UNECE Work session on statistical data confidentiality
Geneva, 9–11 November 2005

Assessing risk in statistical disclosure limitation

Silvia Polettini

Department of Statistics
University of Naples *Federico II*
Italy

Julian Stander

Department of Mathematics and Statistics
University of Plymouth
UK

Disclosure Risk

- The release of microdata files for research may lead to disclosure.

Disclosure Risk

- The release of microdata files for research may lead to disclosure.
- Disclosure: a correct re-identification operation achieved by comparing a target individual in a released sample with a list that contains individual identifiers from an archive or population.

Disclosure Risk

- The release of microdata files for research may lead to disclosure.
- Disclosure: a correct re-identification operation achieved by comparing a target individual in a released sample with a list that contains individual identifiers from an archive or population.

Spy:	Name	KV 1: Sex	KV 2: Age	KV 3: Region

Released:	KV 1: Sex	KV 2: Age	KV 3: Region	Income

Disclosure Risk

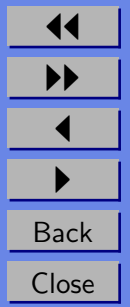
- The release of microdata files for research may lead to disclosure.

▶ Disclosure: a correct re-identification operation achieved by comparing a target individual in a released sample with a list that contains individual identifiers from an archive or population.

Spy:	Name	KV 1: Sex	KV 2: Age	KV 3: Region

Released:	KV 1: Sex	KV 2: Age	KV 3: Region	Income

▶ Social surveys: the released variables are often categorical and usually comprise publicly available variables (sex, age, region).



Disclosure Risk

- The release of microdata files for research may lead to disclosure.
- Disclosure: a correct re-identification operation achieved by comparing a target individual in a released sample with a list that contains individual identifiers from an archive or population.

Spy:	Name	KV 1: Sex	KV 2: Age	KV 3: Region

Released:	KV 1: Sex	KV 2: Age	KV 3: Region	Income

  - Social surveys: the released variables are often categorical and usually comprise publicly available variables (sex, age, region).

- Public variables that allow identification are called key variables (KV).

Disclosure Risk

- The release of microdata files for research may lead to disclosure.
- Disclosure: a correct re-identification operation achieved by comparing a target individual in a released sample with a list that contains individual identifiers from an archive or population.

Spy:	Name	KV 1: Sex	KV 2: Age	KV 3: Region

Released:	KV 1: Sex	KV 2: Age	KV 3: Region	Income

  - Social surveys: the released variables are often categorical and usually comprise publicly available variables (sex, age, region).

- Public variables that allow identification are called key variables (KV).
- **Disclosure risk** is specific to a **cell** in the **contingency table** built by **cross-tabulating** the key variables.

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

- Let f_k be the corresponding observed sample frequency for cell k

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

- Let f_k be the corresponding observed sample frequency for cell k
- Given F_k, the probability of re-identifying an individual coming from cell k is

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

- Let f_k be the corresponding observed sample frequency for cell k

- Given F_k, the probability of re-identifying an individual coming from cell k is

$$1/F_k.$$

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

- Let f_k be the corresponding observed sample frequency for cell k
- Given F_k, the probability of re-identifying an individual coming from cell k is

$$1/F_k.$$

- F_k unknown:

▶ We define the *re-identification risk* as

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

- Let f_k be the corresponding observed sample frequency for cell k
- Given F_k, the probability of re-identifying an individual coming from cell k is

$$1/F_k.$$

- F_k unknown:

▶ We define the *re-identification risk* as

$$\mathrm{E}\left(1/F_k \mid f_1, \ldots, f_K\right).$$

Our **measure of disclosure risk** as follows:

- Let F_k be the number of individuals in the population belonging to cell k, $k = 1, \ldots, K$

  (K: the number of combinations in the population).

- Let f_k be the corresponding observed sample frequency for cell k
- Given F_k, the probability of re-identifying an individual coming from cell k is

$$1/F_k.$$

- F_k unknown:

▶ We define the *re-identification risk* as

$$\mathrm{E}\left(1/F_k \mid f_1, \ldots, f_K\right).$$

[Fienberg and Makov (1998); Omori (1998); Takemura (1998); Forster (2004); Benedetti and Franconi (1998)]

Estimation

Our approach is as follows:

- introduce a superpopulation model that describes the population and sample frequencies $\underline{F} = (F_1, \ldots, F_K), \underline{f} = (f_1, \ldots, f_K)$ (Bayesian hierarchical model)

Estimation

Our approach is as follows:

- introduce a superpopulation model that describes the population and sample frequencies $\underline{F} = (F_1, \ldots, F_K), \underline{f} = (f_1, \ldots, f_K)$ (Bayesian hierarchical model)
- derive the posterior distribution $[F_k \mid f_1, \ldots, f_K]$ of population frequencies given sample frequencies

Estimation

Our approach is as follows:

- introduce a superpopulation model that describes the population and sample frequencies $\underline{F} = (F_1, \ldots, F_K), \underline{f} = (f_1, \ldots, f_K)$ (Bayesian hierarchical model)
- derive the posterior distribution $[F_k \mid f_1, \ldots, f_K]$ of population frequencies given sample frequencies
- use this posterior to estimate the risk $r_k = \mathrm{E}(1/F_k \mid f_1, \ldots, f_K)$

Estimation

Our approach is as follows:

- introduce a superpopulation model that describes the population and sample frequencies $\underline{F} = (F_1, \ldots, F_K), \underline{f} = (f_1, \ldots, f_K)$ (Bayesian hierarchical model)
- derive the posterior distribution $[F_k \mid f_1, \ldots, f_K]$ of population frequencies given sample frequencies
- use this posterior to estimate the risk $r_k = \mathrm{E}(1/F_k \mid f_1, \ldots, f_K)$

EB: use Empirical Bayes approach (Efron and Morris) to estimate model parameters using the observed data distribution; then substitute these estimates into $[F_k|\underline{f}]$ to obtain an estimate of risk.

Some Superpopulation Models

We describe a variety of Bayesian Hierarchical models, all of which share an assumption of independence that implies that r_k is defined as $\mathrm{E}(1/F_k \mid f_k)$ instead of $\mathrm{E}(1/F_k \mid \underline{f})$.

Some Superpopulation Models

We describe a variety of Bayesian Hierarchical models, all of which share an assumption of independence that implies that r_k is defined as $\mathrm{E}(1/F_k \mid f_k)$ instead of $\mathrm{E}(1/F_k \mid \underline{f})$.

First we consider the following models:

Some Superpopulation Models

We describe a variety of Bayesian Hierarchical models, all of which share an assumption of independence that implies that r_k is defined as $\mathrm{E}(1/F_k \mid f_k)$ instead of $\mathrm{E}(1/F_k \mid \underline{f})$.

First we consider the following models:

Model I Benedetti-Franconi (1998) model

Model II Bethlehem, Keller and Pannekoek (1990)-type model

Model III a model by Polettini and Stander (2004)

Model IV a modification of the model by Polettini and Stander (2004)

Some Superpopulation Models

We describe a variety of Bayesian Hierarchical models, all of which share an assumption of independence that implies that r_k is defined as $\mathrm{E}(1/F_k \mid f_k)$ instead of $\mathrm{E}(1/F_k \mid \underline{f})$.

First we consider the following models:

Model I Benedetti-Franconi (1998) model

Model II Bethlehem, Keller and Pannekoek (1990)-type model

Model III a model by Polettini and Stander (2004)

Model IV a modification of the model by Polettini and Stander (2004)

Next we describe the characteristics of these models and compare with

Some Superpopulation Models

We describe a variety of Bayesian Hierarchical models, all of which share an assumption of independence that implies that r_k is defined as $\mathrm{E}(1/F_k \mid f_k)$ instead of $\mathrm{E}(1/F_k \mid \underline{f})$.

First we consider the following models:

Model I Benedetti-Franconi (1998) model

Model II Bethlehem, Keller and Pannekoek (1990)-type model

Model III a model by Polettini and Stander (2004)

Model IV a modification of the model by Polettini and Stander (2004)

Next we describe the characteristics of these models and compare with

Model V a Dirichlet-multinomial-multinomial model.

- Let the microdata file be a random sample of size n drawn from a finite population of N units.

- Let the microdata file be a random sample of size n drawn from a finite population of N units.

  Define

- Let the microdata file be a random sample of size n drawn from a finite population of N units.

  Define

- $\pi_k = P(\text{a member of the population falls into cell } k)$

- Let the microdata file be a random sample of size n drawn from a finite population of N units.

Define

- $\pi_k = P(\text{a member of the population falls into cell } k)$
- $p_k = P(\text{a member of population cell } k \text{ falls into the sample})$

- A model by Benedetti and Franconi (1998):

- A model by Benedetti and Franconi (1998):
- $F_k | f_k, p_k \sim$ negative binomial(f_k, p_k)

- A model by Benedetti and Franconi (1998):
- $F_k | f_k, p_k \sim$ negative binomial(f_k, p_k)

  Rinott (2003) showed that this model can be derived from the following Bayesian hierarchical model:


- A model by Benedetti and Franconi (1998):
- $F_k | f_k, p_k \sim$ negative binomial(f_k, p_k)

  Rinott (2003) showed that this model can be derived from the following Bayesian hierarchical model:

$$\begin{aligned} \pi_k &\sim m(\pi_k) \propto 1/\pi_k \\ F_k | \pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k | F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- A model by Benedetti and Franconi (1998):
- $F_k | f_k, p_k \sim$ negative binomial(f_k, p_k)

  Rinott (2003) showed that this model can be derived from the following Bayesian hierarchical model:

$$\begin{aligned} \pi_k &\sim m(\pi_k) \propto 1/\pi_k \\ F_k | \pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k | F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model I.

- A model by Benedetti and Franconi (1998):
- $F_k|f_k, p_k \sim$ negative binomial(f_k, p_k)

  Rinott (2003) showed that this model can be derived from the following Bayesian hierarchical model:

$$\begin{aligned} \pi_k &\sim m(\pi_k) \propto 1/\pi_k \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model I.

- Note that the hyperprior for π_k is improper, so that EB for parameter estimation is not feasible as $[f_k]$ is also improper.

- A model by Benedetti and Franconi (1998):
- $F_k | f_k, p_k \sim$ negative binomial(f_k, p_k)

  Rinott (2003) showed that this model can be derived from the following Bayesian hierarchical model:

$$\begin{aligned} \pi_k &\sim m(\pi_k) \propto 1/\pi_k \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model I.

- Note that the hyperprior for π_k is improper, so that EB for parameter estimation is not feasible as $[f_k]$ is also improper.
- BF use $\hat{p}_k = f_k/\hat{F}_k^D$, where $\hat{F}_k^D$ is an estimate of F_k using the sampling design weights. This can sometimes be problematic.

- A Bethlehem, Keller and Pannekoek (1990)-type model:

- A Bethlehem, Keller and Pannekoek (1990)-type model:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha); \mathsf{E}[\pi_k] = \frac{1}{K}, \mathsf{Var}[\pi_k] = \frac{1}{K^2\alpha} \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- A Bethlehem, Keller and Pannekoek (1990)-type model:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha); \mathsf{E}[\pi_k] = \frac{1}{K}, \mathsf{Var}[\pi_k] = \frac{1}{K^2\alpha} \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

[in the formalisation due to Rinott]

- A Bethlehem, Keller and Pannekoek (1990)-type model:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha); \mathsf{E}[\pi_k] = \frac{1}{K}, \mathsf{Var}[\pi_k] = \frac{1}{K^2\alpha} \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model II. [in the formalisation due to Rinott]

- A Bethlehem, Keller and Pannekoek (1990)-type model:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha); \mathsf{E}[\pi_k] = \frac{1}{K}, \mathsf{Var}[\pi_k] = \frac{1}{K^2\alpha} \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model II. [in the formalisation due to Rinott]

- If $\alpha \to 0$, then the gamma prior for π_k tends to the improper prior $m(\pi_k) \propto 1/\pi_k$ of Model I.

- A Bethlehem, Keller and Pannekoek (1990)-type model:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha); \mathsf{E}[\pi_k] = \frac{1}{K}, \mathsf{Var}[\pi_k] = \frac{1}{K^2\alpha} \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model II. [in the formalisation due to Rinott]

- If $\alpha \to 0$, then the gamma prior for π_k tends to the improper prior $m(\pi_k) \propto 1/\pi_k$ of Model I.

  Hence, as Rinott (2003) showed, the posterior distribution tends to

$$F_k|f_k, p_k \sim \text{negative binomial}(f_k, p_k)$$

  used by Benedetti and Franconi (1998).

- A Bethlehem, Keller and Pannekoek (1990)-type model:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha); \mathsf{E}[\pi_k] = \frac{1}{K}, \mathsf{Var}[\pi_k] = \frac{1}{K^2\alpha} \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), \ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model II. [in the formalisation due to Rinott]

- If $\alpha \longrightarrow 0$, then the gamma prior for π_k tends to the improper prior $m(\pi_k) \propto 1/\pi_k$ of Model I.

  Hence, as Rinott (2003) showed, the posterior distribution tends to

$$F_k|f_k, p_k \sim \text{negative binomial}(f_k, p_k)$$

  used by Benedetti and Franconi (1998).

- A drawback: gamma hyperprior strongly concentrated on a small mean by the constraints on $[\pi_k]$ (K usually large): low variation across cells. Model I is less constrained.

A New Model

- A modification of Polettini and Stander (2004) (Model III):

A New Model

- A modification of Polettini and Stander (2004) (Model III):

$$
\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, K\alpha) \\
F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\
p_k &\sim \gamma\ \text{beta}(a\,\hat{p}_k, a\,(1-\hat{p}_k)) + (1-\gamma)\ \delta_{\{0\}}(p_k) \\
f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ \textit{independently}\ \text{across cells.}
\end{aligned}
$$

A New Model

- A modification of Polettini and Stander (2004) (Model III):

$$
\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, K\alpha) \\
F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\
p_k &\sim \gamma\ \text{beta}(a\,\hat{p}_k, a\,(1-\hat{p}_k)) + (1-\gamma)\ \delta_{\{0\}}(p_k) \\
f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ \textit{independently across cells.}
\end{aligned}
$$

- We refer to this model as Model IV.

A New Model

- A modification of Polettini and Stander (2004) (Model III):

$$
\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, K\alpha) \\
F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\
p_k &\sim \gamma\ \text{beta}(a\,\hat{p}_k, a\,(1-\hat{p}_k)) + (1-\gamma)\ \delta_{\{0\}}(p_k) \\
f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ \textit{independently} \text{ across cells.}
\end{aligned}
$$

- We refer to this model as Model IV.
- Here extra variation is introduced by also modelling p_k

A New Model

- A modification of Polettini and Stander (2004) (Model III):

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha) \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ p_k &\sim \gamma\ \text{beta}(a\,\hat{p}_k, a\,(1-\hat{p}_k)) + (1-\gamma)\ \delta_{\{0\}}(p_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model IV.

- Here extra variation is introduced by also modelling p_k
- p_k drawn from a mixture of $\begin{cases} \text{a beta}(\ a\hat{p}_k, a(1-\hat{p}_k)) \quad (*) \\ \text{a point mass at zero} \end{cases}$

A New Model

- A modification of Polettini and Stander (2004) (Model III):

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha) \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ p_k &\sim \gamma\ \text{beta}(a\,\hat{p}_k, a\,(1-\hat{p}_k)) + (1-\gamma)\ \delta_{\{0\}}(p_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ \textit{independently} \text{ across cells.} \end{aligned}$$

- We refer to this model as Model IV.

- Here extra variation is introduced by also modelling p_k
- p_k drawn from a mixture of $\begin{cases} \text{a beta}(\ a\hat{p}_k, a(1-\hat{p}_k)) \quad (*) \\ \text{a point mass at zero} \end{cases}$

  with weights γ and $1-\gamma$.

A New Model

- A modification of Polettini and Stander (2004) (Model III):

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha) \\ F_k|\pi_k &\sim \text{Poisson}(N\pi_k) \\ p_k &\sim \gamma\ \text{beta}(a\,\hat{p}_k, a\,(1-\hat{p}_k)) + (1-\gamma)\ \delta_{\{0\}}(p_k) \\ f_k|F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k),\ \textit{independently}\ \text{across cells.} \end{aligned}$$

- We refer to this model as Model IV.

- Here extra variation is introduced by also modelling p_k
- p_k drawn from a mixture of $\begin{cases} \text{a beta}(\ a\hat{p}_k, a(1-\hat{p}_k)) \quad (*) \\ \text{a point mass at zero} \end{cases}$ with weights γ and $1-\gamma$.
- The mean of each beta distribution in $(*)$ is $\hat{p}_k$. Here we make use of the sampling design weights through $\hat{p}_k$.

Some features of superpopulation Model IV:

- Risk is cell-specific.



Back
Close

Some features of superpopulation Model IV:

- Risk is cell-specific.
- Model IV takes into account the features of the sampling scheme (small regions oversampled to get estimates with the same precision across regions)

Some features of superpopulation Model IV:

- Risk is cell-specific.
- Model IV takes into account the features of the sampling scheme (small regions oversampled to get estimates with the same precision across regions)
- Calibration is based on certain population contingency tables. Using calibrated sampling weights effectively relaxes the assumption of independence so introducing association into the model.

Some features of superpopulation Model IV:

- Risk is cell-specific.
- Model IV takes into account the features of the sampling scheme (small regions oversampled to get estimates with the same precision across regions)
- Calibration is based on certain population contingency tables. Using calibrated sampling weights effectively relaxes the assumption of independence so introducing association into the model.

Estimation under Model IV

- The form of $[f_k], [F_k|f_k]$ can be evaluated analytically.

Some features of superpopulation Model IV:

- Risk is cell-specific.
- Model IV takes into account the features of the sampling scheme (small regions oversampled to get estimates with the same precision across regions)
- Calibration is based on certain population contingency tables. Using calibrated sampling weights effectively relaxes the assumption of independence so introducing association into the model.

Estimation under Model IV

- The form of $[f_k], [F_k|f_k]$ can be evaluated analytically.
- We specify α, a and γ, using available information and the loglikelihood to assess our elicitation (EB approach does not work well in Models II$\rightarrow$IV).

Some features of superpopulation Model IV:

- Risk is cell-specific.
- Model IV takes into account the features of the sampling scheme (small regions oversampled to get estimates with the same precision across regions)
- Calibration is based on certain population contingency tables. Using calibrated sampling weights effectively relaxes the assumption of independence so introducing association into the model.

Estimation under Model IV

- The form of $[f_k]$, $[F_k|f_k]$ can be evaluated analytically.
- We specify α, a and γ, using available information and the loglikelihood to assess our elicitation (EB approach does not work well in Models II$\rightarrow$IV).
- Finally, we estimate the risk using mean or mode of $[1/F_k|f_k]$.

An application

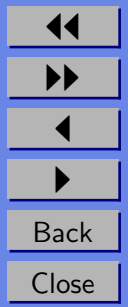
- We applied the proposed methodology to an artificial sample of data drawn from the Italian 1991 Census. We used the sampling scheme of the Labour Force Survey.

An application

- We applied the proposed methodology to an artificial sample of data drawn from the Italian 1991 Census. We used the sampling scheme of the Labour Force Survey.
- $N = 15,142,320$; $n = 53,872$.

An application

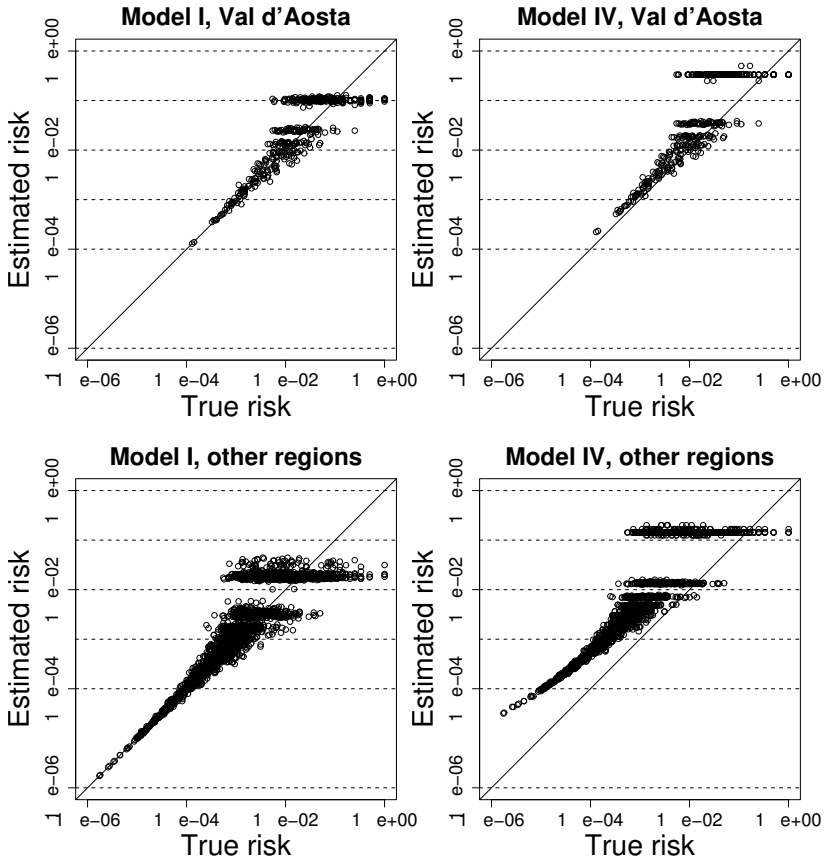
- We applied the proposed methodology to an artificial sample of data drawn from the Italian 1991 Census. We used the sampling scheme of the Labour Force Survey.
- $N = 15,142,320$; $n = 53,872$.
- Key variables:
  - sex (2 categories)
  - age (recorded in 14 classes)
  - region (4: Campania, Lazio, Val d'Aosta, Veneto)
  - position in profession (14 categories)
  - relation with the head of the household (13 categories)



An application

- We applied the proposed methodology to an artificial sample of data drawn from the Italian 1991 Census. We used the sampling scheme of the Labour Force Survey.
- $N = 15,142,320$; $n = 53,872$.
- Key variables:
  - sex (2 categories)
  - age (recorded in 14 classes)
  - region (4: Campania, Lazio, Val d'Aosta, Veneto)
  - position in profession (14 categories)
  - relation with the head of the household (13 categories)
- $K = 20384$ but the number of nonempty cells is 12526 in the population and 2966 in the sample.

Models I and IV give similar patterns:



Model IV performs better with risky cells

Another New Model

- All the above models assume *independence* across cells.

Another New Model

- All the above models assume *independence* across cells.
- We hope that further improvements can be achieved by making some use of the structure of the contingency table.

Another New Model

- All the above models assume *independence* across cells.
- We hope that further improvements can be achieved by making some use of the structure of the contingency table.
- Here is our model:

$$\begin{aligned} \underline{\pi} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ \underline{F}|\underline{\pi} &\sim \text{multinomial}(N; \pi_1, \ldots, \pi_K), \\ \underline{f}|\underline{F} &\sim \text{multinomial}(n; F_1/N, \ldots, F_K/N), \end{aligned}$$

in which $\underline{\pi} = (\pi_1, \ldots, \pi_K)$, *etc.*

Another New Model

- All the above models assume *independence* across cells.
- We hope that further improvements can be achieved by making some use of the structure of the contingency table.

▶ Here is our model:

$$\begin{aligned} \underline{\pi} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ \underline{F}|\underline{\pi} &\sim \text{multinomial}(N; \pi_1, \ldots, \pi_K), \\ \underline{f}|\underline{F} &\sim \text{multinomial}(n; F_1/N, \ldots, F_K/N), \end{aligned}$$

in which $\underline{\pi} = (\pi_1, \ldots, \pi_K)$, *etc.*

▶ We refer to this model as Model V.

Another New Model

- All the above models assume *independence* across cells.
- We hope that further improvements can be achieved by making some use of the structure of the contingency table.

▶ Here is our model:

$$\begin{aligned} \underline{\pi} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ \underline{F}|\underline{\pi} &\sim \text{multinomial}(N; \pi_1, \ldots, \pi_K), \\ \underline{f}|\underline{F} &\sim \text{multinomial}(n; F_1/N, \ldots, F_K/N), \end{aligned}$$

in which $\underline{\pi} = (\pi_1, \ldots, \pi_K)$, *etc.*

▶ We refer to this model as Model V.

- We perform inference using Markov chain Monte Carlo methods, implemented using `WinBUGS` and our own code.

Some Strategies for eliciting the $(\alpha_1, \ldots, \alpha_K)$ *parameters*

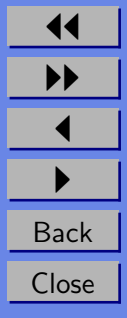
Some Strategies for eliciting the $(\alpha_1, \ldots, \alpha_K)$ parameters

- When no additional information is available, we can make use of the sampling design weights by taking $\alpha_k \propto \hat{F}_k^D$.

Some Strategies for eliciting the $(\alpha_1, \ldots, \alpha_K)$ *parameters*

- When no additional information is available, we can make use of the sampling design weights by taking $\alpha_k \propto \hat{F}_k^D$.
- If data collected at a previous census were available, we could take

$$\alpha_k \propto F_k^{\text{previous}}.$$

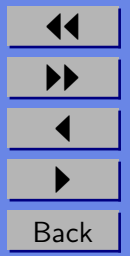


Some Strategies for eliciting the $(\alpha_1, \ldots, \alpha_K)$ parameters

- When no additional information is available, we can make use of the sampling design weights by taking $\alpha_k \propto \hat{F}_k^D$.
- If data collected at a previous census were available, we could take

$$\alpha_k \propto F_k^{\text{previous}}.$$

- If only marginal tables were available, we could specify a conditional independence model corresponding to these marginal tables to elicit the $(\alpha_1, \ldots, \alpha_K)$ parameters.

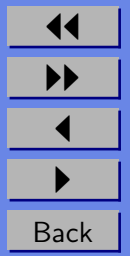


Some Strategies for eliciting the $(\alpha_1, \ldots, \alpha_K)$ parameters

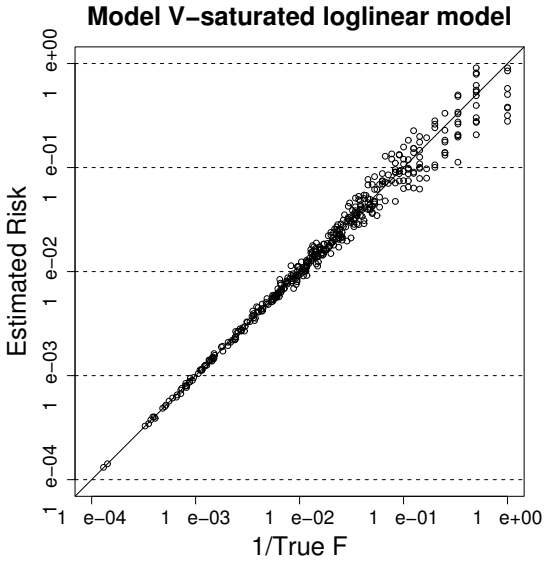
- When no additional information is available, we can make use of the sampling design weights by taking $\alpha_k \propto \hat{F}_k^D$.
- If data collected at a previous census were available, we could take

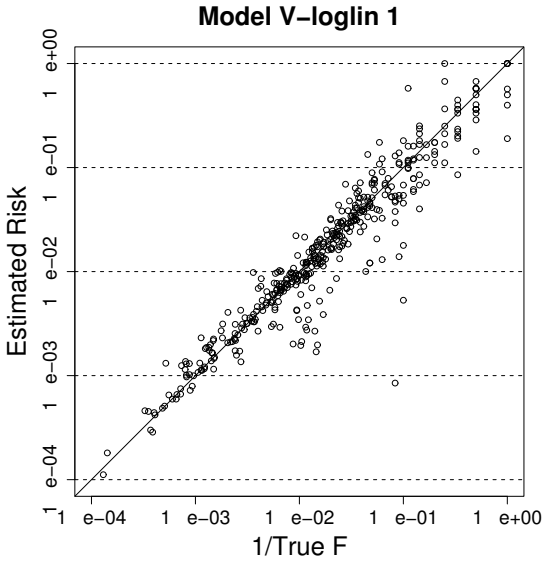
$$\alpha_k \propto F_k^{\text{previous}}.$$

- If only marginal tables were available, we could specify a conditional independence model corresponding to these marginal tables to elicit the $(\alpha_1, \ldots, \alpha_K)$ parameters.
- Loglinear models used (by region):
  - Loglin 1: `sex+(rel+age+posprof)^3`
  - Loglin 2: `rel+(sex+age+posprof)^3`



Results for Val d'Aosta region (44% sample uniques; 77% of sample frequencies in 1-5.)





	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	303	7
$\hat{r}_k > 0.05$	9	102

Sensitivity is 0.94

Specificity is 0.97

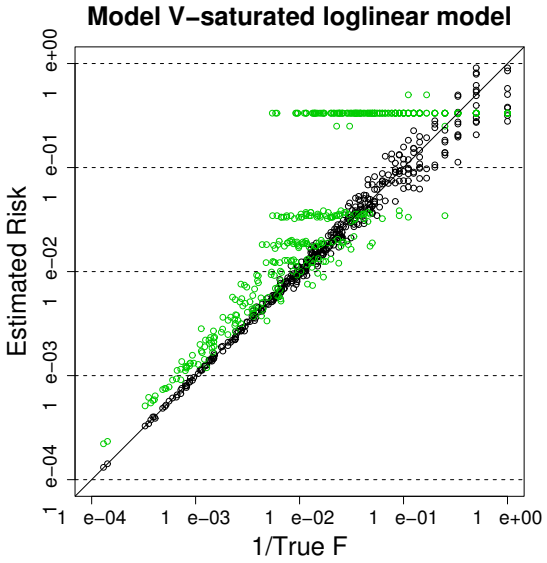
	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	298	17
$\hat{r}_k > 0.05$	14	92

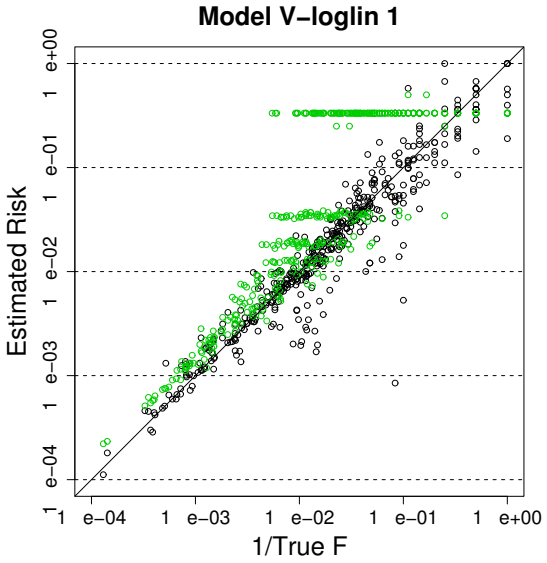
Sensitivity is 0.84

Specificity is 0.96

Back
Close

Results for Val d'Aosta region (44% sample uniques; 77% of sample frequencies in 1-5.)





	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	303	7
$\hat{r}_k > 0.05$	9	102

Sensitivity is 0.94

Specificity is 0.97

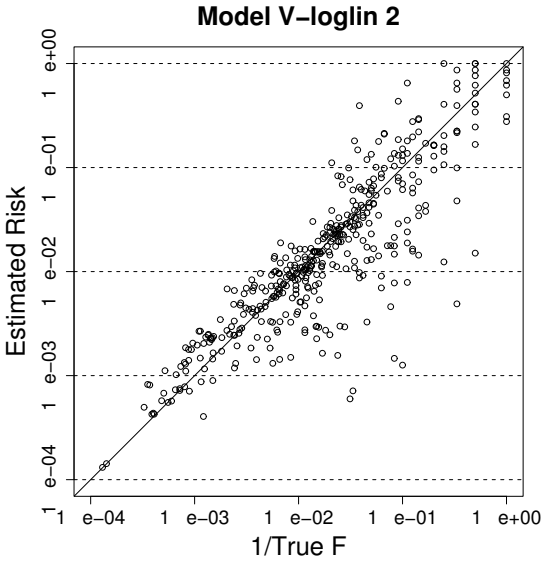
	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	298	17
$\hat{r}_k > 0.05$	14	92

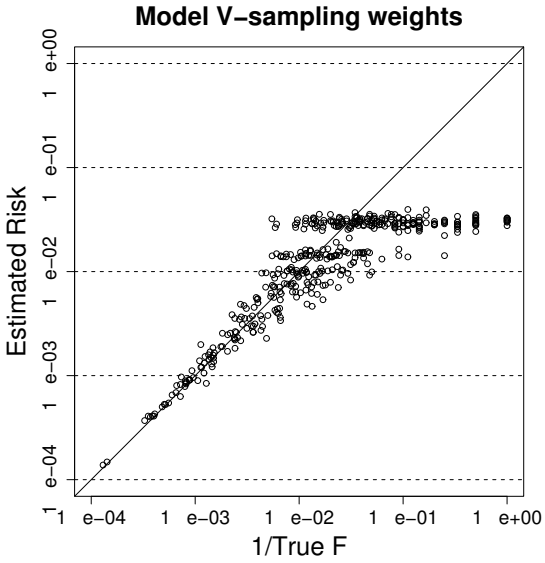
Sensitivity is 0.84

Specificity is 0.96

Back
Close

Results for Val d'Aosta region (44% sample uniques; 77% of sample frequencies in 1-5.)





	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	293	32
$\hat{r}_k > 0.05$	19	77

Sensitivity is 0.71

Specificity is 0.94

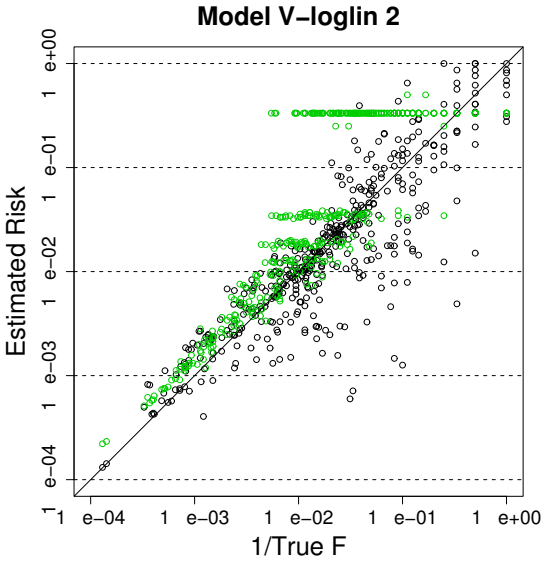
	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	312	109
$\hat{r}_k > 0.05$	0	0

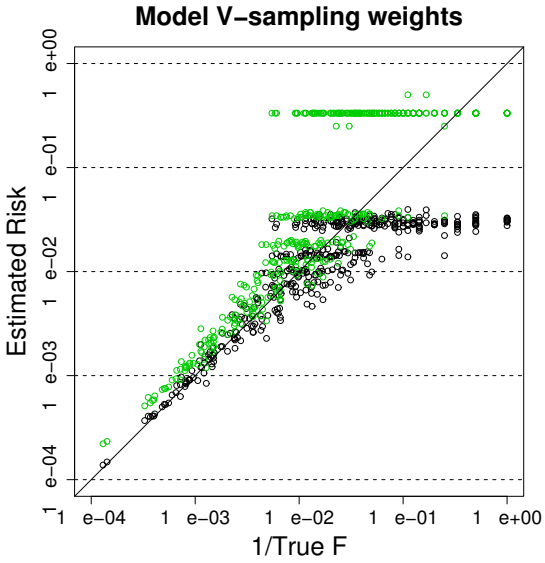
Sensitivity is 0

Specificity is 1

Back
Close

Results for Val d'Aosta region (44% sample uniques; 77% of sample frequencies in 1-5.)





	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	293	32
$\hat{r}_k > 0.05$	19	77

Sensitivity is 0.71

Specificity is 0.94

	$r_k \leq 0.05$	$r_k > 0.05$
$\hat{r}_k \leq 0.05$	312	109
$\hat{r}_k > 0.05$	0	0

Sensitivity is 0

Specificity is 1

Back
Close

Some References

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

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

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

Forster, J.J. and Webb, E.L. (2005). Bayesian Model Averaging for Disclosure Risk Assessment, *Working Paper, University of Southampton*.

Polettini, S. and Stander, J. (2004). A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In Domingo-Ferrer, J. and Torra, V. (Eds.) *Privacy in Statistical Databases*, Berlin: Springer-Verlag, 247–261.

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