL}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - A\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\left[\log\left(\frac{\mu_{i,j}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}}\right)\right]^2 \tag{4}$$

where I is the number of rows and J is the number of columns. A is a penalty constant whose values might be chosen according to some statistical criteria discussed in [9], for example. The function $\mathcal{PL}$ penalizes for missed local independence in the reduced 2x2 tables since $\mathcal{PL}$ takes smaller values when the cross-ratios are much greater (smaller) than 1.

3.3 Properties of the Penalized Likelihood Function

There are several theoretical advantages of this penalized likelihood approach.

First the existence, uniqueness and consistency of the estimators were proved under general conditions, see [9].

Second, the number of parameters to be estimated is greatly reduced with respect to the methodology proposed in [8]. This means that the number of degrees of freedom is kept under control.

Third, the penalized likelihood could be extended to multi-dimensional tables. It is sufficient to use an expression of independence in multi-dimensional tables. For example, in a 3-dimensional table, the penalty could be related to $\log\left(\frac{\mu_{ijm}\mu_{i+1jm}\mu_{ij+1m}\mu_{ijm+1}}{\mu_{i+1j+1m+1}\mu_{i+1j+1m}\mu_{i+1jm+1}\mu_{ij+1m+1}}\right)$. In section 4 some results of an application of this penalty to 3-way tables will be given.

The penalized likelihood approach is an estimation method; in principle, it could be integrated with whatever model. For example, the penalized likelihood could also be integrated with the log-rate models. To obtain unbiased parameter estimates and non-misleading standard errors, the log-rate models introduced in [5] use an offset variable depending on weights, see equation (5). The advantages of using this model in the SDC framework were discussed in [6].

$$\log(\mu_k) = \log(z_k) + \boldsymbol{x}_k^{'}\boldsymbol{\beta} \tag{5}$$

where $z^k = 1/w^k$ is the inverse of the average cell weight $w^k = f_k^w/f_k$.

An interesting feature of (5) is the natural way to deal with the structural zeros. The model may be rewritten in a multiplicative form

$$\mu_k = z_k \exp(\boldsymbol{x}_k^{'}\boldsymbol{\beta})$$

and the z_k may be set equal to zero for all the structural zero cells. This formulation is important especially for the tables containing a large number of structural zeros, as the ones derived from the social surveys. An application of this methodology will be described in section 4.

The penalised likelihood approach could be extended to non-ordinal key variables. The large (full) contingency table could be divided in many 2-way reduced tables. For these reduced tables, the independence could be simply expressed by the cross-ratio. Then the penalty should be expressed in terms of independence in the reduced tables. It should be observed that it is not necessary to assume the smoothness, hence independence, for all the reduced 2-way tables. Only for a subset of such reduced tables the smoothness property might be assumed. This latter extension will be subject to further investigation.

4 Case studies

The penalised likelihood approach was applied to two different datasets. First the Italian 2001 census data was used to generate samples from which 2 and 3-dimensional contingency tables were computed. Secondly, data stemming from the Italian 2001 Labour Force Survey was used to perform some further tests on 2-dimensional tables. The algorithm was implemented in an iterative manner. The penalty constant A was always set equal to 10. The penalized likelihood approach was integrated in the parameter estimation of the independence log-rate models. The structural zeros were taken into account. In these applications, the structural zeros were defined by those cells having a zero value in the contingency table derived from the census data.

4.1 Census data

From the Italian 2001 census data, the variables *Province*, *Gender*, *Age* (14 categories) and *Education* (16 categories) were selected. Variables *Gender* and *Age* were used as stratification variables. For each province, a random municipality was selected among those having more than 500 inhabitants. For different sampling fractions varying from 0.01 to 0.9, a stratified simple random sample was selected. To preserve the population totals by *Gender* and *Age*, the weights were computed using a calibration estimator. The penalized likelihood method was firstly applied to the 2-way contingency tables defined by *Age* and *Education*. For a selection of provinces, the results are illustrated in figure 1. It may be observed that the estimated τ_1 is much closer to the real τ_1 than to the number of sample uniques. For the other sampling fractions, the same qualitative conclusion holds, except that the distance between the black, red and blue lines decreases as the sampling fraction increases. For the 3-dimensional tests on census data, the same settings were used. *Gender*, *Age* (14 categories) and *Education* (16 categories) were the categorical key variables. 200 samples were generated for each municipality and for each sampling fraction. In table 2, the results obtained in several provinces are shown. Indeed, for each province and for each sampling fraction, the percentage of times the absolute difference $|\hat{\tau}_1 - \tau_1|$ is greater than 3 is shown in the third column. Moreover, the

Table 2: Simulated tables from the Italian 2001 census. Percentage of times when the indicated criteria is satisfied.

Province	π	$\lvert\hat{\tau}_1 - \tau_1\rvert > 3$	$\left\lvert\frac{\hat{\tau}_1-\tau_1}{\tau_1}\right\rvert > 0.5$	$\min \tau_1$	$\bar{\tau_1}$	$\max \tau_1$
TRENTO	0.05	8	62	0	2.4	6
TRENTO	0.10	4	12	1	4.15	7
TRENTO	0.30	16	0	6	11.93	19
TRENTO	0.50	49	0	11	20.51	27
TRENTO	0.70	65	0	22	28.84	35
TRENTO	0.90	28	0	32	37.19	41
GENOVA	0.01	0	79	0	0.73	3
GENOVA	0.05	20	62	0	2.50	6
GENOVA	0.10	28	34	0	5.72	10
GENOVA	0.50	77	0	17	29.04	38
FIRENZE	0.01	0	100	0	0.54	3
FIRENZE	0.05	8	39	1	3.48	9
FIRENZE	0.10	12	16	1	6.61	13
FIRENZE	0.50	78	0	23	31.62	43
PALERMO	0.01	1	100	0	0.70	4
PALERMO	0.05	10	27	1	3.70	10
PALERMO	0.10	26	9	3	7.45	16
PALERMO	0.50	98	0	23	36.23	47
TORINO	0.01	2	100	0	0.63	4
TORINO	0.05	2	26	0	3.19	7

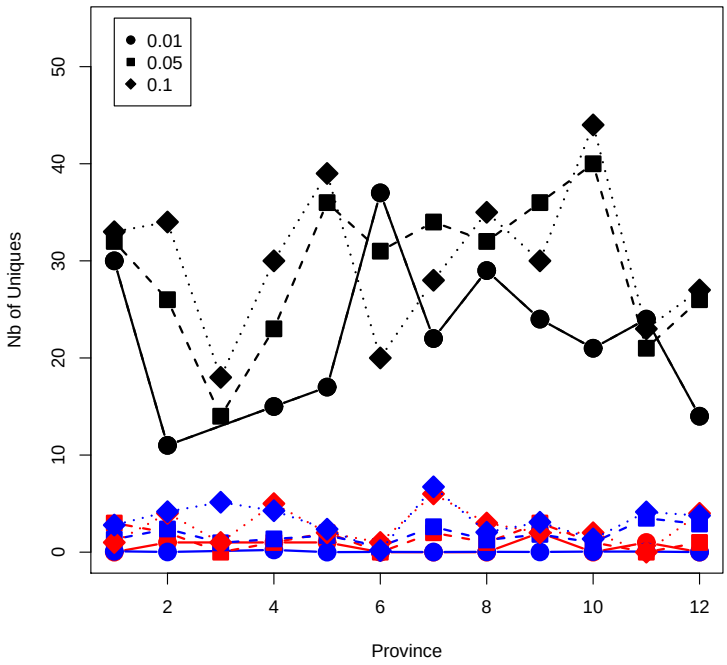

Figure 1: Results obtained on samples of census data. The black lines represent the number of sample uniques, the red lines represent the number of sample and population uniques (τ_1); the blue lines represent the estimated number of sample uniques that are also population uniques ($\hat{\tau}_1$); different symbols represent different sampling fractions.

percentage of times the absolute relative error $\left|\frac{\hat{\tau}_1-\tau_1}{\tau_1}\right|$ is greater than 0.5 is indicated in the fourth column. For brevity, only a selection of provinces and sampling fractions is shown. For the $\pi = 0.01$, the 100% values in the fourth columns are due to the extremely low values of the number of generated sample and population uniques (τ_1). The minimum, mean and maximum number of generated sample and population uniques are also given in table 2. The high percentages in the third column are associated to higher values of τ_1, so the relative error is quite low.

4.2 Labour Force Survey

A second experiment was conducted using the Labour Force Survey 2001 data. For this survey a two-stage stratified sampling was used and the applied stratification technique involved also the dimension of municipalities, a variable that could be hardly seen as a key variable in practical disclosure scenarios. For each province, the contingency tables were analyzed for each *Occupation* (10 categories) combination. *Age* (120 categories) and *Education* (8 categories) were considered as key variables. As remarked in [4], [6] or [11], this is probably the most complicated situation, when a stratification (or calibration) variable is not included among the key variables. In table 3, a selection of results is presented. Only some *Occupation*

Table 3: Labour Force Survey. τ_1 estimation by province and occupation domains. SU = number of sample uniques, SPU = number of sample uniques that are population uniques, NbIter = number of iterations used in the implementation of the penalized likelihood estimation method, Log = τ_1 estimation when the classical independence log-linear model is used.

Province	Occupation	SU	SPU	NbIter1	NbIter5	NbIter10	Log
TORINO	1	49	1	3.39	2.92	2.83	0.00
TORINO	2	88	7	13.42	12.98	9.88	0.00
TORINO	3	37	3	1.35	1.33	1.23	0.01
AOSTA	1	44	2	2.35	2.35	2.35	2.17
AOSTA	2	20	2	4.72	4.48	2.71	0.01
AOSTA	3	31	4	4.39	3.78	3.27	1.50
AOSTA	4	12	1	0.00	0.01	0.13	0.00
AOSTA	5	74	3	4.50	4.34	3.91	2.91
TRENTO	1	18	2	1.00	0.87	1.43	0.00
TRENTO	2	56	1	1.00	1.01	1.01	1.47
TRENTO	3	9	1	0.05	0.06	0.06	0.00
VENEZIA	1	16	2	0.00	0.18	0.77	0.00
VENEZIA	2	9	1	0.00	0.00	0.00	0.00
VENEZIA	3	65	9	11.49	11.17	7.56	0.00
TRIESTE	1	17	3	0.00	0.01	0.02	0.01
TRIESTE	2	56	3	3.00	2.77	2.70	0.01
TRIESTE	3	7	2	0.01	0.03	0.00	0.00
TRIESTE	4	7	1	0.00	0.16	0.15	0.00
BOLOGNA	1	16	2	0.00	0.08	0.82	0.00
BOLOGNA	2	68	3	6.44	4.39	6.16	0.00
ANCONA	1	22	1	0.63	1.32	1.38	0.00
ANCONA	2	55	5	0.22	0.18	4.06	0.28
ANCONA	3	9	2	0.00	0.00	0.03	0.00
CAMPOBASSO	2	48	6	3.64	3.06	2.78	0.22
CAMPOBASSO	3	5	1	0.00	0.08	0.17	0.00
CAMPOBASSO	4	56	3	0.19	0.13	4.26	0.08
CATANZARO	2	74	2	3.54	3.16	2.96	0.01
CAGLIARI	3	17	1	1.27	0.49	0.22	0.00
FIRENZE	1	83	5	3.26	3.48	3.68	0.19
PALERMO	2	58	4	7.15	6.96	6.72	0.14
PALERMO	4	50	2	2.03	1.59	1.32	0.11
NAPOLI	1	64	4	0.00	0.00	2.31	0.06
NAPOLI	2	78	4	5.04	4.84	3.78	0.08
GENOVA	1	55	1	0.90	0.91	0.92	0.01
GENOVA	2	39	2	3.38	2.65	2.19	0.01
GENOVA	3	76	6	6.57	6.36	6.12	0.49
PERUGIA	2	64	1	1.18	1.18	1.18	0.14
PERUGIA	3	42	1	1.46	1.44	1.38	1.46
ROMA	2	66	5	7.73	6.86	6.84	0.02

categories having at least one sample and population unique are presented. Since the penalized likelihood approach was implemented in an iterative manner, the effect of the number of iterations (NbIter) was assessed. In future testing, the algorithm will be implemented using a deviation criteria. From table 3, it may be observed that the simple independence log-linear model might not be sufficient to estimate τ_1. As expected, $\hat{\tau}_1$ generally approaches τ_1 as the number of iterations increases. When the difference between the number of sample uniques (SU) and the number of sample and population uniques (SPU) increases, the latter statement might not hold, especially when the number of SPU is very low. It should be also observed that a greater number of sample uniques is related to the table sparsity.

5 Conclusions

In the SDC framework, table smoothness is particularly important since the estimation of any disclosure risk measure might be performed by borrowing information from the neighboring cells. A penalized likelihood approach was proposed to deal with the smoothness characteristic of the tables. The penalty function was expressed in terms of independence constraints. The methodology was applied to both simulated and real datasets. In all these tests, the estimated risk value was much closer to the real risk value than to the number of sample uniques. Depending on the sampling fractions, the proposed approach to estimate τ_1 proved to be a reliable estimation method.

Since the sparsity of the contingency tables is still a problematic issue, other smoothness/independence measures will be investigated. Especially for multi-dimensional tables, usage of more suitable independence measures could improve the estimation of the global risk measure. Moreover, the estimation of other disclosure risk measures will be taken into account

Acknowledgments.

Istat is not responsible for any views or results presented. The author was supported by the Network of Excellence in the European Statistical System in the field of SDC (ESSnet-SDC).

References

[1] Agresti, A., Yang, M.: An Empirical Investigation of Some Effects of Sparness in Contingency Tables. Computational Statistics and Data Analysis **5** (1987) 9–21.

[2] Elamir, E., Skinner, C.: Record Level Measures of Disclosure Risk for Survey Microdata. Journal of Official Statistics, **22(3)** (2006) 525–539.

[3] Fienberg, S.E., Holland, P.W.: On the Choice of Flattening Constants for Estimating Multinomial Probabilities. Journal of Multivariate Analysis **2** (1972) 127–134.

[4] Franconi, L. Polettini, S.: Individual Risk Estimation in μ-ARGUS: a review. In Domingo-Ferrer, J. and Torra, V. (eds.), PSD 2004, LNCS, vol. 3050, pp. 262-272, Springer Heidelberg (2004).

[5] Haberman, S. J.: Analysis of Qualitative Data, Vol 2. New Developments. : Academic Press, New York (1979).

[6] Ichim D.: Extensions of the Re-identification Risk Measures Based on Log-linear Models. In Domingo-Ferrer, J. and Saygm, Y. (eds.), PSD 2008, LNCS, vol. 5262, pp. 203-212, Springer-Verlag Berlin Heidelberg (2008).

[7] Polettini, S.: Some Remarks on the Individual Risk Methodology. Monographs of Official Statistics. Work Session on Statistical Data Confidentiality. European Comission (2003).

[8] Rinott, Y., Shlomo, N.: A Smoothing Model for Sample Disclosure Risk Estimation. IMS Lecture NotesMonograph Series Complex Datasets and Inverse Problems: Tomography, Networks and Beyond Vol. 54 (2007) 161-171.

[9] Simonoff, J. S.: A Penalty Function Approach to Smoothing Large Sparse Contingency Tables. The Annals of Statistics **11** (1983) 208–218.

[10] Skinner, C., Holmes, D.: Estimating The Re-Identification Risk per Record in Microdata. Journal of Official Statistics **14** (1998) 361–372.

[11] Skinner, C., Shlomo, N. : Assessing Identification Risk in Survey Microdata Using Log-Linear Models. Journal of the American Statistical Association **103** (2008) 989–1001.