Individual Risk Assessment of Social Microdata

Giovanni Seri
ISTAT, Servizio della Metodologia di Base per la Produzione Statistica
via Cesare Balbo 16
00184 ROMA, ITALY
seri@istat.it

Loredana Di Consiglio
ISTAT, Servizio della Metodologia di Base per la Produzione Statistica
via Cesare Balbo 16
00184 ROMA, ITALY
diconsig@istat.it

Luisa Franconi
ISTAT, Servizio della Metodologia di Base per la Produzione Statistica
via Cesare Balbo 16
00184 ROMA, ITALY
franconi@istat.it

1. Introduction

Prior to any microdata release an assessment of disclosure risk should be carried out. The aim of this paper is to conduct further experiments (a first investigation was reported in Di Consiglio *et al*., 2003) to assess the behaviour of the individual risk of disclosure initially proposed by Benedetti *et al*. (1999). This methodology that estimates disclosure risk for social microdata is currently in the process of being implemented in the software μ-Argus as part of the CASC project. The assessment intends to reproduce, as much as possible, real instances. For such reason a subset of the 1991 Italian population census was chosen and several hundreds samples using the Labour Force Survey (LFS) design were selected. This assessment aims at investigating especially the behaviour of the methods in sample rare units. In Section 2 we outline the individual risk model. In Section 3 we describe the experiment and present some results.

2. The individual risk methodology

We outline the individual risk model used in this experiment; for further information and details see Benedetti et al. (1999), Benedetti and Franconi (1998) and recent development in Polettini (2003). The aim of the individual risk model is ordering each unit i in the random sample to be released. Units that present a value of the risk higher than a predefined threshold, the maximum tolerable risk, will be subjected to protection (e.g. local soppression); see Polettini, 2003 on ways to choose it. The way in which an intruder may identify a unit is by mean of the key variables, variables that allows identification and are publicly available. The individual risk for unit i in the sample is defined as r_i=P(i correctly linked to individual i^* in the population | sample). Under the negative binomial distribution assumption the risk can be seen as an expected value and an analytic form for it can be evaluated, see also Polettini, (2003). The risk turns out to be a function of the final weight, w_i, attached to each unit in the sample and the frequency in the population, F_k, for the combination of key variables k corresponding to unit i.

3. The experiment and results

We consider the 1991 Italian Population Census data from 4 administrative Italian regions (Val D'Aosta, Veneto, Lazio and Campania) as the source of information for the experiment. Then we sample from such regions several hundreds samples according to the LFS sampling design. This design is a two-stage design with stratification of the municipalities and systematic selection of the households. The final weight of each household is derived by means of the calibration process (Deville and Särndal, 1992) in order to preserve some known population totals. In the LFS the known totals we consider are age (in fourteen classes) by sex by region of residence. We call these variables the design variables.

We inspected several key variables, e.g. sex (2 categories), age in years (from 0 to 110), region of residence (4 selected regions in this study), position in profession (14 categories) and relationship with the head of the household (13 categories) and others.

If the intruder has the whole of the population, the probability of linking one unit in the sample to one individual in the population is equal to $1/F_k$ where k is the corresponding combination to which the unit belongs to. We call this quantity the *real risk, R*. One of the aim of this simulation study is to analyse the effectiveness of the protection procedure based on the estimates of the individual risk. To do this we compare the estimates of the risk with the corresponding real risk for different sets of key variables and design variables. When there is perfect agreement between design and key variables the method overestimates the individual risk of units with rare combination in the sample (mainly sample uniques and two cases). When the design variables is a subset of the key variables then the pattern is not so clear and underestimation as well as overestimation is present. In this paper we investigate the reasons for this and try to give advice on this matter.

Acknowledgements

We gratefully acknowledge the financial support of the European Union CASC project IST-2000-25069. The views expressed are those of the authors only and do not necessarily reflect the policies of the Istituto Nazionale di Statistica.

REFERENCES

Benedetti, R. and Franconi, L. (1998), Statistical and technological solutions for controlled data dissemination, Pre-proceedings of New Techniques and Technologies for Statistics–Sorrento, 4-6 November 1998, vol.1, 225-232.

Benedetti, R., Franconi, L. and Piersimoni, F. (1999), Per-record risk of disclosure in dependent data, Proceedings of the Conference on Statistical Data Protection, Lisbon 25-27 March 1998. European Communities, Luxembourg.

Deville, J. C., Särndal, C. E., (1992), Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, 87, pp. 367-382.

Di Consiglio, L., Franconi, L. and Seri, G. (2003), Assessing the individual risk of disclosure: an experiment, Presented at the Joint ECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003).

Polettini, S. (2003), Some remarks on the individual risk methodology, Presented at the Joint ECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003)

RÉSUMÉ

Dans cette article nous conduisons une expérience pour évaluer les performances de l'estimation du risque individuel de violation initialement proposé par Benedetti and Franconi (1998). Cette méthodologie est actuellement en train d'être appliquée dans le logiciel μ-Argus, comme partie du projet européenne CASC (Computational Aspects of Statistical Confidentiality).