Working Paper No. 16
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

ON MODELS FOR STATISTICAL DISCLOSURE RISK ESTIMATION

Invited paper

Submitted by the Hebrew University (Israel)[1]

[1] Prepared by Yosef Rinott (rinott@mscc.huji.ac.il). This work was supported by a grant of the Central Bureau of Statistics, Israel.

On models for statistical disclosure risk estimation

Yosef Rinott *

March 15, 2003

Abstract

Risk estimation in the context of statistical disclosure control is usually model based. Many models and related methods have been proposed in recent years, and some programs are distributed to implement them. However, the literature and the program manuals rarely discuss statistical questions such as the sensitivity or robustness of the estimates relative to the assumptions of the models, goodness of fit tests for the validity of the model, and variance estimates of the risk measures proposed.

In order to deal with these issues in concrete terms we chose here to discuss two well-known models and methods for risk estimation. We study variations on a model of Bethlehem et. al. (1990), to which we shall refer as the B model, and on a model due to Benedetti and Franconi (1998), to which we shall refer as the ARGUS model or the A model for brevity, since it forms the basis for risk estimation in the ARGUS program. In particular the goodness of fit tests proposed here are based on an embedding of the A model into the B model which is demonstrated in this paper.

We come to analyze and contribute to these models, not to bury nor to praise them.

Keywords and phrases: global risk measures, population uniques, key variables, sensitive variables, goodness of fit tests.

---

*Department of Statistics, Hebrew University, Jerusalem Israel. This work was supported by a grant of the CBS.

1 Introduction

Consider a file containing a sample of microdata, that is, data on several variables for a sample of individuals given in the form of a contingency table. We distinguish between *key* variables which may help an intruder identify certain individuals in the file, and *sensitive* variables about which an intruder may want to learn from the sample data if identification of individuals is possible. This distinction may depend on the intruder, and some variables may be both key and sensitive. It is assumed that the intruder has some knowledge on the key variables for individuals of interest. In the sequel we are concerned only with key variables, and with estimating the risk (or probability) that an intruder correctly matches an individual in the sample to an identified person in the population on the basis of key variables, either by realizing that this individual is unique in the population with a certain combination of key values, or by guessing, which might work if such a combination is rare.

The agency which considers releasing the file knows the file itself, which consists of a sample from some population, and perhaps has some partial knowledge on the whole population. On the basis of this knowledge the agency wants to estimate the disclosure risk involved in releasing the file.

2 μ-ARGUS: a Negative Binomial model

Let F_k and f_k denote the population and sample frequencies of cell k (in the key variables), $k = 1, \ldots, K$. The μ-ARGUS manual proposes a model of Benedetti and Franconi (1998) according to which one assumes that $F_k \,|\, f_k \sim NB(f_k, p_k)$, the Negative Binomial distribution counting the number of **coin tosses** until f_k Heads are observed, with probability of a Head being p_k. Clearly $F_k \geq f_k$.

In addition it is assumed that for each individual in the sample there is a known inflation factor w_i. F_k is estimated by $\hat{F}_k = \sum_{i:k(i)=k} w_i$, where $k(i) = k$ indicates that the ith individual belongs to the kth cell defined by the key variables.

(Here we propose to consider the possibility of incorporating a model into the latter estimates. For example, if a particular log linear model representing some independence or conditional independence in the table is reasonable, then the above estimates $\hat{F}_k$ can be readjusted to fit the model in standard ways. This combination of ideas of Fienberg and Makov (1998) with the ARGUS model is discussed elsewhere.)

The above ARGUS or A model is an unusual setup in the sense that the parameters F_k which we wish to estimate are modelled as random variables given the data, with some partial information on them through the inflation factors, which are not part of

the probabilistic model. Without a prior on the F_k's, this is neither an ordinary nor an empirical Bayes model, and of course it is not a standard frequentist model.

This lack of framework makes it difficult to perform standard statistical analyses of the method, e.g., set a likelihood function, compute MLE's, assess the variability of estimates, construct some kind of confidence intervals to the proposed measures, and most importantly, test the validity of the model. Indeed, to our knowledge, none of these issues are seriously addressed in the ARGUS related literature. For example, since the model is conditioned on the observations f_k, it says nothing about their distribution, so goodness of fit of the data cannot be tested.

We show that much light can be shed on the A model by embedding in an empirical Bayes model of Bethlehem at al. (1990), and in particular this approach may lead to goodness of fit tests. Also, We shall indicate a direction which can produce variance estimates and confidence intervals, but the latter need to be properly defined in this model of random parameters. We discuss this model in detail because it seems to work in certain real data situations, but to fail in others, and we would like to understand the conditions under which it is applicable, and because it seems to be a candidate for use by European statistical agencies in the future. Our experiments with the A model are described elsewhere.

In its simplest formulation (which is sufficiently general for the sake of this discussion), the risk in cell k is defined as $r_k = 1/F_k$, the probability of correct matching if the intruder matches at random an individual from cell k in the sample with one of the F_k individuals in the kth population cell.

The F_k's are unknown and estimates are needed. It looks like a candidate for an estimate of $1/F_k$ could be $1/\hat{F}_k$. It should be a good one if $\hat{F}_k$ is a good estimate. In terms of the A model, F_k is random and $r_k = E_{p_k}[1/F_k \,|\, f_k]$. ARGUS estimates r_k by first estimating p_k. It is easy to see that under the Negative Binomial model, $E_{p_k}[F_k \,|\, f_k] = f_k/p_k$. This leads to the "moment estimate" $\hat{p}_k = f_k/\hat{F}_k = f_k/\sum_{i:k(i)=k} w_i$. (The estimate f_k/F_k is MLE if F_k is assumed known, but this is not really the case here.) From the original w_i's we have now computed an estimate for the sampling fraction in cell k which is the harmonic mean of the individual sampling fractions or inclusion probabilities $\pi_i = 1/w_i$.

With the estimate $\hat{p}_k$, ARGUS estimates the risk r_k by $\hat{r}_k = E_{\hat{p}_k}[1/F_k \,|\, f_k]$.

This expectation is easy to compute. As proposed in Benedetti and Franconi (1998), a standard calculation (integrating the moment generating function of a variable X to obtain $E\frac{1}{X}$ and substitution in the integral) lead to

$$\hat{r}_k = (\frac{\hat{p}_k}{1-\hat{p}_k})^{f_k} \int_1^{1/\hat{p}_k} \frac{(u-1)^{f_k-1}}{u} du\,.$$

A straightforward expansion of $(u-1)^m$ and integration shows that

$$\int_1^a \frac{(u-1)^m}{u} du = (-1)^m \log a + \sum_{i=1}^{m} \frac{1}{i} \binom{m}{i} (-1)^{m-i} (a^i - 1).$$

When $f_k = 1$ we obtain $\hat{r}_k = -\frac{\hat{p}_k}{1-\hat{p}_k} \log(\hat{p}_k)$. Thus $\hat{r}_k$ is evaluated easily once p_k is estimated (the above sum runs in no time on a Matlab program, for example; for large m, it can also be approximated in standard ways).

ARGUS manuals, and Benedetti and Franconi (1998), Benedetti et al. (2003) focus on the individual risk measure $\hat{r}_k$. Global risk measures in other models were presented by numerous authors (see references). It is natural to invoke them here as well. The present authors (Technical Report, 2002, and lecture at ISTAT July 2002) and Polettini and Seri (2003) proposed to use the $\hat{r}_k$'s to construct global risk measures, which in general have the form $\sum_k a(f_k)\hat{r}_k$ (or more involved forms which we avoid for simplicity of this paper). For example, the measure $\hat{\tau} = \sum_k I(f_k = 1)\hat{r}_k$, which we use here to demonstrate some of our points, is an estimate of $\tau = \sum_k I(f_k = 1)1/F_k$, the expected number of correct guesses by an intruder who bases his matching on sample uniques.

Returning to the basic estimates: ARGUS estimates $1/F_k$ by first estimating F_k, from it p_k is estimated in a particular way and then expectation with the latter estimate as parameter is computed in order to estimate $1/F_k$.

Why and when should this approach work? How sensitive is it to the assumptions? Given the sample file, how do we test that the assumptions hold and the model fits?

The Negative Binomial model is akin to the inverse sampling method (Haldane, 1945), which is not commonly used by statistical agencies. The model may be relevant, but cannot be taken for granted. The Negative Binomial assumption can be interpreted as follows: sample and population are created together. When an individual in the population cell k is created, it enters the sample with probability p_k, otherwise it is counted only in the population, not in the sample. When the sample cell reaches the count f_k, then no more individuals of type k are created in the population, and its value F_k is reached. Note that strictly speaking, this seems to be inconsistent with the possibility of having different sample inclusion probabilities π_i for individuals in cell k.

The first issue we discuss is the overestimation tendency of ARGUS of individual risk under certain conditions due to a simple mathematical fact (Jensen's inequality). The following discussion also indicates another inconsistency of the model: if the inflation factors provide perfect information about the F_k's (see details below), a consistent model should produce $\hat{r}_k = 1/F_k$. With ARGUS this is not the case. Let us consider for simplicity an extreme situation where the weighting class partition co-

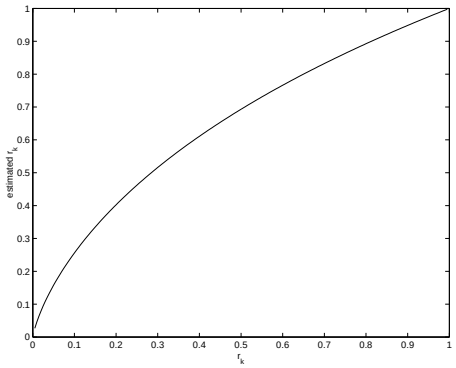
incides with the key variable partition. (It should be clear that the problem discussed exists more generally, not just in this simple case.) In this case the population key variables, or equivalently the F_k's are completely known to the agency. Therefore, in this case the estimates satisfy $\hat{F}_k = F_k$ so that $\hat{F}_k$ are correct and should be used. Since ARGUS is based on the estimates $\hat{F}_k$, it seems that it should work well when these estimates are good. Rather than use these good estimates directly, ARGUS would use its own $\hat{r}_k$ with $\hat{p}_k = f_k/\hat{F}_k = f_k/F_k$. Since for any positive random variable $E[1/X] > 1/EX$, it follows readily that

$$\hat{r}_k = E_{\hat{p}_k}[F_k^{-1} \mid f_k] > 1/E_{\hat{p}_k}[F_k \mid f_k] = 1/[f_k/\hat{p}_k] = 1/F_k.$$

Thus ARGUS will provide overestimates to the correct values of *individual* risk. How serious is this systematic bias?

For the present situation of full information, Figure 1 shows the values of $r_k = 1/F_k = p_k/f_k$ against their ARGUS estimates $\hat{r}_k$ for $0 \leq p_k \leq 1$ and $f_k = 1$, corresponding to sample uniques. In this case $r_k = p_k$. Note that for small p_k's the overestimation is severe. For example, for $p_k = 1/200 = 0.005$ the ARGUS estimate $\hat{r}_k = 0.0266$ which is an overestimate by a factor of $0.0266/.005 = 5.3$. Thus in the range of sampling fraction of $1/100 - 1/200$, which is common in the Israeli CBS, global measures which look at sample uniques ($f_k = 1$) may overestimate individual risk by a factor of 5. The same may happen, perhaps more mildly for larger sampling fractions and for measures not based on uniques. Thus ARGUS seems problematic (even?) when perfect estimates of its parameters are available. Can it be better with less information?

Figure 1: correct vs ARGUS estimated risks, Note the over estimation



In the above discussion of overestimation bias, the F_k's were taken as known, and a similar discussion holds if they are known up to a relatively small error. If we denote the information on the population table (which often consists of certain marginals) by M, then rather than r_k, the right risk measure is $E_{p_k}[1/F_k \mid f_k,\ M]$, with

an appropriate estimate of p_k. This would prevent the bias issue discussed above. We hope to pursue this direction elsewhere.

Another source of bias, this time towards underestimation, is related to the fact that in standard inflation models empty cells cannot be inflated. This problem, and a possible remedy in connection with the B model, are discussed towards the end of the paper.

While the A model overestimates r_k when it is actually known, recall that these parameters are in general random. Using the model we can therefore write $E_{p_k}(r_k)$ or equivalently $E_{p_k}[1/F_k \mid f_k]$, which equals $\hat{r}_k$ provided the latter is computed with a correct value of $\hat{p}_k$. Thus we have unbiasedness in a reversed but still very reasonable sense: the expectation of the parameter equals the estimator (under certain conditions). Next we consider global measures in this reversed sense. Assuming $F_k \mid f_k \sim NB(f_k, p_k)$ and again that the $\hat{r}_k$'s are computed with a good estimate $\hat{p}_k$'s, then the random variable $\tau = \sum_k I(f_k = 1)1/F_k$ (which is a quantity we are trying to estimate) has expectation $\hat{\tau} = \sum_k I(f_k = 1)\hat{r}_k$. Moreover, under certain conditions, which include the very necessary assumption of *conditional independence* of $\{F_k\}$ given $\{f_k\}$, (not mentioned explicitly in Benedetti et al.; some mention of group independence appears in the ARGUS related literature), the laws of large numbers imply convergence of τ to $\hat{\tau}$, when both are properly normalized. Note again the unusual situation where the parameter's expectation equals the estimator, and converges to it, which is a reversal of the usual situation. Below we do some variance calculations for τ and $\hat{\tau}$.

In order to further understand global measures, we concentrate for simplicity on a single weighting class C. We assume that it contains several cells k, and therefore all individuals in these cells have the same inflation factor w. We assume without loss of generality that $f_k = 1$ for all $k \in C$. Let $\tau_C = \sum_{k \in C} 1/F_k$ (a risk measure for the class C) and let $\hat{\tau}_C = \sum_{k \in C} \hat{r}_k$ be its ARGUS type estimate. Since C is a weighting class, $F_C := \sum_{k \in C} F_k$ is known, and we can estimate a single $p = p_k$ for all $k \in C$ by $\hat{p} = |C|/F_C$, where $|\cdot|$ denotes cardinality. This is a good estimate under the model if $|C|$ is not small, the above Negative Binomial assumption holds, and in addition if F_k are conditionally independent, an important assumption also for other arguments. This coincides with the ARGUS estimates (for $f_k = 1$), $\hat{p}_k = 1/w$. Also, if $|C|$ is not small, conditioning on the known value of F_C does not change the joint distribution of $\{F_k \mid f_k\}$ by much and they remain nearly independent Negative Binomials. Thus, by the above unbiasedness, $\hat{\tau}$ and τ tend to be close due to summing (or averaging). The two quantities have small variances: under the model of Negative Binomial and conditional independence, $E(\tau) = -|C|\frac{p}{q}\log(p)$, $\mathrm{Var}(\tau) = |C|\{\frac{p}{q}\sum_{j=1}^{\infty} q^i/i^2 - [\frac{p}{q}\log(p)]^2\}$, which is bounded by $\frac{\pi^2}{6}\frac{p}{q}|C|$. For $1/10 > p > 1/10000$, we have roughly

$-\sqrt{p}\,\log(p) > .1$, so the coefficient of variation, $-\pi\sqrt{q}/[\sqrt{6}\sqrt{p}\log(p)\sqrt{|C|}\,]$ is of order no bigger than about $13/\sqrt{|C|}$.

In this case, where $f_k = 1$ for all k, we have $\hat{\tau} = |C|\hat{r}$ where from the integral expression of $\hat{r}$ or directly we obtain $\hat{r} = -\frac{\hat{p}}{\hat{q}}\log(\hat{p})$ and $\hat{p} = |C|/F_C$. Using the δ method, we obtain after some calculations $\mathrm{Var}(\hat{\tau}) \approx |C|\,[p^2(1+\log(p)-p)^2/(1-p)^3]$.

For example, when $|C| = 100$ and $p = .05$ this variance estimate is 1.22, while $\mathrm{E}(\tau)$ is about 16. In summary, global measures are quite stable if they average on large enough sets C under reasonable conditions and under the A model, and the estimates have small variances. Large sets C correspond to coarse weighting classes, and for those ARGUS seems to work well, under suitable conditions. For small sets C ARGUS overestimates the risk. Finally we mention that calculations of this kind can be extended to produce confidence intervals for the risk, though these need to be properly defined for the A model. We present further results in this direction elsewhere.

Let us turn to the Negative Binomial assumption. We will provide here one example of what happens when this assumption is violated. Suppose that the true underlying model is similar to the above, and $F_k \mid f_k$ is distributed as $\sum_{i=1}^{f_k} X_i$ where $X_i \geq 1$ are iid. If $X_i \sim \mathrm{Geometric}(p_k)$ we have the above Negative Binomial model. Consider $X_i \sim Poisson(\lambda_k) + 1$. This is the size biased Poisson distribution (see below). How will the ARGUS estimates perform for data satisfying this assumption, for example rather than the A model assumption?

We comment that if one makes the not unreasonable assumption that family size has a Poisson distribution, and if sampling is done from a list of individuals and then all their family members enter the sample, then the above size biased Poisson model obtains. Neither this scheme, nor inverse sampling are used in samples such as the Family Expenditure Survey at the CBS in Israel, and neither model or any other can be assumed without further justification.

We compare the two models, the ARGUS' Negative Binomial, and the above Poisson model.
For the first we have $E_{p_k}[F_k \mid f_k] = f_k/p_k$ and $\mathrm{Var}_{p_k}[F_k \mid f_k] = f_k(1-p_k)/p_k^2$. For the second $E_{\lambda_k}[F_k \mid f_k] = f_k(1+\lambda_k)$ and $\mathrm{Var}_{\lambda_k}[F_k \mid f_k] = f_k\lambda_k$.
For $p_k = 1/(1+\lambda_k)$ the expectations are equal but the Negative Binomial variance is larger by a factor of $1+\lambda_k$. Roughly speaking, this is the order of the inflation factor, so for sampling fraction of 1/100, say, the Negative Binomial variance is 100 times larger. Can we expect any program to be indifferent to the variance difference between the two models?

In order to learn how ARGUS performs when its assumptions do not hold, we

generate $F_k \mid f_k = 1$ as a Poisson$(\lambda = 1/p_k - 1) + 1$ variable. As before, we concentrate on a single weighting class C, and assume without loss of generality that $f_k = 1$ for all $k \in C$. Since C is a weighting class, $\sum_{k \in C} F_k$ is known, and therefore ARGUS estimates $\hat{p}_k = |C| / \sum_{k \in C} F_k$ for all $k \in C$ where $|\cdot|$ denotes cardinality. Taking for example $|C| = 500$ and $p = 1/100$ (or $\lambda = 99$) we obtained that the real risk generated by sample uniques in C, defined by $\tau = \sum_{k \in C} 1/F_k$ varied very slightly in repeated simulations around $\tau = 5$. Its ARGUS type estimate $\hat{\tau} = \sum_{k \in C} \hat{r}_k$ varied very slightly around 23, a huge overestimate.

This is not surprising. The real risk is smaller since for parameters in the range of interest, the Poisson+1 distribution does not concentrate much mass around 1, unlike the Geometric whose mode is at 1, generating small and risky cells with high probability. ARGUS overestimates, thinking Negative Binomial.

It can be demonstrated in many other ways that deviations from the marginal Negative Binomial model, or introduction of some dependence among cells even under the marginal Negative Binomial assumption (see below), can produce arbitrarily large errors in the ARGUS estimates. Thus it is essential to test the model before relying on it. ARGUS and other models have been tried on certain data sets, however such programs do not usually supply means for testing their hypotheses on new samples, and it is not clear how a user is expected to know whether the assumptions hold.

Next we discuss the conditional independence assumption, and show that it is indeed necessary. Consider the simple situation where F_0 and G_k are independent $Geometric(p)$ and for some fixed s

$$F_k = \begin{cases} F_0 & F_0 \leq s \\ s + G_k & otherwise. \end{cases} \tag{2.1}$$

Then $F_k \sim NB(1, p)$, so the assumption of ARGUS (with $f_k = 1$) hold, however the F_k's are (conditionally on $f_k = 1$) dependent (in a positive way). In simulation runs of this model, say with 100 cells, $s = 15$ and $p = .05$, the ARGUS type estimate $\hat{\tau}$ overestimates τ by a factor varying between 2 and 3. Positive dependence implies overestimation. Positive dependence in not necessarily artificial. It distributes the population into the cells more evenly than suggested by the model. A relatively even distribution in tables is often obtained by design, trying to avoid small and huge cells.

In the next section we discuss an earlier model which turns out to be related to the ARGUS model. In particular we show that it allows, in some sense, to test goodness of fit of the A model.

3 A model of Bethlehem et al.

Bethlehem, Keller, and Pannekoek (1990) proposed an empirical Bayes model. We present it with some extensions (calling it henceforth the B model). Let γ_k be independent (random) parameters, and conditionally on γ_k the population cell frequencies F_k are independent where

$$\gamma_k \sim \Gamma(\alpha, \beta), \text{ and } \quad F_k|\gamma_k \sim Poisson(N\gamma_k)$$

with α and β such that $E\gamma_k = \alpha\beta = 1/K$, hence $EF_k = N/K$ and $E\sum_{k=1}^{K} F_k = N$. It follows that $P(F_k = x) = \frac{\Gamma(x+\alpha)}{\Gamma(\alpha)\Gamma(x+1)}(\frac{N\beta}{1+N\beta})^x(\frac{1}{1+N\beta})^\alpha$, $x = 0, 1, 2, \ldots$. We denote this distribution as $\widetilde{NB}(\alpha, \frac{1}{1+N\beta})$. In general the parameter α and β can be any positive number. Here, $\alpha\beta = 1/K$. For a natural number α this is the distribution of the number of **failures** (not coin tosses, note the difference from the above NB) until α successes are observed, with success probability being $p = \frac{1}{1+N\beta}$. We have $EF_k = \alpha(1-p)/p = N\alpha\beta$, and $\mathrm{Var}F_k = \alpha(1-p)/p^2 = N\alpha\beta(1+N\beta) = \frac{N}{K}(1+N\beta)$.

Assuming Poisson sampling, that is, $f_k \,|\, F_k \sim Bin(F_k, \pi_k)$, we have $f_k \,|\, \gamma_k \sim Poisson(N\gamma_k\pi_k)$ so that for f_k the same results as for F_k, with $N\pi_k$ replacing N. In particular $f_k \sim \widetilde{NB}(\alpha, \frac{1}{1+N\pi_k\beta})$ and $\mathrm{Var}f_k = \frac{N\pi_k}{K}(1+N\pi_k\beta)$. Bethlehem et al. considered only the case of sampling with equal probabilities (Bernoulli sampling), that is, $\pi_k = n/N$. We now continue with this assumption. Taking s^2 to be an estimate of $\mathrm{Var}(f_k)$ we obtain a moment estimate for β of the form $\hat{\beta} = \frac{s^2K-n}{n^2}$. An estimate of α is then obtained from $\alpha\beta = 1/K$. For s^2 we can take the usual unbiased sample estimate based on the f_k's, or using $Ef_k = n/K$, we can take $s^2 = \frac{1}{K}\sum_{k=1}^{K}(f_k - n/K)^2$.

Based on the B model, which is a standard empirical Bayes model, we propose to estimate the parameters of the prior as above and use them to test goodness of fit of the f_k's to the estimated Poisson (or mixture of Poissons) distribution.

Bethlehem et al. considered only the marginal distribution of F_k and used it to estimate the number of population uniques. We compute for $F \geq f$,

$$P(F_k = F|f_k = f) = P(f|F)\frac{P(F)}{P(f)} = \binom{\alpha+F-1}{F-f}(\frac{N\pi_k + 1/\beta}{N+1/\beta})^{\alpha+f}(\frac{N(1-\pi_k)}{N+1/\beta})^{F-f}.$$

This means that $F_k \,|\, f_k \sim NB(\alpha + f_k, \frac{N\pi_k+1/\beta}{N+1/\beta})$, with $F_k \geq f_k$.

As $\alpha \to 0$ (and hence $\beta \to \infty$) we obtain exactly the ARGUS assumption $F_k \,|\, f_k \sim NB(f_k, \pi_k)$. With this observation we see that the A model can be embedded in the empirical Bayes B model. For equal sampling probabilities π_k and equal inflation factors, the A model becomes a special case of the B model. ARGUS adds the varying sampling probabilities and inflation factors to the model.

As mentioned above, Bethlehem et al. assumed $\pi_k = n/N$ and obtained an estimate for β (and hence for α). In this case we propose to test goodness of fit of the data to the model by testing that the f_k's are iid $\widetilde{NB}(\alpha, \frac{1}{1+n\beta})$, for the estimated α and β. If the data fits the model with a small α, then the A model is justified (however, it is a sufficient condition only). If it fits for some other α, we can still use the model in a similar way to ARGUS (though it will no longer be the original A model). In general, for varying but known π_k's, we showed that $f_k \sim \widetilde{NB}(\alpha, \frac{1}{1+N\pi_k\beta})$. We discuss estimation and goodness of fit for this situation elsewhere. This approach allows testing the A model, and more general ones.

A difficulty that arises in the B model is the need to define K, the number of cells in the population table. Should structural zeros and other population cells which are known to be empty be counted? A closely related issue arises also in the A model. When the estimates $\hat{F}_k$ are constructed from the inflation factors, empty sample cells are not inflated even if they are not structural zeros, and therefore nonempty cells must be over inflated on the average. This biases risk estimates towards underestimation. In our experiments with both models this seemed to be a serious issue. One way to correct for this problem in ARGUS seems to be related to the need to define the non structural sample zeros, and inflate them too, or somehow reduce the inflation of nonempty cells. Another attempt to address this issue in the A model relies on the connection to the B model, and ideas suggested in Skinner et al. (1994) in terms of the B model. This is described briefly next (the details are given elsewhere).

Let K denote the number of non structural zeros. In the original B model it is considered known. Skinner at al. (1994) propose to take K to be an unknown parameter which will be estimated. The variance estimate $s^2 = \mathrm{Var}(f_k)$ must take the number of k's with $f_k = 0$ into account. However, since the number of empty cells may depend on the table definition, which may be arbitrary, we propose to estimate the (related) parameters α, β and K on the basis of the observed non empty cell frequencies f_k, having the $\widetilde{NB}$ distribution conditioned on being non zero. Skinner et al. (1994) propose a similar approach. Once K is estimated, the inflation factors in the A model can be corrected, and goodness of fit is possible along the lines described before for both models.

As mentioned above, Bethlehem et al. used their model to estimate the number of population uniques. A more relevant global risk measure considered in Skinner et al. (1994) in connection with the B model with $\pi_k = n/N$, and by others in other models, is $\tau_1 = \sum I(f_k = 1, F_k = 1)$, the number of population uniques which are in the sample. Of course, other relevant global measures under this model can be studied along the same line. We propose to study this measure also for varying π_k.

From previous calculation, $P(F_k = 1 \mid f_k = 1) = (\frac{N\pi_k + 1/\beta}{N + 1/\beta})^{\alpha+1}$; we can compute an estimate $\hat{\alpha}$ and using it we obtain a natural estimator of τ in the case $\pi_k = n/N$, is $\hat{\tau}_1 = \sum I(f_k = 1)(\frac{N\pi_k + 1/\hat{\beta}}{N + 1/\hat{\beta}})^{\hat{\alpha}+1}$. For $\pi_k = n/N$ this is the estimate proposed by Skinner et al.

Note that τ_1 corresponds to an intruder who does not guess, but rather looks only to match population uniques. If τ_1 is high it means that matching without guessing is possible in many cells, so that intruders can operate reliably. We discuss measure of intruders' reliability corresponding to risk measures elsewhere.

Finally we mention that Skinner et al. (1994) express doubts about the utility of the B model in practice and its fit to real data, on the basis of various experiments. Some of our experiments with ARGUS yielded good results, in others it performed poorly. In the second experiment in Benedetti et al. (2003), ARGUS estimates the low risks much better than the more important high risks. In our minds, these findings emphasize the need for statistical model testing prior to using model based disclosure risk measures, and for the use of a diversity of measures and models, rather than programs based on a single model.

Acknowledgement This work was prepared in collaboration with Natalie Shlomo; Theodor Yitzkov's contribution is also acknowledged. Both are from CBS, Israel. Micha Mandel has made thoughtful contributions to this work in several conversations.

References

Benedetti, R. and Franconi, L. (1998). An estimation method for individual risk of disclosure based on sampling design.

Benedetti, R., Capobianchi, A. and Franconi, L. (2003). Individual risk of disclosure using sampling design information

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

Fienberg, S. E. and U.E.Makov(1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. Journal of Official Statistics 14(4), 385-397.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998). Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics 14 (4), 485-502.

Hundepool, A. and L. Willenborg (1999, March). ARGUS: Software from the SDC project. Presented at the Joint ECE/Eurostat Work Session on Statistical Data Confdentiality, Thessaloniki, Greece, Working Paper No. 7.

Skinner, C.J. and M.J. Elliot (2001). A Measure of disclosure risk for microdata. http://www.ccsr.ac.uk/publications/occasion/occ23.pdf.

Skinner, C.J., Marsh C., Openshaw S., and Wymer, C. (1994). Disclosure control for census microdata. J.Official Statist.,10, 31-51.

Samuels, S.M. (1998) A Bayesian, species-sampling-inspired approach to the uniques problems in microdata disclosure risk assessment. J.Official Statist., 14, 373-383.

Willenborg, L. And T. de Waal (2001). Elements of Statistical Disclosure Control. Statistics Netherlands Springer Verlag Series: Lecture Notes in Statistics. VOL. 155.

Willenborg, L. and T. de Waal (1996). Statistical Disclosure Control in Practice Statistics Netherlands Springer Verlag Series: Lecture Notes in Statistics. VOL. 111.