A note on the individual risk of disclosure

Silvia Polettini*

Abstract. Individual risk estimation was one of the issues that the European Union project CASC targeted. The software μ-Argus contains now a routine, that has been implemented by CBS Netherlands in cooperation with ISTAT, for computing the Benedetti-Franconi or individual risk of disclosure. This note proposes an alternative expression for the Benedetti-Franconi risk, that can be exploited to produce more stable numerical evaluations. Such an expression provides an interesting relation with the Gauss Hypergeometric function.

1 The individual risk

The definition of the individual risk measure is based on the concept of re-identification disclosure (e.g. Chen and Keller-McNulty 1998, Fienberg and Makov 1998, Skinner and Holmes 1998, Duncan and Lambert 1986, Willenborg and de Waal 2001), and is appropriate for samples of microdata stemming from social surveys. By re-identification we mean that the unit in the released file and a unit in the register that an intruder has access to belong to the same individual in the population. The underlying hypothesis is that the intruder will always try to match a record and a unit in the register using public domain variables (*key variables*) only. In social data, categorical key variables allow us to tackle the problem of disclosure limitation via the concept of unique or rare *combinations* in the sample. A combination is a cell in the contingency table obtained by cross-tabulating the key variables. A key issue is to be able to distinguish between combinations that are at risk, for example sample uniques corresponding to rare combinations in the population, and combinations that are not at risk, for example sample uniques corresponding to combinations that are common in the population. To this aim, a step of inference from the sample to the population is performed. Instead of focusing on the sample frequency of combinations of key variables, the individual risk of disclosure is defined as the probability that a sampled record is re-identified, i.e. recognised as corresponding to a particular unit in the population. This value can then be estimated for each record in the released file on the basis of the observed sample. In the last few years a number of proposals have been made: Fienberg and Makov (1998), Skinner and Holmes (1998) and Elamir and Skinner (2003) define, with different motivations, a log linear model for the estimation of the individual risk. Benedetti and Franconi (1998) propose a methodology to estimate a measure of risk per record using the sampling weights, as the usual instrument that national statistical institutes adopt to allow for inference from the sample to the population. Further discussion of the approach is in Di Consiglio, Franconi and Seri (2003), Polettini (2003), Rinott (2003). A related approach is described in Carlson (2002).

1.1 Some notation

Let the released file be a random sample s of size n drawn from a finite population P consisting of N units. For a generic unit i in the population, we denote w_i^{-1} its probability of being included in the sample. Under the hypothesis that the key variables are discrete, cross-tabulating the key variables produces a set of combinations $\{1, \ldots, K\}$. A combination k is defined to be the k-th cell in the cross-tabulation. The set of combinations defines a partition of both the population and the sample and the sample values of the key variables on unit i will classify such a record into one combination. We denote by $k = k(i)$ the cell into which the sampled record i falls. Let f_k and F_k denote the size of the k-th cell in the sample and population, respectively. Retaining only the observed combinations -combinations with zero sample frequency being omitted- does not alter the above partition of the sample.

*ISTAT, DCMT, Via C. Balbo 16 00184 Roma

1.2 Definition of the individual risk

Assume for simplicity that there is complete agreement between the sample and the external archive available to the intruder, as far as the key variables are concerned (for a more general setting, see Polettini 2003). We first note that if we were to know the population frequency of the k-th combination, F_k, we would define the re-identification risk simply by $\frac{1}{F_k}$, for each record that is classified in combination k (i.e. $\forall i : k(i) = k$). The population frequencies are generally unknown, therefore an inferential step is to be performed. In the proposal by Benedetti and Franconi (1998) the uncertainty on F_k is accounted for in a Bayesian fashion by introducing the distribution of the population frequencies given the sample frequencies. The risk is then measured as the (posterior) mean of $1/F_k$ with respect to the distribution of $F_k|f_k$:

$$r_k = E\left(\frac{1}{F_k}\,|f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \Pr\left(F_k = h\,|f_k\right) \ . \tag{1}$$

To determine the probability mass function of $F_k|f_k$, the following superpopulation approach is introduced (see Bethlehem, Keller and Pannekoek 1990, Rinott 2003, Polettini 2003):

$$\begin{aligned} \pi_k \sim [\pi_k] \propto 1/\pi_k,\ \pi_k > 0,\ k = 1, \ldots, K \ , \\ F_k|\pi_k \sim \text{Poisson}(N\pi_k),\ F_k = 0, 1, \ldots \ , \\ f_k|F_k, \pi_k, p_k \sim \text{binomial}(F_k, p_k) \ ,\ f_k = 0, 1, \ldots, F_k \ . \end{aligned} \tag{2}$$

Under these assumptions, the posterior distribution of $F_k|f_k$ is negative binomial with success probability p_k and number of successes f_k. In general, the probability mass function of a negative binomial variable F_k counting the number of trials before the j-th success, each with probability p_k, is the following:

$$\Pr\left(F_k = h|f_k = j\right) = \binom{h-1}{j-1} p_k^j \left(1 - p_k\right)^{h-j}, \quad h \geq j \ .$$

In Benedetti and Franconi (1998) it is shown that under the negative binomial distribution the risk (1) can be expressed as

$$r_k = E(F_k^{-1}|f_k) = \int_0^\infty \left\{\frac{p_k \exp\left(-t\right)}{1 - q_k \exp\left(-t\right)}\right\}^{f_k} dt \ , \tag{3}$$

where $q_k = 1 - p_k$.

Substitution of an estimate of p_k in (3) can lead to an estimate of the individual risk of disclosure (1).

Given F_k, the maximum likelihood estimator of p_k under the binomial model in (2) is $\hat{p}_k = f_k/F_k$. F_k being not observable, Benedetti and Franconi (1998) propose to use

$$\hat{p}_k = \frac{f_k}{\sum\limits_{i \in k(i)} w_i} \ , \tag{4}$$

where $\sum\limits_{i \in k(i)} w_i$ is an estimate of F_k based on the sampling design.

2 An alternative expression of the individual risk of disclosure

Benedetti and Franconi (1998) propose to use the transformation $y = \left(1 - q_k \exp\left(-t\right)\right)^{-1}$ to obtain the following expression:

$$r_k = \left(\frac{p_k}{q_k}\right)^{f_k} \int_1^{1/p_k} \frac{(y-1)^{f_k - 1}}{y}\, dy,$$

but this is numerically unstable for values of p_k close to 0 and 1. Using the expression above, the authors propose an approximation, based on the Binomial theorem. The approximation has the same problems as (4) for p_k close to the extremes of the unit interval.

Instead, the transformation $\exp(-t) = y$ in (3) gives the integral

$$r_k = p_k^{f_k} \int_0^1 t^{f_k - 1} (1 - t q_k)^{-f_k} dt. \tag{5}$$

The previous formula can be expressed via the integral representation the Hypergeometric function ${}_2F_1(a,b;c;z)$ as

$$r_k = \frac{p_k^{f_k}}{f_k} {}_2F_1(f_k, f_k; f_k+1; q_k) \tag{6}$$

where

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a} dt$$

is the integral representation (valid for $(\Re(c) > \Re(b) > 0)$ of the Gauss Hypergeometric series

$${}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{n=0}^{\infty} \frac{\Gamma(a+n)\Gamma(b+n)}{\Gamma(c+n)} \frac{z^n}{n!}; \tag{7}$$

for definition and properties, see Abramowitz and Stegun (1965). The Gauss Hypergeometric Series (7) has circle of convergence the unit circle $|z| = 1$. The series may be divergent ($\Re(c-a-b) \leq -1$), convergent ($\Re(c-a-b) > 0$), conditionally convergent ($-1 \leq \Re(c-a-b) < 0$), except for $|z| = 1$.

Estimation of the individual risk can be accomplished by computation of the integral in (6). To this aim, special properties of the Hypergeometric function can be exploited; for example, it is easily seen that for $f_k = 1$ the function equals ${}_2F_1(1,1,2,q_k) = -\log(p_k)/q_k$, so that the risk becomes

$$r_k = \frac{p_k}{1-p_k} \log \frac{1}{p_k}.$$

For $f_k = 2$ we have

$$r_k = \frac{p_k^2}{2} {}_2F_1(2,2,3,q_k) = p_k^2 \frac{(q_k-1)\log(p_k) - q_k}{q_k^2(q_k-1)} = \frac{p_k}{(1-p_k)^2} \left[p_k \log p_k + 1 - p_k\right].$$

Similarly for $f_k = 3$:

$$r_k = \frac{p_k^3}{3} {}_2F_1(3,3,4,q_k) = \frac{p_k^3}{3} \frac{3q_k(3q_k-2) - 6(q_k-1)^2\log(1-q_k)}{2(q_k-1)^2 q_k^3} = \frac{p_k}{2q_k^2} \left[q_k(3q_k-2) - 2p_k^2 \log p_k\right].$$

In the definition of the individual risk methodology it is assumed that for each k $q_k, p_k \in (0,1)$. Recall that $\hat{p}_k = \frac{f_k}{\sum_{i:k(i)=k} w_i}$, w_i being the sampling weights. Whereas by definition $0 < p_k < 1$, estimates of this quantity might attain the extremes of the unit interval. Although we never deal with $\hat{p}_k = 0$ (corresponding to $f_k = 0$), when $\hat{p}_k = 1$ we have ${}_2F_1(f_k, f_k, f_k+1, 0) = 1$, so that the individual risk equals $1/f_k$.

3 Approximating the individual risk

In order to obtain an approximation to the individual risk for large f_k, contiguity relations for the Hypergeometric function (see Abramowitz and Stegun 1965) were exploited. In particular,

$${}_2F_1(f_k, f_k, f_k+1, 0) = \frac{1}{(1-q_k)^{f_k-1}} {}_2F_1(1,1,f_k+1,q_k).$$

Using the series representation (7) of the Hypergeometric function we get

$$\begin{aligned} {}_2F_1(f_k, f_k, f_k+1, p_k) =& \frac{1}{(1-q_k)^{f_k-1}} \left(1 + \frac{q_k}{f_k+1} + \frac{2^2 q_k^2}{2(f_k+1)(f_k+2)} + \frac{6^2 q_k^3}{6(f_k+1)(f_k+2)(f_k+3)} + \dots\right) \\ =& \frac{1}{(1-q_k)^{f_k-1}} \left(1 + \frac{q_k}{f_k+1} + \frac{2^2 q_k^2}{2(f_k+1)(f_k+2)} + O(f_k^{-3})\right). \end{aligned} \tag{8}$$

For large f_k the risk can therefore be approximated by

$$r_k \approx \frac{p_k}{f_k} \left(1 + \frac{q_k}{f_k+1} + \frac{2^2 q_k^2}{2(f_k+1)(f_k+2)}\right). \tag{9}$$

In most cases, the first order approximation

$$r_k \approx \frac{p_k}{f_k} \left(1 + \frac{q_k}{f_k+1}\right) \tag{10}$$

will be satisfactory. In practice, estimated p_k, q_k that depend on the observed frequencies f_k are used in formulas (9) and (10), therefore the order of magnitude of the remainders is of order $O(f_k^{-2})$ and $O(f_k^{-1})$ respectively. The first approximation is therefore recommended for moderate f_k. A simple check can be conducted on the term $6\hat{q}_k^3/[(f_k+1)(f_k+2)(f_k+3)]$: when this is not negligible, the first approximation is more accurate and recommended.

The approximation always leads to underestimating the risk. The error depends on the remainder. When using approximation (9) the error is of order $O(f_k^{-2})$. Alternatively, a better accuracy may be achieved by introducing additional terms in the representation: adding terms up to the third power of q_k, the error has order of magnitude $O(f_k^{-3})$, and so on. In general, in order to achieve an absolute error of approximation lower than ϵ with an observed frequency f_k, the order of the polynomial to be used for the approximation has to be an integer j such that

$$j > -\frac{\log(\epsilon)}{\log(f_k)}$$

Therefore an appropriate choice of the order of the approximating polynomial makes it possible to achieve the desired accuracy.

The approximations provided are based on the series representation of the Hypergeometric function ${}_2F_1(1, 1, f_k + 1, q_k)$. Its series representation is divergent when $f_k \leq 0$, therefore divergence is never of concern in practice. Absolute convergence of the series is guaranteed for $f_k > 1$.

Acknowledgements

The author would like to thank Luisa Franconi for for helpful conversations.

The author gratefully acknowledges the partial financial support of the European Union project IST-2000-25069 CASC on Computational Aspects of Statistical Confidentiality.

References

Abramowitz, M. and Stegun, I. A. 1965. *Handbook of Mathematical Functions*. Dover. New York.

Benedetti, R. and Franconi, L. 1998. Statistical and technological solutions for controlled data dissemination. *Pre-proceedings of New Techniques and Technologies for Statistics*. Vol. 1. Sorrento. pp. 225–232.

Bethlehem, J., Keller, W. and Pannekoek, J. 1990. Disclosure control of microdata. *Journal of the American Statistical Association* 85, 38–45.

Carlson, M. 2002. Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition* 5, 901–925.

Chen, G. and Keller-McNulty, S. 1998. Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* 14, 79–95.

Di Consiglio, L., Franconi, L. and Seri, G. 2003. Assessing individual risk of disclosure: an experiment. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg.

Duncan, G. T. and Lambert, D. 1986. Disclosure-limited data dissemination (with comments). *Journal of the American Statistical Association* 81, 10–27.

Elamir, E. and Skinner, C. 2003. Modeling the re-identification risk per record in microdata. *54th Session of the International Statistical Institute*. Berlin.

Fienberg, S. E. and Makov, U. E. 1998. Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* 14, 385–397.

Polettini, S. 2003. Some remarks on the individual risk methodology. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg.

Rinott, Y. 2003. On models for statistical disclosure risk estimation. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg.

Skinner, C. J. and Holmes, D. J. 1998. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics* 14, 361–372.

Willenborg, L. and de Waal, T. 2001. *Elements of Statistical Disclosure Control*. Springer. New York.