Statistica Applicata Vol. 6, n. 3, 1994

SOME TECHNIQUES TO CONTROL DISCLOSURE RISK

Simone Borra, Fabio Crescenzi

ISTAT, Roma.

National Institutes of Statistics (NSis) meet an increasing demand of data. One of the most important challenge that NSis have to face is the research of a balance between data access and data protection.

The release of more and more detailed information may lead, as undesiderable consequence, to violate the individual right to privacy. The problem may arise both in the release of micro data and in the release of tabular or macro data.

The aim of this paper is to indicate new tools to control disclosure risk. The proportion of population uniques is the major factor that affects the risk. The use of a Pareto or Discrete Pareto model gives the opportunity to build an indicator based on the amount of the released information and on the concentration of the units with respect to the variablses that can be used to violate data confidentiality. The result is a direct consequence of the link between proportion of uniques and amount of information released and data concentration.

1. INTRODUCTION

The increase of data exchange requires initiatives capable to ensure a more effective respect of confidentiality rules. The strategy of releasing data anonymized by removing all obvious identifiers of respondents such as name and address, it is not sufficient to garantee from the risk of disclosure (Dalenius, 1986).

Data can be released either as an anonymized microdata sample, or as a file in tabular or aggregated form. In both cases techniques to prevent the linkage with the respondents must be used.

A major factor affecting the risk is the proportion of unique elements in the population and the procedures to estimate this proportion are extremely important.

A function that links the number of uniques to the size of the data set could be an useful tool to model the effects of reducing the geographical datail. Such a function is called prediction function (PF).

Some results on the methodology to build PF and to estimate PF parameters are shown in Crescenzi 1992a, Crescenzi 1992b and Coccia 1992.

In Crescenzi 1992b three different functions are compared and it is shown that a modification of the function based on the Negative binomial – Gamma distribution leads to better results.

In this paper we consider the use of a Zipf or Pareto Discrete model. Unfortunately these models can't be used to build PFs, but some important conclusions to improve the control of the number of unique elements can be achieved. The most relevant is that the percentage of uniques is strictly related to an indicator of the amount and to an indicator of the concentration of the released information.

2. CONTINGENCY TABLE AND SUMMARY TABLE

Let us consider some definitions on contingency tables. Both in the case of macro–data release and in the case of micro–data release the object of the study is a contingency table. If the released file is an anonimysed micro–data sample, we may refer to the contingency table built on the key– variables, (the variables that may lead to the identification by the linkage to external files). If the file is in aggregated form, the data are already set as a contingency table.

We call cell a conbination of the categories for a given set of variables, key is the set of all combinations. Our interest is to synthetize in some simple indicators the distribution of the information over the table with reference to the total information and to the number of risky cells. In this paper we regard as risky the one–unit cells, however it must be noted that in some cases cells with more than one unit have also to be considered as risky.

M is the number of non empty cells, F_i is the number of units in the cell i $(i=1,\ldots,M)$, $\sum_{i=1}^{M} F_i = N$ is the total number of units in the table, $\pi_i = F_i/N$ is the relative frequency of units in the cell i. We consider the distribution arising by grouping cells that contain the same number of units. W_1 is the number of one–unit cells, W_2 number of two–units cells, and so on. We call this distribution summary table(ST) since it contains all the relevant information on the risk associated to the contingency table. $\sum_{j=1}^{\max j} W_j = M$ is the total number of non empty cells and $\sum_{j=1}^{\max j} j \cdot W_j = N$ is the total number of units.

The study of the summary table is extremely important because it summarises the concentration of units in the contingency table. If the units are widespread, there will be many one–unit cells and the summary distribution is steeppy decreasing, if the units are more concentrated is some cells, the ST show a lower slope.

A common approach to the problem is to find a theoretical distribution to describe the summary table depending on a set of parameters. In this case the parameters will reflect all the relevant information about the contingency table. The choice of the theoretical distribution is very important. Some models were introduced in the last years, as the well known model based on the Poisson–Gamma distribution (Bethlehem et al. 1990), or the Poisson–lognormal (Skinner–Holmes 1992).

An important result is reported in Biggeri–Zannella (1991). The authors showed that some dispersion indices (especially the entropy) are adequate indicators of the potential disclosure risk and that there is a clear relationship between the percentage of uniqes and M, the number of non empty cells.

In this work we show that: i) Distributions like Zipf or Pareto discrete fit ST data in a satisfying way; ii) Under the Zipf or Pareto discrete assumptions, the percentage of uniques is connected with the indicator $\left(-\sum_{i=1}^{M}\log(\pi_i)\right)\Big/M$.

3. THE ZIPF AND THE DISCRETE PARETO DISTRIBUTION

3.1. THE ZIP DISTRIBUTION

According to the hypothesis that the number of units of a cell is a realization from a Zipf distribution, $\frac{W_j}{M} \cong Z_1(j)$, where $Z_1(j) = \frac{j^{-(\rho+1)}}{\zeta(\rho+1)}$ $j = 1,2,....$ and $\zeta(\rho+1) = \sum_{j=1}^{\infty} j^{-(\rho+1)}$ is the zeta function.

$Z_1(j)$ is often used as a discrete form of Pareto distribution. It can be shown that under the hypothesis of a Zipf distribution it is:

$$\frac{\log(W_j) - \log(W_i)}{\log(i) - \log(j)} \cong (\rho+1), \quad i \neq j. \tag{1}$$

The frequency of one–unit cells then is:

$$PU = \frac{W_1}{M} \cong \frac{1}{\zeta(\rho+1)}. \tag{2}$$

If we cal $H = \left(M \Big/ \sum_{j=1}^{max\ j} W_j \log(j) \right)$, and $\zeta'(\hat{\rho}+1)$, is the first derivative by ρ of $\zeta(\hat{\rho}+1)$, the maximum likelihood equation to estimate ρ is:

$$H = -\zeta(\hat{\rho}+1)/\zeta'(\hat{\rho}+1). \tag{3}$$

If can be shown that $-\zeta(\rho+1)/\zeta'(\rho+1)$ is an increasing function of ρ (Johnson, Kotz, 1969). When PU increases, ρ increases, and also H increases.

H is an increasing function of PU and can be used as an indicator of the risk.

3. 2. THE PARETO DISCRETE DISTRIBUTION

Let us considerer the Pareto distribution:

$$f(z) = \rho \cdot \lambda^{\rho} \cdot z^{-(\rho+1)},\ \lambda \le z < \infty,\ \lambda > 0,\ \rho > 0. \tag{4}$$

Another way to build a discrete distribution from f(z) is to consider the intervals $(j + \lambda - 1)$, $(j + \lambda)$:

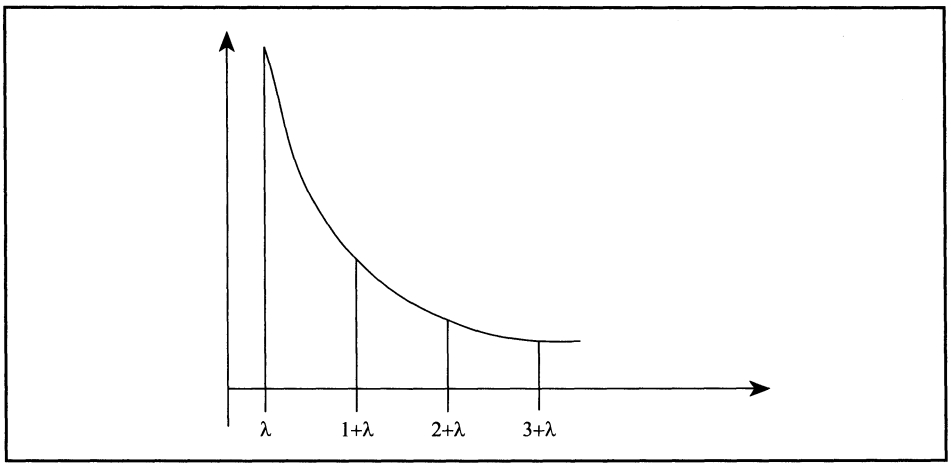

Fig.1: Pareto distribution.

The area between $(j + \lambda - 1)$ and $(j + \lambda)$ is:

$$Z_2(j) = \Pr\{(j+\lambda-1) \le z < j+\lambda\} = \left(\frac{\lambda}{(j+\lambda-1)}\right)^{\rho} - \left(\frac{\lambda}{j+\lambda}\right)^{\rho} \quad j = 1,2,\ldots \tag{5}$$

and this can be used as a discrete form of Pareto distribution. It can be shown that,

if $\frac{W_j}{M} \cong Z_2(j)$, then:

$$\frac{\log(1-U_j)}{\log(\lambda)-\log(j+\lambda)} \cong \rho \quad \text{where} \quad U_j = \left[\sum_{j=1}^{j} W_i \Big/ M\right]. \tag{6}$$

The frequency of one–unit cells is:

$$PU = \frac{W_1}{M} \cong 1-\left(\frac{\lambda}{1+\lambda}\right)^{\rho}. \tag{7}$$

It is interesting to note that, given the scale parameter λ, the proportion of uniques depends on the parameter ρ in the following way:

$$\log\left(1-\frac{W_1}{M}\right) \cong \rho \cdot \log\left(\frac{\lambda}{1+\lambda}\right), \tag{8}$$

$\log\left(\frac{\lambda}{1+\lambda}\right) < 0$, so if we fix the scale parameter λ, an increase of ρ is linked to an increase of PU. Let us consider the profile maximum likelihood estimator of ρ:

$$\hat{\rho} = \left(M \Big/ \sum_{j=1}^{\max j} W_j \log(j/\hat{\lambda})\right). \tag{9}$$

If $\hat{\lambda} \cong 1$, then $\hat{\rho} = H$ and

$$\log(1-PU) \cong H \cdot \log\left(\frac{1}{2}\right). \tag{10}$$

In this case, as in the Zipf distribution, H is an increasing function of PU and can be used to control the risk.

EXAMPLE 1:

Let us consdider the 2% sample form the 1981 Italian Population Census and the variables (number of categories): S = Sex (2); A = Age (110);
R = Relationship to the family head (9); M = Marital status (5); E = Educational level (6); O = Occupational group (9); P = Professional status (16).

In the first step we build the contingency table associated to the KEY: S × A

× R × M × E × O × P, in the second the ST.

From the observed ST we may estimate the Zipf and Pareto parameters on using a method based on non–linear least squares (Proc Model, SAS Institute, 1990).

Table I shows the observed values compared values compared to the Zipf and Dscrete Pareto estimates.

Tab. I.

Size of the cell	Observed (Frequency)	Zipf Estimates (Frequency)	Pareto Estimates (Frequency)
1	18,878	18,879	18,866
2	6,488	6,555	6,516
3	3,536	3,530	3,497
4	2,223	2,276	2,239
5	1,621	1,619	1,580
6	1,199	1,226	1,186
> 6	12,398	11,059	11,260
Total	45,144	45,144	45,144

EXAMPLE 2:

Let us consider the data of the 1989 ISTAT survey on the accounts system of the enterprises with 20 or more employees (42, 622 records) and the variables (number of categories): R = Region (21); E = Employes (5); S = Economic Sector (48); G = Value added class (14).

Following the same procedure used in example 1 we build the contingency table, the ST associated to the KEY: R × E × S × G and the Zipf and Pareto estimates.

Table II shows the observed values compared to the Zipf and Discrete Pareto estimates.

Tab. II.

Size of the cell	Observed (Frequency)	Zipf Estimates (Frequency)	Pareto Estimates (Frequency)
1	2,181	2,192	2,182
2	842	750	830
3	461	401	451
4	269	257	284
5	182	182	196
6	143	137	144
>6	1,044	1,203	1,026
Total	5,122	5,122	5,122

4. THE CONTROL OF THE NUMBER OF UNIQUES

The percentage PU is linked to H, so if we consider two different methods to reduce uniques on the contingency table, we may use H to measure the efficacy of the methods.

Let us consider more deeply the indicator $H = \left(M \Big/ \sum_{j=1}^{\max j} W_j \cdot \log(j) \right)$, composed by two factors: a) M, the total number of non empty cells; b) $\sum_{j=1}^{\max j} W_j \cdot \log(j)$.

It can be observed that b) can be written in tems of π_i:

$$\sum_{j=1}^{\max j} W_j \cdot \log(j) = \sum_{i=1}^{M} \log(F_i) = \sum_{i=1}^{M} \log(\pi_i) + M \cdot \log(N), \tag{11}$$

$$H = \left(\frac{M}{M \cdot \log(N) + \sum_{i=1}^{M} \log(\pi_i)} \right) = \left(\log(N) + \frac{\sum_{i=1}^{M} \log(\pi_i)}{M} \right)^{-1}. \tag{12}$$

Given N, the total number of units, H is an increasing function of

$$L = -\frac{\sum_{i=1}^{M} \log(\pi_i)}{M}.$$

The L numerator is a measure of the dispersion of units in the contingency table that reminds the entropy $E = \left(-\sum_{i=1}^{M} \pi_i \cdot \log(\pi_i) \right)$. This is standardized by the total number of non emply cells.

We have already shown that H is an increasing function of PU. L is an increasing function of H, so can be used to control alternative options to reduce uniques.

EXAMPLE 1: (CONTINUED)

If we consider the three keys:

KEY 1: S × A × R × M × E × O × P × ES

KEY 2: S × A × R × M × E × O × P

KEY 3: S × A × R × M × E × O

Variables (number of categories): S = Sex (2); A = Age (110); R = Relationship to the family head (9); M = Marital status (5); E = Educational level (6); O = Occupational group (9); P = Professional status (16); ES = Economic Sector (10).

Key 2 and Key 3 can be thought as aggregation from key 1 by suppression of variables.

Tab. III.

Key	N	M	PU%	L
1	1,095,412	81,124	49.4	12.994
2	1,095,412	45,144	41.8	9.306
3	1,095,412	27,742	36.5	12.536

EXAMPLE 2: (CONTINUED)

If we consider the three keys:

KEY 1: R × E × S × G

KEY 2: R × S × G

KEY 3: R × E × S

Variables (number of categories): R = Region (21); E = Employees (5); S = Economic Sector (48); G = Value added class (14).

As in example 1 Key 2 and Key 3 can be thought as aggregation from key 1 by suppression of variables.

Tab. IV.

Key	N	M	PU%	L
1	42,622	5,122	42.6	9.667
2	42,622	3,183	31.6	12.737
3	42,622	2,102	24.8	9.055

5. CONCLUSIONS

The information contained in the ST can be used to test the efficacy of the techniques to reduce the percentage of uniqe elements in a contingency table. The information contained in the ST can be even more summmarised. It is possible to

find indicators that synthetize the relevant information related to: i) the dispersion of units; ii) the size of the table.

We found that $\left(-\sum_{i=1}^{M} \log(\pi_i) \right)$ and M can be used to this purpose.

It can be useful to remark that the results obtained do not complete the research aimed to build a system of data protection, but represent a step in this work. We wish to conclude suggesting some open questions on which it will be useful to investigate:

a) What will be the consequences of considering as risky cells also two unit and three units cells?

b) The Zipf or Pareto discrete fitting is satisfying when the number of categories is big enough. Wath happens in the case of contingency tables of small size? Is there a limit size to accept to work on the models proposed?

c) How to estimate the variability of $\hat{\rho}$?

d) To facilitate the aggregation of variable's categories, are there indicators based on the marginals of the contingency table? (see Borra, Crescenzi, 1993a, 1993b).

REFERENCES

Bethlem J.G., Keller W.J., Pannekoek J., 1990, Disclosure Control of Microdata, *Journal of the American Statistical Association,* vol. 85, pp. 38–45.

Biggeri L., Zannella F., 1991, Release of microdata and statistical disclosure control in the new national system of Italy: main problems, some technical solutions, experiments, *Proceedings of the 48th ISI session,* Cairo.

Borra S., Crescenzi F., 1993a, Protecting enterprises data agaunst disclosure. A procedure to aggregate variables categories *Bullettin of the International Statistical Institute – Contributed Papers,* 49th Session, Firenze.

Borra S., Crescenzi F., 1993b, The control of disclosure of risk. a simple indicator to sect variables categories for the aggregation *Bullettin of the International Statistical Institute – Contributed Papers,* 49th Session, Firenze.

Coccia G., 1992, Disclosure risk in the Italian Current Population Survey, *Proceedings of International Seminar on Statistical Confidentiality,* Eurostat, Dublin, Ireland, pp. 415–423.

Crescenzi F., 1992 a, Un metodo per stimare il numero di casi unici della popolazione dalle informazioni di un campione osservato, *Atti della XXXVI Riunione Scientifica della SIS,* vol. 2, n. 3, pp. 357–364.

Crescenzi F., 1992 b, On Estimating Population Uniques. Methodological Proposals and Applications on Italian Census Data, *Proceedings of International Seminar on Statistical Confidentially,* Eurostat, Dublin, Ireland, pp. 247–260.

Dalenius T., 1986, Finding a Neddle In a Haystack (or Identifying Anonymous Census Records), *Journal of Official Statistics,* Vol. 2 n. 3, pp. 329–336.

Johnson N., Kotz S., 1969, *Distributions in statistics: Discrete distributions,* John Wiley & Sons, New York.

SAS Institute, INC., 1990, SAS/ETS *user's guide, version 6,* First Edition, Cary, NC, USA, pp. 315–397.

Skinner C.J., Holmes D.J., 1992, Modelling Population Uniqueness, *Proceedings of International Seminar on Statistical Confidentiality,* Eurostat, Dublin, Ireland, pp. 175–199.

RIASSUNTO

Gli Istituti Nazionali di Statistica incontrano una domanda crescente di dati e una delle più importanti sfide che gli Istituti devono fronteggiare è la ricerca di un equilibrio fra l'accesso ai dati e la loro protezione.

Il rilascio di informazioni sempre più dettagliate può portare, come conseguenza indesiderabile, alla violazione del diritto individuale alla riservatezza. Il problema può sorgere sia nel rilascio di micro dati che nel rilascio di dati tabulari o macro dati.

L'obiettivo di questo lavoro è quello di indicare nuovi metodi per controllare il rischio di violazione.

La proporzione di casi unici è il maggiore fattore che influenza il rischio. L'impiego di un modello Discreto di Pareto consente la costruzione di un indicatore basato sull'ammontare dell'informazione rilasciata e sulla concentrazione delle unità rispetto alle variabili che possono essere usate per violare la riservatezza dei dati. Il risultato è una diretta conseguenza e pone in risalto il legame esistente fra proporzione di casi nici, ammontare dell'informazione rilasciata e concentrazione dei dati.