Variances and Confidence Intervals for Sample Disclosure Risk Measures.

Rinott, Yosef
The Hebrew University
Jerusalem 91905, Israel
E-mail: rinott@mscc.huji.ac.il

Shlomo, Natalie
Central Bureau of Statistics and Hebrew University of Jerusalem
Kanfey Nesharim 66
Jerusalem 95464, Israel
E-mail: NatalieShlomo@cc.huji.ac.il

1. Introduction

Statistical Agencies have to assess the disclosure risk involved in the release of sample micro-data when the population is unknown or only partially known based on marginal distributions from current population estimates. When the sample is given in the form of a frequency table, *disclosure risk* arises when both the sample and the population have small counts in some cells defined by cross-classifying identifying key variables (i.e., sex, age, marital status, ethnicity, place of residence, etc.). This allows an "intruder" who has the sample data and access to some information on the population to identify an individual in the sample with high probability.

Various individual and global disclosure risk measures estimation methods have been proposed in the literature based on probabilistic models, see e.g., Bethlehem (1990), Benedetti, Capobianchi and Franconi (1998), Skinner and Holmes (1998), Elamir and Skinner (2006), Rinott (2003), Rinott and Shlomo (2006). In the context of such methods, the question of computing the variance of a given risk estimator, and a related confidence interval is natural, and was raised frequently in conferences.

In this paper we present a simple method for calculating approximate confidence intervals for global risk measures under probabilistic models. In Section 2 we provide a short description of the Poisson log-linear model which we use here, and describe our variance estimation procedure, based on the model selection method of Skinner and Shlomo (2006) which we briefly review. Section 3 provides simulation results based on a real data set drawn from the 2001 UK Census. We conclude in Section 4 with a discussion.

Calculating precise estimates and confidence interval for global disclosure risk measures is a hard problem due to the fact that risk measures are not ordinary parameters. In fact they depend both on the sample counts (random, observable data) and on the population counts (unobservable parameters), and therefore are not parameters in the classical sense. Furthermore, model based risk estimation requires model selection in order to obtain unbiased risk estimates in a sense described in Section 2, which is a very hard problem in the relevant sparse frequency tables, and as a result one must settle for approximations, and cannot expect very precise results.

An agency which considers releasing the sample is interested in the risk of the given sample and a confidence interval for it, and not in the potential risk of other samples of this type. As mentioned above, risk measures depend both on the sample and the population. A risk measure in a given sample and its variance may depend, for example, on the number of sample uniques and population uniques. Given the sample and a risk measure estimate, its variability and the need for a confidence interval is due to the fact that the population is unknown and assumed random in our model. A confidence interval should provide information on how precise the risk estimate is for this sample, and not across all possible random samples. Therefore,

for the problem at hand, it is natural to provide *conditional confidence intervals* with a coverage probability which is conditional on the given sample. Since our assumed structure is of a Bayesian type, we can consider confidence (or credible) intervals based on the posterior distribution of the parameter given the sample. However, since this distribution contains unknown parameters which are estimated from the sample, we are led naturally to *empirical Bayes confidence intervals*. For a general discussion of such intervals see, e.g., Laird and Louis (1989) and references therein, and for a discussion and references on the issue of estimating parameters of the type considered here which involve both the known sample and unknown parameters, with a brief discussion of the relevance to disclosure control, see Zhang (2005).

2. Risk Estimates and Probabilistic Models

Using the notation of Skinner and Shlomo (2006) and Rinott and Shlomo (2006), let $\mathbf{f} = \{f_k\}$ denote an *m*-way frequency table, which is a sample from a population table $\mathbf{F} = \{F_k\}$, where $k = (k_1,...,k_m)$ indicates a cell and f_k and F_k denote the frequency in the sample and in the population cell *k*, respectively. Formally the sample and population sizes in our models are random and their expectations are denoted by *n* and *N*, respectively and the number of cells by *K*. We can either assume that *n* and *N* are known, or that they are estimated by their natural estimators: the actual sample and population sizes, assumed to be known. Throughout this paper, when we write *n* or *N,* we formally refer to their expectations.

We assume that the *m* attributes in the table are the identifying key variables, i.e. variables which are accessible to the public or to potential intruders and can be used to identify individuals. Disclosure risk arises from cells in which both f_k and F_k are positive and small, and in particular when $f_k = F_k = 1$ (sample and population uniques). We focus here on two global disclosure risk measures which are based on sample uniques, but the approach is easily extended to similar measures. We consider

$$\tau_1 = \sum_k I(f_k = 1, F_k = 1), \quad \tau_2 = \sum_k I(f_k = 1) 1/F_k,$$

where *I* denotes the indicator function. Note that τ_1 counts the number of sample uniques which are also population uniques and τ_2 is the expected number of correct guesses if each sample unique is matched to a randomly chosen individual from the same population cell. Note also that τ_i, $i = 1,2$ involve both the sample and the population, and therefore they are not ordinary parameters.

We consider the case that $\mathbf{f}$ is known, and $\mathbf{F}$ is an unknown parameter (on which there may be some partial information, and a prior family of distributions) and the quantities τ_1 and τ_2 should be estimated. The methods discussed in this section consist of modeling the conditional distribution of $\mathbf{F} \mid \mathbf{f}$, estimating parameters in this distribution and then using estimates of the form:

$$\hat{\tau}_1 = \sum_k I(f_k = 1) \hat{P}(F_k = 1 \mid f_k = 1), \quad \hat{\tau}_2 = \sum_k I(f_k = 1) \hat{E}[1/F_k \mid f_k = 1], \tag{1}$$

where $\hat{P}$ and $\hat{E}$ denote estimates of the relevant conditional probability and expectation. Note that the quantities in (1) are in fact estimates of $E(\tau_i \mid f)$ for $i = 1,2$ rather than τ_i.

The method described in this paper is based on the Poisson distribution and the use of log-linear models to estimate the parameters of the distribution, however other models could be considered with the present approach.

2.1 Variance Estimates and Confidence Intervals

We first describe the assumptions and method in general, and in Section 2.2, we describe the Poisson log-linear model which we shall use.

In general we assume that $\{F_k \mid f_k\}$ are (conditionally) independent with a distribution depending on a parameter λ_k which can be estimated from the observations $\{f_k\}$. Consider first $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$. Given the sample $\{f_k\}$ it is a sum of Bernoulli random variables over sample uniques, taking the value one with probability $P(F_k = 1 \mid f_k = 1)$. Thus

$$Var(\tau_1 \mid \mathbf{f}) = \sum_k I(f_k = 1) P(F_k = 1 \mid f_k = 1)(1 - P(F_k = 1 \mid f_k = 1)). \tag{2}$$

As in (1) we estimate the variance of (2) by

$$\hat{Var}(\tau_1 \mid \mathbf{f}) = \sum_k I(f_k = 1) \hat{P}(F_k = 1 \mid f_k = 1)(1 - \hat{P}(F_k = 1 \mid f_k = 1)), \tag{3}$$

provided we can estimate the indicated conditional probabilities (see Section 2.2). In a similar way we have

$$Var(\tau_2 \mid \mathbf{f}) = \sum_k I(f_k = 1) Var(1/F_k \mid f_k = 1), \quad (4)$$

which is estimated by replacing the latter conditional variance by its estimate based on the estimated conditional distribution of $\{F_k \mid f_k\}$ to yield

$$\hat{Var}(\tau_2 \mid \mathbf{f}) = \sum_k I(f_k = 1) \hat{Var}(1/F_k \mid f_k = 1). \quad (5)$$

In order to construct confidence intervals, we use the approximation that conditionally on $\mathbf{f} = \{f_k\}$, $\tau_i - E(\tau_i \mid \mathrm{f})$ is normally distributed with variance as in (2) for τ_1 and (4) for τ_2. Since we do not observe $E(\tau_i \mid \mathbf{f})$ we replace them by their estimates $\hat{\tau}_i \quad i = 1,2$ given in (1). Following Skinner and Shlomo (2006), see details in Section 2.2, we assume a Poisson log-linear model, and select the model which minimizes the bias (or rather an estimate thereof)

$$B_i = \hat{\tau}_i - E(\tau_i \mid \mathrm{f}) \quad (6)$$

and then use the approximate confidence interval of the type

$$\hat{\tau}_i \pm 2\sqrt{\hat{Var}(\tau_i \mid \mathbf{f})}, \quad (7)$$

where the variance estimates are those of (3) and (5). Since we replace $E(\tau_i \mid \mathrm{f})$ by $\hat{\tau}_i$, this is a reasonable approximation provided B_i is indeed small, and thus the utility of this approximation depends on the quality of the model selection and parameter estimation. Here we rely on the method of Skinner and Shlomo (2006), however, the approach we present here can be adapted to other models and model selection methods.

2.2 The Poisson Log-Linear Model

A common assumption in the frequency table literature is $F_k \sim Poisson(\lambda_k)$, independently, where $\sum_k F_k = N$ is a random parameter. Binomial (or Poisson) sampling from F_k means that $f_k \mid F_k \sim Bin(F_k, \pi_k)$ independently, where π_k is the sampling fraction in cell *k*. By standard calculations we then have:

$$f_k \sim Poisson(\lambda_k \pi_k) \text{ and, } F_k \mid f_k \sim f_k + Poisson(\lambda_k (1 - \pi_k)), \quad (8)$$

where $F_k \mid f_k$ are conditionally independent.

Assuming a simple random sample design where $\pi_k = \pi = n/N$ and setting $\mu_k = \lambda_k \pi$, Skinner and Holmes (1998) and Elamir and Skinner (2006) proposed using the sample counts $\{f_k\}$ to fit a log-linear model: $\log \mu_k = x_k' \beta$ in order to obtain estimates for the parameters: $\hat{\lambda}_k = \hat{\mu}_k / \pi$. Assuming that f_k are the outcomes of independent $Poisson(\pi \lambda_k)$ random variables, the maximum likelihood (MLE) estimator $\hat{\beta}$ may be obtained by solving the score equations: $\sum_k [f_k - \exp(x_k' \beta)] x_k = 0$.

Using the second part of (8), it is easy to compute the expected individual disclosure risk measures for cell *k*, defined by:

$$P_{\lambda_k}(F_k = 1 \mid f_k = 1) = e^{-\lambda_k (1-\pi)}, \quad E_{\lambda_k}(1/F_k \mid f_k = 1) = [1 - e^{-\lambda_k (1-\pi)}]/[\lambda_k (1-\pi)]. \quad (9)$$

Plugging $\hat{\lambda}_k$ for λ_k in (3) leads to the desired estimates $\hat{P}_{\hat{\lambda}_k}(F_k = 1 \mid f_k = 1)$ and $\hat{E}_{\hat{\lambda}_k}[1/F_k \mid f_k = 1]$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1).

2.3 Model Selection

Skinner and Shlomo (2006) developed a method of selecting the model for risk estimation based on estimating and (approximately) minimizing the bias $B_i, \ i = 1,2$ of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. The bias of $\hat{\tau}_i, \ i = 1,2$ is estimated by statistics denoted by $\hat{B}_i, \ i = 1,2$ which are asymptotically normal with variances which can also be estimated under the model.

We briefly review the procedure of Skinner and Shlomo (2006), with a slight change of notation.

Write $E(\tau \mid \mathbf{f}) = \sum_k I(f_k = 1) h(\lambda_k)$ and $\hat{\tau} = \sum_k I(f_k = 1) h(\hat{\lambda}_k)$ where for τ_1 we take $h(\lambda_k) = P(F_k = 1 \mid f_k = 1)$ and $h(\hat{\lambda}_k) = \hat{P}(F_k = 1 \mid f_k = 1)$, the estimate of $h(\lambda_k)$ obtained by plugging in the estimated parameters $\hat{\lambda}_k$. Similarly for τ_2 set $h(\lambda_k) = E(1/F_k \mid f_k = 1)$ and $h(\hat{\lambda}_k) = \hat{E}(1/F_k \mid f_k = 1)$. We have the approximation

$$B = \sum_k E[I(f_k = 1)][h(\hat{\lambda}_k) - h(\lambda_k)]. \quad (10)$$

A Taylor expansion of h, leads to the approximation

$$B \approx \sum_k \pi\lambda_k \exp(-\lambda_k)[h'(\lambda_k)(\hat{\lambda}_k - \lambda_k) + h''(\lambda_k)(\hat{\lambda}_k - \lambda_k)^2 / 2]$$

and the relations $Ef_k = \pi\lambda_k$ and $E[(f_k - \pi\hat{\lambda}_k)^2 - f_k] = \pi^2 E(\lambda_k - \hat{\lambda}_k)^2$ lead to a further approximation of B of the form

$$\hat{B} \approx \sum_k \hat{\lambda}_k \exp(-\pi\hat{\lambda}_k)[-h'(\hat{\lambda}_k)(f_k - \pi\hat{\lambda}_k) + h''(\hat{\lambda}_k)[(f_k - \pi\hat{\lambda}_k)^2 - f_k]/(2\pi)].$$

The method of Skinner and Shlomo (2006) selects the model which minimizes the standardized bias estimate $\hat{B}_i / \sqrt{\hat{v}_i}$, $i = 1,2$, where $\hat{v}_i$ are estimates of the variance of $\hat{B}_i$ which will not be discussed here.

2.4 Variance Estimates and Confidence Intervals under The Poisson Model

Using the second part of (8), we have $P(F_k = 1 \mid f_k = 1) = \exp(-\lambda_k(1-\pi))$ and therefore we obtain

$$\hat{Var}(\tau_1 \mid \mathrm{f}) = \sum_k I(f_k = 1)\exp(-\hat{\lambda}_k(1-\pi)[1 - \exp(-\hat{\lambda}_k(1-\pi)]. \quad (11)$$

For τ_2 we use (8) again to compute a series approximation of $Var(1/F_k \mid f_k = 1)$ as a function of λ_k and plug in $\hat{\lambda}_k$ to obtain the estimate $\hat{Var}(\tau_2 \mid \mathbf{f})$.

Now the confidence intervals of (7) can be computed using $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1) (using (9) as described above), and the variance $\hat{Var}(\tau_1 \mid \mathbf{f})$ of (11) and the corresponding $\hat{Var}(\tau_2 \mid \mathrm{f})$.

Taking into account that we have done many different approximations such as Taylor series and normal approximations, and used plug-in estimates, we must expect a rather imprecise coverage level of the intervals. This will be studied by simulations.

3. A Simulation Study

We assess the coverage rate and utility of the proposed confidence interval method by three types of simulation experiments which we now describe.

Part 1. A frequency table of real data taken from the 2001 UK Census representing a population of size N was constructed. A random sample of size *n* was drawn from this population. Since both the population and the sample are known, we can compute the true risk measures τ_1 and τ_2.

A model selection procedure as in Section 2.3 was performed on the basis of the sample only, (however, due to complexity of computations we sometimes used heuristic algorithms which do not look at all the models, similar to the well-known forward or backward selection algorithms). The corresponding parameters $\hat{\lambda}_k$ were estimated. The risk measure estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1) and (9) and the variance $\hat{Var}(\tau_1 \mid \mathbf{f})$ of (11) and the corresponding $\hat{Var}(\tau_2 \mid \mathbf{f})$ were then computed along with the resulting confidence intervals of (7). Since τ_1 and τ_2 are known, we can see if the confidence intervals cover them, or whether they seem to be in the ballpark. This was repeated several times and the results are given in Table 1.

The error in the coverage rate is due to many approximations as described above and the fact that the Census population used cannot be expected to follow any Poisson log-linear model precisely. The next simulation is aimed at eliminating the latter reason.

Part 2. We chose a log-linear model which fits the population in one of the examples of Part 1, and estimated its parameters. Using these parameters we generated a new population using $F_k \sim Poisson(\lambda_k)$ independently. We now have a simulated population which satisfies the assumptions of the model. We continue as in Part 1 above, that is, we draw a sample from the generated population, select a model for risk estimation and estimate the parameters, construct the confidence intervals, and check their coverage rate by repeating the whole experiment 100 times, that is, with the same parameters we generated 99 more populations, samples, etc., (except that we did not select a new model for the risk estimation in each repetition, but rather used the model selected based on the first sample; this is expected to reduce the coverage rate). The intervals are based on a normal approximation, model selection and parameter estimates, and are therefore approximate; however, the population now satisfies the assumptions of the model.

Part 3. Now we generate the population as in Part 2 and draw a sample. However, the risk estimates are computed using the same parameters that were used to generate the population and not their estimates. This eliminates another source of noise, and should improve the coverage rate, which is again estimated by repeating the experiment 100 times.

In Table 1, we present results of Part 1 for some real data sets taken from the 2001 UK Census. The population size *N*, the sample size *n*, the number of cells in the table *K* and the attributes (key variables) are given in each table with the number of categories in each attribute in parentheses. The table shows the resulting parameters τ_i and their estimates $\hat{\tau}_i$, the selection parameters $\hat{B}_i / \sqrt{\hat{v}_i}$, and the confidence intervals. The confidence intervals contain the true τ_1 in all but one experiment, while τ_2 is contained in about 50% of the experiments, however when it is outside the interval, it is very close and the approximation seems reasonable and useful.

Table 1: Coverage rates for the estimated confidence intervals of the disclosure risk measures of samples from a UK Census population as described in Part 1.

Model	τ_1	$\hat{\tau}_1$	$\hat{B}_1 / \sqrt{\hat{v}_1}$	$\hat{\tau}_1 \pm 2\sqrt{v(\hat{\tau}_1 \mid \mathbf{f})}$	τ_2	$\hat{\tau}_2$	$\hat{B}_2 / \sqrt{\hat{v}_2}$	$\hat{\tau}_2 \pm 2\sqrt{v(\hat{\tau}_2 \mid \mathbf{f})}$
Example 1: N=1,468,255 K=5,563,080 area(3),sex(2),age(101),marital status(6),ethnicity(17),work status(10),religion(9)								
n= 7,341	212	195.7	1.31	178.8 - 212.6	444.3	421.4	0.80	410.4 - 432.4
n=14,683	359	384.0	1.10	360.6 - 407.4	783.5	800.3	0.74	784.9 - 815.7
Example 2: N=1,468,255 K=618,120 area(3),sex(2),age(101),marital status(6),ethnicity(17),work status(10)								
n= 7,341	88	87.4	0.87	75.3 - 99.5	225.2	211.8	1.16	203.9 - 219.7
n=14,683	167	167.5	-0.17	150.4 - 184.6	408.1	423.2	0.53	411.9 - 434.7
n=14,683	147	171.5	0.60	154.4 - 188.6	403.6	413.3	0.57	402.2 - 424.4
n=14,683	180	190.0	0.90	172.9 - 207.1	435.6	423.6	0.27	412.6 - 434.6
Example 3: N=1,468,255 K=741,744 area(36),sex(2),age(101),marital status(6),ethnicity(17)								
n=14,683	142	142.2	0.09	126.4 - 158.0	379.0	376.2	1.51	365.7 - 386.7
n=14,683	136	135.7	1.02	120.1 - 151.3	364.2	366.0	0.55	355.6 - 376.4
n=14,683	152	146.5	-0.65	130.7 - 162.3	373.0	373.5	-0.03	363.0 - 384.0
Example 4: N=944,793 K=412,080 area(2),sex(2),age(101),marital status(6),ethnicity(17),work status(10)								
n= 9,448	159	152.2	0.88	136.5 - 167.9	355.9	343.3	1.73	332.1 - 354.5
n=18,896	263	277.0	-0.27	255.7 - 298.3	628.9	638.1	0.92	625.8 - 650.9
Example 5: N=51,620,597, K=443,520(*) region (11), age (96), sex (2), number of residents(7), marital status(6), number of cars(5)								
n= 16,651	18.0	18.6	-0.21	13.0 - 24.2	83.0	83.9	0.06	29.9 - 87.9
n= 54,560	24	33.8	1.24	25 - 42.6	220.3	211.0	0.54	204.6 - 217.4
n=119,618	64	74.4	0.56	61.6 - 87.2	446.6	441.7	0.03	432.4 - 451.0
n=357,888	211	189.5	-0.08	168.4 - 210.6	1193.8	1147.2	0.21	1131.6 - 1162.8

(*) These samples are based on a complex survey design based on clustered samples.

Table 2 shows the results of the simulations in Part 2. The data was generated using the model which was selected in the first experiment of Example 2 of Table 1 among those based on *n=14,683.* This model contains six interaction terms, {area*work status}, {sex*work status}, {age*marital status}, {sex*marital status}, {age*ethnicity}, {age*work status} and parameters estimated from the whole original population. For each population generated, a sample of *n=14,683* was selected and the risk measures estimated. We selected separate models for the estimated risk measures τ_1 and τ_2 in order to ensure a small (and positive) $\hat{B}_i / \sqrt{\hat{v}_i}$, *i*=1,2 statistic for each of the measures as described in Skinner and Shlomo (2006). Based on the chosen model the experiment was repeated a further 99 times as described in Part 2. The models chosen were:

- For τ_1 , the following interactions were included in the model: {area*age}, {area*ethnicity}, {area*work status}, {sex*marital status}, {sex*work status}, {age*marital status}, {age*ethnicity}, {marital status*ethnicity}, {ethnicity*work status}. This produced for a true $\tau_1 = 256$ an estimate of $\hat{\tau}_1 = 246.1$ $(\hat{B}_1 / \sqrt{\hat{v}_1} = 0.30)$.
- For τ_2, the following interactions were included in the model: {area*work status}, {sex*marital status}, {sex*work status}, {age*ethnicity}, {age*work status}, {marital status*ethnicity}. This produced for a true $\tau_2 = 504.2$ an estimate of $\hat{\tau}_2 = 504.3$ $(\hat{B}_2 / \sqrt{\hat{v}_2} = 0.98)$.

As shown in Table 2, 74% of the repeated samples had the true risk measure τ_1 within the confidence interval of $\hat{\tau}_1 \pm 2\sqrt{\hat{V}ar(\tau_1 \mid \mathrm{f})}$, and 93% in the interval of $\hat{\tau}_1 \pm 3\sqrt{\hat{V}ar(\tau_1 \mid \mathbf{f})}$. The results for the risk measure τ_2 are similar. From these results, we suggest using confidence intervals based on three standard deviations, say, in order to take into account the numerous approximations of the method.

Table 2: Coverage rates for the confidence intervals of the disclosure risk measures for Part 2.

Disclosure risk measures	Number of samples within confidence interval out of 100		
	$\pm 2\sqrt{\hat{V}ar(\tau_i \mid \mathrm{f})}$	$\pm 2.5\sqrt{\hat{V}ar(\tau_i \mid \mathbf{f})}$	$\pm 3\sqrt{\hat{V}ar(\tau_i \mid \mathbf{f})}$
τ_1	74%	89%	93%
τ_2	76%	84%	93%

For Part 3 one should expect coverage close to 95% for confidence intervals of $\pm$ two standard deviations, since the only approximations are that of τ_i to the normal, and the strong law approximation of replacing the risk measure by its expectation based on the true (rather than estimated as in Part 2) parameters. These are good approximations in the presence of a large number of sample uniques. Indeed we obtained that the percentage of samples that were in the confidence interval out of 100 repeated experiments was 95% for τ_1 and 94% for τ_2.

4. Discussion

The precision of an estimate is often described by a confidence interval. For risk measures estimates, confidence intervals would allow to compare the risk in different samples, and to identify samples which are significantly more risky than others.

We presented a method of computing approximate confidence intervals for estimated risk measures of a given sample. The risk measures involve both the given sample and the unknown population which is modeled as random. The confidence level is conditional on the given sample. The method is based on a Poisson log-linear model, and the model selection procedure of Skinner and Shlomo (2006). Besides the assumptions of the model, the method involves various approximations such as a normal approximation to a

sum of random variables, the approximations involved in the model selection procedure and the subsequent parameter estimation, and in using plug-in estimators of the risk measure and their variances.

Nevertheless, the simulations indicate that these rough approximations yield reasonable results, and given a risk measure estimate for a given sample, we can obtain a useful measure on its precision by using the proposed confidence intervals.

The method was tried also on other models such as the Poisson and Negative Binomial smoothing polynomial described in Rinott and Shlomo (2006). Our impression is that given a good model selection procedure, which in the latter paper means selection of the degree of the smoothing polynomial and neighborhood sizes (or weights) for the fit, confidence intervals can be constructed in the same way. In fact the present approach applies to any probabilistic approach, such as Argus, see Benedetti, et al. (1998), in cases where its assumptions are realistic, and provided it comes with a method which yields roughly unbiased estimates.

REFERENCES

Benedetti, R., Capobianchi, A., and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design. Contributi Istat.

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

Elamir, E. and Skinner, C.J. (2006) Record-Level Measures of Disclosure Risk for Survey Micro-data. *Journal of Official Statistics, 22*, 525-539.

Laird, N.M., and Louis, T.A. (1989) Empirical Bayes Confidence Intervals for a Series of Related Experiments. *Biometrics* 45, 481-495.

Rinott, Y. (2003) On Models for Statistical Disclosure Risk Estimation. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxemburg, 275-285.

Rinott, Y. and Shlomo, N (2006) A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In *PSD'2006 Privacy in Statistical Databases*, (Eds. J. Domingo-Ferrer and L. Franconi), Springer LNCS 4302, 82-93.

Skinner, C. J. and Holmes, D. (1998) Estimating the Re-identification Risk per Record in Microdata. *Journal of Official Statistics 14*, 361-372.

Skinner, C. and Shlomo, N. (2006) Assessing Identification Risk in Survey Microdata Using Log-linear Models. Submitted. See: Southampton Statistical Sciences Research Institute, Working Paper M06/14.

Zhang, C-H (2005) Estimation of Sums of Random Variables: Examples and Information Bounds. *Annals of Statistics 33*, 2022-2041.

RÉSUMÉ (ABSTRACT)

Nous étudions l'évaluation du risque de révélation d'un échantillon de micro-données sous forme d'une table de fréquence. Lorsque l'on calcule des estimateurs de mesures de risque, la question de leur précision se pose naturellement. Nous répondons à cette question en fournissant des estimateurs de la variance et des intervalles de confiance approchés pour une certaine classe d'estimateurs de mesure de risque basés sur la structure de la table et sur un modèle probabiliste Bayésien. Les intervalles sont basés sur la distribution a posteriori de la mesure de risque estimée, qui implique une estimation de paramètres, et sont par conséquent des intervalles de Bayes empiriques.

Les niveaux de couverture approchés de ces intervalles sont fonction de l'échantillon. Une étude par simulation sur des données synthétiques et des données réelles montre l'utilité de la méthode.