Working Paper No. 18
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (v): Risk assessment

SOME REMARKS ON THE INDIVIDUAL RISK METHODOLOGY

Invited paper

Submitted by the National Institute of Statistics (ISTAT), Italy [1]

[1] Prepared by Silvia Polettini (polettin@istat.it).

Some remarks on the individual risk methodology

Silvia Polettini
Istat, Servizio della metodologia di base per la produzione statistica
Via C. Balbo 16, I-00184 Rome, Italy

Abstract

The paper discusses some aspects of the individual risk methodology, that was initially proposed by Benedetti and Franconi (1998). The original formulation defines a record-level measure of re-identification, called the *individual risk*, that can be estimated exploiting information on the sampling design. This methodology is currently implemented in the testing version of the software μ-Argus, developed under the European project CASC.
When dealing with social surveys, that is the context where the individual risk methodology is best suited, it is reasonable to hypothesise that identification, which consists of linking a sample unit to a population unit, is performed based on a set of known identifying or *key* categorical variables. This implies that the individual risk depends on the joint distribution of the key variables, e.g. on the size f_k, F_k of subgroups of units having a given *combination* of key variables in the sample and population, respectively. Unlike the approaches that propose a record-level measure of risk based on the concept of sample uniques (e.g. Skinner and Elliot, 2002), the risk is defined for any record in the sample. The measure also differs from those based on the sample frequency of combinations, because inference on the sizes F_k of population subgroups is performed. The method shares with the above mentioned strategies the inferential nature and the approach to protection, respectively. Indeed, having estimated the individual risk for each record in the sample, protection is ensured by applying local suppression to high risk individuals only.

The paper discusses the formalisation of the individual risk function for files of independent units. Upon defining the disclosure scenario, the individual risk measure is linked to the probability of re-identification of a single record given information on a set of key variables observed on the whole population. Based on such connection, an overall measure of risk, called the *re-identification rate*, is proposed. Although this is a measure at the file level like the ones discussed by Skinner and Elliot (2002), it exploits the probability of re-identification of each sampled record. In particular, it is defined in terms of the expected number of re-identifications in the file to be released. Whenever the individual risk methodology is used to protect a sample by local suppression, the user is requested to select a risk threshold that classifies individuals into safe or unsafe. The paper investigates how the re-identification rate may be exploited for selection of a proper risk threshold using a measure of target "safety" of the whole file.

Introduction

This paper describes some aspects of the *individual risk methodology* as introduced in an initial paper by Benedetti and Franconi (1998). These authors propose a methodology for individual risk estimation based on the sampling weights; such procedure has been implemented into the beta version of the software μ-ARGUS. In this methodology by *disclosure* it is meant a correct *record re-identification* that is achieved by an intruder upon comparing a target individual in a sample with an available list of units that contains individual identifiers such as name, address and so on.

The basic concept is that of introducing a risk function defined *at the individual level*, instead of an overall measure defined on the whole file. The *individual risk* of unit *i* in the sample is the probability of it being correctly re-identified. Such probability is estimated, as we will show below, based on the sampled data. After the risk has been estimated, the main approach consists in fixing a threshold in terms of risk, e.g. the probability of re-identification. Units exceeding such threshold are defined at risk, and local suppressions will be applied to high risk individuals only, so as to lower their probability of being re-identified. In practice, risk estimation has the aim of introducing a *"disclosure" ordering* of units, which is then exploited to apply protection *selectively*.

1. Notation

Let the released file be a random sample of size *n* drawn from a finite population of *N* units. For generic unit *i* in the population, we denote as w_i^{-1} its probability to be included in the sample. For each record *i* the released file contains a set of key variables i.e. variables that allow identification and are accessible to the public, and the sensitive variables. Under the hypothesis that the key variables are discrete, a situation

which is classical in household surveys and in population censuses, we can focus the analysis on each of the k=1, …, K subpopulations defined by all the possible combinations of values of such variables; note that the maximum number of such combinations, K, can be quite high (in the order of hundred thousand). Let f_k and F_k be, respectively, the number of records in the released file and the number of units in the population with the k-th combination of categories of the key variables; F_k is unknown for each k. In the sample to be released only a subset of the total number K of combinations will be observed and only this subset, for whom $f_k>0$, is of interest to the disclosure risk estimation problem.

2. Disclosure scenario

The definition of *disclosure* follows the strategy of an intruder trying to establish a link between a unit in the sample s to be released and a unit in an available archive R. Such an archive, or *register*, contains individual direct identifiers (*name, ID number, phone number*…) plus a set of variables called *identifying* or *key variables (sex, age, marital status…).* The intruder tries to link unit $i \in s$ to a unit in R by comparing the values of the key variables. An *identification* occurs when based on this comparison, a unit i^* in the register is selected as a match to i and this link is correct, e.g. i^* corresponds to i, or otherwise stated, i^* is the labelling of unit i in the population.

The following assumptions are made, that define the *disclosure scenario*:

1. the archive available to the intruder covers *the whole population*; this is a conservative approach, in that the worst case is considered; this assumption implies that for each $i \in s$, the corresponding unit $i^* \in R$ does always exist;
2. the data to be released are a sample from a larger population, and sampling weights are available;
3. besides the individual direct identifiers, the archive contains a set of *key variables* that are also present in the sample;
4. the intruder tries to link a unit i in the sample to a unit i^* in the population register by comparing the values of the key variables in the two files;
5. the intruder has no extra information other than the one contained in the register;
6. a re-identification occurs when a link between a sample unit i and a population unit i^* is established and i^* is actually the individual of the population from which the sampled unit i was derived; e.g. the link has to be a *correct link* before an identification takes place.

3. Definitions

Within the above described scenario, the individual risk of disclosure r is defined as the *probability of re-identification* of a unit in the sample. In symbols:

$$r_i = P\left(i \text{ correctly linked with } i^* \mid s, R\right).$$

Clearly the probability that $i \in s$ is correctly linked with $i^* \in R$ is null if the intruder does not perform any link. Therefore conditioning on the event $L_i=1$ if the intruder attempts a re-identification of unit $i \in s$, and $L_i=0$ otherwise we have

$$r_i = P\left(i \text{ correctly linked with } i^* \mid s, R, L_i\right) P(L_i),$$

where $P(L_i)$ represents the probability that the intruder tries to establish a link between unit $i \in s$ and some unit in R.

The register containing the individual direct identifiers of a unit i^* that is included in the sample as unit i may be such that the values of the key variables recorded for i^* in R differ from those recorded for i in s. When trying to match a unit i in the released sample with a unit in the population register, the intruder compares the values of the key variables. Record re-identification is more likely if the key variables are recorded without discrepancies for $i \in s$ and its corresponding $i^* \in R$. If the key variables are recorded with error or missing values in either s or R, less information is available to the intruder for re-identification purposes, therefore such an identification is less likely to occur. Denote by V_i the 0/1 variable describing whether or not there is agreement in the values of the key variables between unit $i \in s$ and its corresponding $i^* \in R$. $V_i=0$ means perfect agreement, as far as the key variables are concerned, between unit i in the sample and its corresponding unit in the population archive. We can decompose the re-identification risk as:

$$r_i = P(L_i)\Big[P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 0\big) P(V_i = 0) + P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 1\big) P(V_i = 1) \Big] \quad (1)$$

As we already noticed,

$$P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 1\big) \le P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 0\big),$$

and formula (1) can be bounded by $r_i^* = P(L_i) P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 0\big)$:

$$r_i \le r_i^* = P(L_i) P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 0\big) \quad (2)$$

This means that estimating r_i^* in place *of* r_i is *prudential*, e.g. the actual risk is lower than the one we are estimating. We refer to r_i^* as the *individual risk of re-identification*. Recall that this is an upper bound to the probability of re-identification of unit *i* in the sample. Details on estimation of the individual risk are provided in the next section.

A prudential approach may lead to assume $P(L_i) = 1$, e.g. the intruder tries to re-identify each unit in the sample, so that the individual risk r_i^* simplifies to

$$r_i = P\big(i \text{ correctly linked with } i^* \mid s, R, L_i, V_i = 0\big). \quad (3)$$

This is referred to as the *base individual risk*, and is actually what is estimated within μ-Argus.

4. Estimation of the individual risk

Consider cross-tabulating the key variables, or collapsing the set of key variables to a single one. In either cases a set of *combinations* {1, …, *k*, …, K} is produced. A *combination k* is defined as the specific value *k* taken by this single variable, or the *k*-th cell in the cross-tabulation. The set of combinations defines a *partition* of both the population and the sample into *subdomains.* Retaining only the observed combinations (combinations with zero sample frequency being omitted) does not alter the above partition of the sample. Typically, we expect to find several sampled units within the same subdomain, e.g. combination *k* of key variables. Observing the values of the key variables on individual $i \in s$ will classify such individual into one subdomain. We denote by $k = k(i)$ the index of the subdomain into which individual $i \in s$ is classified based on the values of the key variables.

We take into account formula (3), that is, the upper bound to the probability of re-identification of unit *i* in the sample when the key variables exactly agree on the two data archives *s* and *R* and any sampled individual is matched to one individual in *R*. Once the intruder has found in the sample f_k individuals with the same combination *k* of key variables, k=1, … K, out of F_k individuals in the register *R* with such combination of key variables, these individuals are exchangeable for identification, e.g. each of the f_k can be linked to any of the F_k, assuming that no discrepancies occur between the sample and the register as far as the key variables are concerned. In this case the probability of a correct link conditional to a re-identification attempt would be simply $1/F_k$. Recall that the agency that distributes the data may not have access to the population register, and may not know the subpopulation sizes F_k. Therefore F_k is estimated via sampling design information. The *individual risk is therefore the agency estimate of the upper bound (2) to the intruder's re-identification probability* (see Trottini, 2001). According to the above consideration, risk as a shorthand for *risk from the point of view of the agency.* Thus r_i^* appearing in formula (2) is estimated as

$$\hat{r}_i^* = \sum_{h \ge f_k} \frac{1}{h} P(F_k = h \mid f_k) P(L_i). \quad (4)$$

This estimate can be split in two components, of whom the main term, $\sum_{h \geq f_k} \frac{1}{h} P(F_k = h \mid f_k) \stackrel{\text{def}}{=} \hat{r}_i$, is the estimate of the base individual risk r_i, whereas the second, $P(L_i)$, should be modelled according to the attack model, which is a component of the disclosure scenario.

4.1 Estimation of the base individual risk

In order to evaluate the *base* individual risk of re-identification $\hat{r}_i$, the distribution of $F_k|f_k$ has to be further modelled. For population surveys, Benedetti and Franconi (1998) assume that $F_k|f_k$ is distributed according to a negative binomial distribution with success probability p_k and number of successes f_k:

$$\Pr(F_k = h \mid f_k = j) = \binom{h-1}{j-1} p_k{}^j (1 - p_k)^{h-j} \quad h \geq j$$

Such framework stems from previous work by Bethlehem et al. (1990), who proposed the following hierarchical model:

$$\begin{aligned} p_k &\sim \text{Gamma}(\lambda, \alpha), \quad (\lambda\alpha = K) \\ F_k \mid p_k &\sim \text{Pois}(N p_k) \\ f_k \mid F_k &\sim \text{Bin}(F_k, \frac{n}{N}) \end{aligned}$$

in which K is the total number of combinations.

Benedetti and Franconi directly model $F_k \mid f_k$ by a negative binomial distribution with parameters f_k, p_k. Rinott (2002, personal communication) criticised choice of the Negative Binomial distribution as "the" model for $F_k \mid f_k$. However he showed simulations from several different models and finally concluded that, surprisingly, the original negative binomial model performed best. He also stated that under a modification of Bethlehem et al. (1990) model, namely

$$\begin{aligned} p_k &\sim \text{Gamma}(\lambda, \alpha), \quad (\lambda\alpha = K) \\ F_k \mid p_k &\sim \text{Pois}(N p_k) \\ f_k \mid F_k &\sim \text{Bin}(F_k, p_k) \end{aligned} \tag{5}$$

when $\lambda \to 0$ in the hyper-prior, then

$$F_k \mid f_k \sim \text{NBin}(f_k, p_k) \tag{6}$$

We obtained the same result (6) in a very similar framework, e.g. using model (5) except for giving p_k a diffuse (improper) prior $p(p_k) \propto 1/p_k$:

$$\begin{aligned} p_k &: p(p_k) \propto 1/p_k \\ F_k \mid p_k &\sim \text{Pois}(N p_k) \\ f_k \mid F_k &\sim \text{Bin}(F_k, p_k) \end{aligned} \tag{5'}$$

The stated superiority of the negative binomial model, at least in selected circumstances, can be explained in light of the result we obtained. In fact use of noninformative priors allows matching Bayesian and frequentist results and we therefore expect that model (6) has "good" frequentist properties. Note that use of a parameter p_k in the binomial distribution for $f_k \mid F_k$ accounts for stratified sampling, under which simple random samples are drawn independently from each stratum.

The result we obtained is in practice the same of Rinott (2002). Indeed for the Gamma distribution it can be proved that when $\lambda \to 0$ and $\alpha \to 0$ with the same order, the Gamma density $p(p_k; \lambda, \alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} p_k{}^{\alpha-1} e^{-\lambda p_k}$ tends to $1/p_k$, which is a diffuse prior. Use of noninformative prior is also convenient as knowledge of *K* is not required.

As the parameter p_k is estimated based on the observed frequencies, model (6) is therefore justified as a noninformative empirical Bayes procedure. The maximum likelihood estimator of p_k under model (5') is $\hat{p}_k = \frac{f_k}{F_k}$. As under this framework F_k is not observable, Benedetti and Franconi (1998) propose to estimate it as $\hat{p}_k = \frac{f_k}{\sum_{i \in k(i)} w_i}$. Properties of ratio estimators like $\hat{p}_k$ are described e.g. in Särndal et al. (1992).

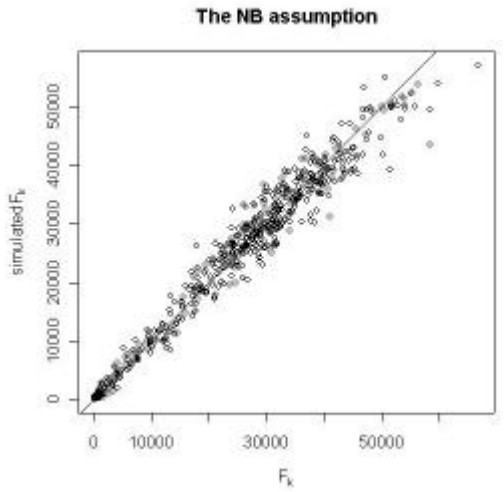

Figure 1. Population frequencies vs. simulations from negative binomial distributions

Note also that the model introduced by Benedetti and Franconi is a *mixture model*, in which a *size 1* sample from each subpopulation is observed. For this reason, unless assuming that a sufficiently high number of classes can be collapsed, it is also difficult to test the assumption (5') or (6) based on the observed data. Figure 1 provides a simple simulation experiment in which for fixed combinations, their population frequencies F_k are compared to independent samples of size 1 from negative binomial distributions of type (6), where $\hat{p}_k = \frac{f_k}{F_k}$ (F_k is known in this experiment). Of course this cannot be interpreted as a formal test.

Under model (6), the base individual risk equals $\sum_{h \geq f_k} \frac{1}{h} P(F_k = h \mid f_k) = E(F_k^{-1} \mid f_k)$; for estimation we can exploit the analytic expression valid for the above expectation under the negative binomial distribution, e.g.

$$E(F_k^{-1} \mid f_k) = p_k^{f_k} \int_0^1 \frac{y^{f_k - 1}}{(1 - p_k y)^{f_k}} \, dy \qquad (7)$$

The expression in formula (7) may be computed, and approximated for large f_k.

Substitution of an estimate $\hat{p}_k$ in formula (7) above leads to estimation of the base individual risk of disclosure $\hat{r}_i$.

For the purpose of estimating the base individual risk, Benedetti and Franconi (1998) rewrite (7) by an expansion based on the Binomial theorem; however this might lead to unstable estimates as it depends on the ratio $[\hat{p}_k / (1 - \hat{p}_k)]^{f_k}$ Alternative expressions based on the Hypergeometric function

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

may be exploited. Indeed

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

We also note that approximations of the base individual risk exploiting the above representation can be provided. To this aim we used the series representation of the Hypergeometric function. In this case, convergence of the series is guaranteed by f_k being always greater than zero.

Note that $0 < p_k < 1$, but estimates might attain the extremes of the unit interval. Whereas we never deal with $\hat{p}_k = 0$, which is obtained only when $f_k = 0$, it can still happen that $\hat{p}_k = 1$. In this case ${}_2F_1(f_k, f_k, f_k+1, 0) = 1$, and the base individual risk equals $1/f_k$. Details are provided in a technical report (Polettini, 2003).

Estimating the *individual risk* using relation $\hat{r}_i^* = P(L_i)\ \hat{r}_i$ requires specifying the attack model; each choice leads to different estimates. Recall that under any of the attack models discussed in the next section the individual risk $\hat{r}_i^*$ will estimate an *upper bound* to the probability of correctly linking record i of the sample to a unit i^* in the population archive.

4.2 Modelling $P(L_i)$: attack models

There are several *attack models* that one can conceive:

M1. The simplest approach is the *random attack* model, under which the intruder selects a unit at random from the sample; denoting the sample size by n, we have $P(L_i)=n^{-1}$ which is constant across the whole sample.

M2. A different approach is to hypothesize that the intruder compares the sample frequencies f_k with the register or population frequencies F_k and then tries to identify first those population individuals that are highly represented in the sample, e.g. those scoring high on the ratio f_k/F_k. In this case we have $P(L_i)= f_k/F_k$, which is constant over a single combination of key variables.

M3. A prudential assumption is to hypothesize that the intruder tries to match any single record in the sample to a unit in the register. In this case $P(L_i)=1$ for all i.

M4. According to different hypotheses, constant probability other than 1, n^{-1} can be assumed for the linking operation: $P(L_i)=p$ for all $i \in s$. This might be the case when a microdata file for research is to be released. In this case we might assume that a recognised researcher receiving the data upon contract would attempt to disclose information with low constant probability.

M5. Finally, another attack model assumes that, before trying to link a unit in the sample with a unit in the population, the intruder scans the sampling weights and first attempts to re-identify those individuals whose inclusion probability is highest. In this case $P(L_i)$ can be modelled as $P(L_i)=p_i=w_i^{-1}$, w_i representing the sampling weight. This in general will vary across individuals and combinations. If the key variables are those defining the strata in the sampling design, attack model M2 and M5 should coincide.

5. Assessing the risk of the whole file: global risk

The individual risk provides a measure of risk *at the individual level.* A *global* measure of disclosure risk for the whole file can be expressed in terms of the expected number of re-identifications in the file. In this section we introduce the *expected number of re-identifications* and the *re-identification rate* as relative and absolute measures of disclosure, respectively. These measures are not new in the literature of disclosure limitation, see for example Lambert (1993) and Dobra et al. (2002). Here these measures are introduced with the sole aim to provide a tool for setting a threshold for the individual risk, which in turn is needed to protect the data by the individual risk methodology. In our view measures at the individual level are more helpful than overall measures, even those based on uniques, in that they allow applying protection to selected records.

Define a dichotomous random variable Φ, assuming value 1 if the re-identification is correct and 0 if the re-identification is not correct. In general for each unit in the sample one such variable Φ_i is defined, assuming value 1 with at most probability $r_i^* = P(L_i)\ r_i$. In the discussion we behave as if such probability

were *exactly* r_i^*. Generally speaking, the random variables Φ_i are not *i.i.d.*. Under any attack model but model M5, $P(L_i)$ is in fact constant over the subdomain and therefore $P(L_i)\,r_i$ is constant. This implies that for subdomain k defined by a given combination k of key variables, we have f_k *i.i.d.* such random variables Φ_i assuming value 1 if the re-identification is correct with probability $j_k = P(L_i)\,r_i = r_i^*$. All these quantities can be estimated by plugging in the estimates $\hat{r}_i^*$. The above probabilities can be exploited to derive the *expected number of re-identifications per subdomain*, which equals $f_k\, j_k$.
In general, the overall *expected number of re-identifications over the whole sample* is the expected value $\mathrm{ER} = \sum_{i=1}^{n} E(\Phi_i) = \sum_{i=1}^{n} r_i^*$, that can be further specified according to the attack model:

M1. for random attack model, where $P(L_i)=1/n$, $\mathrm{ER}_1 = \frac{1}{n}\sum_{k=1}^{K} f_k r_k$;

M2. for the attack model in which the intruder compares the subdomain sizes in the sample and in the population, e.g. when $P(L_i)=f_k/F_k$, the expected number of re-identifications over the whole file is $\mathrm{ER}_2 = \sum_{k=1}^{K} \frac{f_k^{\,2}}{F_k} r_k$;

M3. for the pessimistic attack model according to which $P(L_i)=1$, the expected number of re-identifications is simply $\mathrm{ER}_3 = \sum_{k=1}^{K} f_k r_k$

M4. analogously to model M3, the constant probability attack model according to which $P(L_i)=p$, for any i, $p \in (0,1)$, has an expected number of re-identifications $\mathrm{ER}_4 = p \sum_{k=1}^{K} f_k r_k = p \cdot \mathrm{ER}_3$;

M5. when the probability of attack differs among sampled units as in model M5, the probability of re-identification varies across sub-domains, and the expected number of re-identifications equals in this case $\mathrm{ER}_5 = \sum_{i=1}^{n} p_i r_{k(i)} = \sum_{k=1}^{K} r_k \Pi_k$, where $\Pi_k = \sum_{i \in k(i)} p_i$ is the total inclusion probability of units falling into subdomain k.

If we define the *re-identification rate* x as

x = expected number of re-identifications/n,

then x provides a measure of *global risk, i.e.* a measure of disclosure for the whole file.
As it is clear from the above discussion, the risk assessment of the sample is affected by the particular assumption made on the attack model. For example, if our hypothesis is that the intruder tries to re-identify *any single* sampled individual as in model M3, a confidentiality breach is more likely to take place than under attack model M1, when the intruder selects one unit at random from the sample. Whereas in the former case the expected number of re-identifications is

$$\mathrm{ER}_3 = \sum_{k=1}^{K} f_k r_k \,, \qquad \text{and consequently} \qquad x_3 = \frac{\sum_{k=1}^{K} f_k r_k}{n} \,,$$

under model M1 there is eventually *only one* individual out of n that is under attack, and accordingly the expected number of re-identifications equals $\mathrm{ER}_1 = \frac{1}{n}\sum_{k=1}^{K} f_k r_k$. In such case we have a re-identification rate $x_3 = \frac{\sum_{k=1}^{K} f_k r_k}{n^2}$, that is $1/n$-th the one under the previous model.

6. Remarks on estimation of the individual risk

The procedure relies on the assumption that the available data are a sample from a larger population. The sampling design is assumed known, as far as the sampling weights are concerned at least.
If the sampling weights are not available, or if data represent the whole population, the strategy used to estimate the individual risk is not meaningful. Moreover the significance of the estimates carries over the risk estimate. In order to get reliable results, before applying the individual risk methodology, it is advisable to recode hierarchical variables (e.g. geography) up to the minimum level these are significant, and then run the individual risk methodology so that it produces significant risk estimates.
Assessment of the analytic properties of the risk estimators is still under investigation (see also Di Consiglio et al., 2003). In general due to averaging over the combinations, the overall measures discussed in the previous paragraph are more stable than the individual measures. Properties of the latter depend on size of the sample subdomain.

7. Practical issues in the application of the individual risk methodology: Threshold setting

When going to protect the microdata file, users must have in mind a threshold, e.g. a level of *acceptable risk*, representing a risk value under which an individual can be considered safe.
In determining such a level, users can refer to the expected number of correct re-identifications in the file. If the expected number (or percentage) of correct re-identifications is below a level the user considers acceptable, then he/she can decide to release the file as it is. Once fixed a tolerable percentage of re-identifications in the file to release, a threshold in terms of probability of re-identification can be selected (see below) and the individuals that exceed such threshold are defined at risk. Specifically, such individuals are at risk because their probability of re-identification is high enough to give rise to a large expected proportion of correct re-identifications over the whole file. Suppressions will be applied only to those individuals whose risk exceeds the selected threshold. Clearly choice of the threshold affects the quality of the resulting "safe" file, and before saving the output file, the threshold can be changed. This will allow for assessment of the risk of the file, number of consequent suppressions and therefore quality of the "safe" file before data release.

As already discussed, a threshold for the individual risk can be determined by choosing the maximum tolerable *re-identification rate* in the sample. We give an example using the worst method of attack (attack model M3), which also means that we can refer to the *base* individual risk (3). According to the worst method of attack, the expected number of re-identifications is the sum $\mathrm{ER}_3 = \sum_{k=1}^{\mathrm{K}} r_k f_k$ (see Section 5). Units can therefore be sorted according to their individual risk. Since units belonging to the same subdomain k have the same individual risk, *strata* can be arranged in increasing order of such risk. The subscript k can be used to denote the so sorted strata. Each stratum is responsible for a certain amount $r_k f_k$ of expected re-identifications.
Once a threshold r^t has been set on the individual risk, the individuals whose risk exceeds the threshold will undergo protection via local suppression. Protected units will have individual risk lower than the selected threshold; for this reason the individuals that will be locally suppressed, will have *after protection*, $r_i \le r^t$. This means that the expected number of re-identifications in the released sample after protection will be certainly smaller than

$$\sum_{k=1}^{\mathrm{K}^t} f_k r_k + r^t \sum_{k=\mathrm{K}^t+1}^{\mathrm{K}} f_k \,.$$

In the above formula r^t and K^t are in one-to-one correspondence, i.e. K^t indexes the subdomain whose units have individual risk r^t. In fact the threshold r^t can be picked from the discrete set of *observed* risk values, as choosing a threshold that is bracketed by two consecutive observed risk levels would not change the expected number of re-identifications.
The previous formula allows setting r^t so that *after protection* the upper bound on the expected number of re-identifications in the released file does not exceed a chosen level t, e.g. $\sum_{k=1}^{\mathrm{K}^t} f_k r_k + r^t \sum_{k=\mathrm{K}^t+1}^{\mathrm{K}} f_k \le t$. This provides a guideline in setting an appropriate threshold on the individual risk based on an objective

global measure of risk for the whole file. If as it is reasonable an absolute measure is preferable, the desired level for the re-identification rate x can be used to derive the appropriate level for the expected number of re-identifications and to set the threshold in the same way as it was just discussed.

Acknowledgments
The author thanks Luisa Franconi for helpful comments and Giovanni Seri also for providing data sets extracted from census files.
Model (6) as a consequence of (5') stems from the analogous result (formulae (5) and (6)) J. Rinott stated in a seminar he gave at Istat.
This work was partially supported by European project IST-2000-25069 "Computational Aspects of Statistical Confidentiality".

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.

Di Consiglio, L., Franconi, L. and Seri, G. (2003) Assessing individual risk of disclosure: an experiment. To be presented at the Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003)

Dobra, A., Fienberg, S.E., and Trottini, M. (2002), Assessing the Risk of Disclosure of Confidential Categorical Data. To appear in J. Bernardo et al., eds., Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics, Oxford University Press.

Lambert, D. (1993) Measures of disclosure risk and harm, Journal of Official Statistics, 9, 313-333.

Rinott, J. (2002) Seminar on "Modeling, estimation and speculation in statistical disclosure control: a preliminary report" held at Istat, July 2002

Polettini, S. (2003) An alternative expression for the individual risk of disclosure. Technical report.

Trottini, M. (2001) A decision-theoretic approach to data disclosure problems Research in Official Statistics, 4, 7-22

Särndal, C-E, Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer, New York.

Skinner, C.J. and Elliot, M.J. (2002): A measure of disclosure risk for microdata, Journal of the Royal Statistical Society, Series B, 64, 855-867