On sample Disclosure Risk: definitions, models and estimation

Yosi Rinott
Dept of Statistics
Hebrew university

July 2006

Apology: this talk represent the type of work of some statisticians in this area. You might disagree with the basic tenets. Still maybe there are some statistical issues of interest.

Israel STATISTICS ORDINANCE (1972)

No information collected for the purposes of this Ordinance and derived from an individual return or the answer to a question... **shall be so published as to enable the identification of the person to whom it relates.**

.

.

.

A person who publishes or communicates to any person any information which **to his knowledge** has been disclosed in contravention of this Ordinance shall be liable to **imprisonment for a term of three years.**

http://www.cbs.gov.il/ordinanc.htm

brocard ignorantia legis non excusat

(=**ignorance** of the law **is** no **excuse**).

Example of sensitive data :
NATIONAL HEALTH SURVEY 2003
Source: The Central Bureau of Statistics and the Ministry of Health

The survey population includes all those aged 21 and over in the permanent population of Israel. 7,075 people were sampled for the survey; of whom 4,859 were interviewed. The data was collected in face-to-face interviews... The survey questionnaire includes about 1,400! items in a number of questionnaires ...

The dataset received in the ISDC includes 6 data files:
1. The core questionnaire 2. Mental and emotional disorders : depression, affective bi-polar disorder, manic episodes, dysthymia, panic disorder, agoraphobia, generalized anxiety disorder, suicidality 3. Post-traumatic stress disorder 4. Alcohol and drug addiction 5. Use of medication and use of health services 6. Mental diagnoses.

*Israel Social Sciences Data Center

Confidentiality (privacy)

Many papers discuss protection methods without a precise definition of privacy, e.g., Adam and Wortman (1989). Various types of perturbations may bias the data, and reduce its utility.

Formal definitions require additional structure such as a prior (Bayesian stucture).

It seems that a precise definition of privacy leads to the conclusion that Noise should be added to *any* data released (query).

Noise should be big enough to mask individuals (small queries) but small enough not to distort totals (large queries).

Grunau Committee at the Israel Central Bureau of Statistics: data should never be perturbed by random noise.

For large queries of size N, noise of size $< \sqrt{N}$ is insignificant, but significantly perturbs small queries.

Provides protection against a scenario of differencing (tracking), for example.

What about collusion and averaging over noise?

When the data is a (sparse) sample, natural queries may be relatively small.

Main *Scenario* of Bureaus of Statistics: no queries (but this is rapidly changing), single file (often a sample) to be released. To protect data, variables are coarsened (rounded), and Microdata becomes a Frequency Table. (Coarsening may also bias the data.)

Rules for coarsening: no cells smaller than 3 ... This is query restriction.

The 3-rule provides protection against certain scenarios, but may reduce utility of the data "too much".

Whoever wishes to keep a secret must hide the fact that he possesses one –Johann Wolfgang von Goethe

This talk: A sample frequency table is to be released by agency. Population table unknown or partially known to the agency.

Disclosure Risk arises from small population cells which are represented in the sample and in particular population uniques which are also in the sample (hence sample uniques).

Agency wants to assess risk under relevant scenarios and modify table if risk is high.

Sample (size n): $\mathbf{f} = \{f_k : k = 1, \ldots, K\}$
Population (size N): $\mathbf{F} = \{F_k : k = 1, \ldots, K\}$,
tables with K cells, $k = (k_1, \ldots, k_m)$,
m-way table.

Agency intends to publish the sample.

"Intruder" (adversary, snooper): tries to match individuals in sample with population on the basis of variables with which he is familiar (Key Variables) and then infer on other variables in the table.

Agency: on the basis of the sample and (usually) partial knowledge on the population, agency **estimates** Disclosure Risk, that is, some measure of the intruder's chance of success.

Statistical models for contingency tables:

$\mathbf{F} = \{F_k : k = 1, \ldots, K\}$, a parameter.
$\mathbf{f} = \{f_k : k = 1, \ldots, K\}$ data.
$n/N = \pi =$ sampling fraction. $\rho_k = F_k/N$
$\{f_k\}|\{F_k\} \sim \text{Multinomial}_K(n, \{\rho_k\})$, or $f_k \sim \text{Poisson}(n\rho_k)$.

Models: attributes are independent ($\rho_k = \prod_{i=1}^m p_{k_i}^{[i]}$), or conditionally independent ...

$\mathsf{E} f_k = n\rho_k = n \exp(\mathbf{x}'_k \boldsymbol{\theta})$; for example
$= n \exp(\sum_{i=1}^m \theta_{k_i}^{[i]}) = n \prod_{i=1}^m p_{k_i}^{[i]}$
means independent attributes.
($\boldsymbol{\theta} = \{\theta_j^{[i]}\}$, $1 \le i, j \le m$, $x_{k,i} = \delta_{i,k_i}$).

Two-way interactions: $\mathsf{E} f_k = \exp(\sum_{i,j=1}^m \theta_{k_i,k_j}^{[i,j]})$.

Statistics: estimate parameters (MLE), distribution of estimates (confidence intervals) for large n, test hypotheses (e.g., that attributes are independent).

Is this relevant to SDC?

Risk Measures: two simple examples

$$\tau_1 = \sum I(f_k = 1, F_k = 1)$$

$$\tau_2 = \sum I(f_k = 1) 1/F_k$$

=the expected number of correct matches of sample uniques. Clearly $\tau_1 < \tau_2$.

May want normalize by size of file, or its information value?

In the statistics literature, $\not\exists$ formal definition of safe file.

Other Risk measures:

$\theta_1 = \sum_k I(f_k = 1, F_k = 1) / \sum_k I(f_k = 1)$
$= P(pu \,|\, su)$ = probability that a randomly chosen sample unique is a population unique.

$\theta_2 = \sum_{k=1}^{K} I(f_k = 1) F_k^{-1} / \sum_{k=1}^{K} I(f_k = 1)$
=average probability of a correct match.

Skinner and Elliot (2002) :

$$\theta_{SE} = \sum_{k=1}^{K} I(f_k = 1) / \sum_{k=1}^{K} F_k I(f_k = 1)\,.$$

Probability of a correct match if intruder chooses at random an individual from all population cells which are sample uniques and matches him to the sample unique.

The "parameters" τ_i and θ_i are of the form $\sum_k U(f_k, F_k)$, or changing notation $\sum_i U(X_i, \theta_i)$, Robbins and Zhang (2000), Zhang (2005).

Estimation of risk measures are based on the conditional distribution of $\mathbf{F}|\mathbf{f}$.

Estimates:

$$\hat{\tau}_1 = \sum I(f_k = 1)\hat{P}(F_k = 1|f_k = 1)$$

$$\hat{\tau}_2 = \sum I(f_k = 1)\hat{E}[1/F_k|f_k = 1]$$

where $\hat{P}$ and $\hat{E}$ are estimates, efficient under certain conditions, Zhang (2005).

Statistical models for disclosure risk

Common (natural?) Assumptions

$F_k|\gamma_k \sim Poisson(N\gamma_k)$ ind. $\sum \gamma_k = 1$

$f_k|F_k \sim Bin(F_k, \pi_k)$, Bernoulli or Poisson sampling.

$\Downarrow$

$f_k \sim Poisson(N\gamma_k\pi_k)$. observed

$\Downarrow$

$F_k \,|\, f_k \;\sim\; f_k + Poisson(\lambda_k = N\gamma_k(1-\pi_k))$ (*)

In particular

$F_k \,|\, f_k = 1 \;\sim\; 1 + Poisson(\lambda_k)$

ADD assumption

$\gamma_k \sim Gamma(\alpha, \beta)$ ind

$\Downarrow$

$f_k \sim NB(\alpha, p_k = \frac{1}{1+N\pi_k\beta})$ observed (♣)

$F_k \,|\, f_k \sim f_k + NB(\alpha + f_k, \frac{N\pi_k + 1/\beta}{N+1/\beta})$ (**)

As $\alpha \to 0$ and $\beta \to \infty$ we obtain the μ-ARGUS assumption $F_k \,|\, f_k \sim f_k + NB(f_k, \pi_k)$.
As $\alpha \to \infty$ and $\alpha\beta^2 \to 0$ we obtain the above Poisson (*).

Estimation:

(**) **ARGUS** (Benedetti, Capobianchi and Franconi 1998): $F_k \,|\, f_k \sim f_k + NB(f_k, \pi_k)$.

Let $w_i =$ "sampling weight" of individual i, obtained from design or post-stratification.

$\hat{\pi}_k = f_k / \hat{F}_k$, $\hat{F}_k = \sum_{i \in \text{sample cell } k} w_i$.

$f_k = 0 \Rightarrow \hat{F}_k = 0$, $\quad \sum_k \hat{F}_k = \sum_i w_i = N$

$\Rightarrow \hat{\pi}_k \neq 0$ are overestimated $\Rightarrow$ Risk is underestimated.

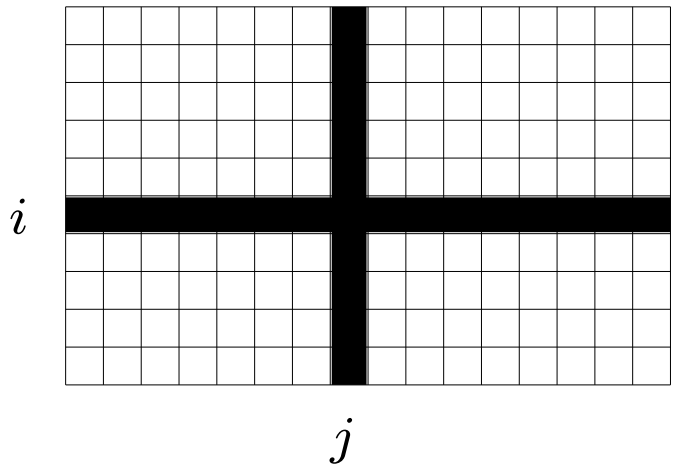
Monotonicity: if we replace $f_k = 0$ by some ε, Risk estimates increase to the correct level in ε, but how do we estimate ε?

(*) **Poisson Log-linear Models** Skinner and Holmes (1998), Elamir and Skinner (2005), Skinner and Shlomo (2005):
$\mathrm{E} f_k = \exp\{(x_k' \beta\}$.

"Monotonicity" in the size of the model. Saturated ("big" model) $\Rightarrow$ Risk underestimation, Independence ("small" model) $\Rightarrow$ Risk overestimation.

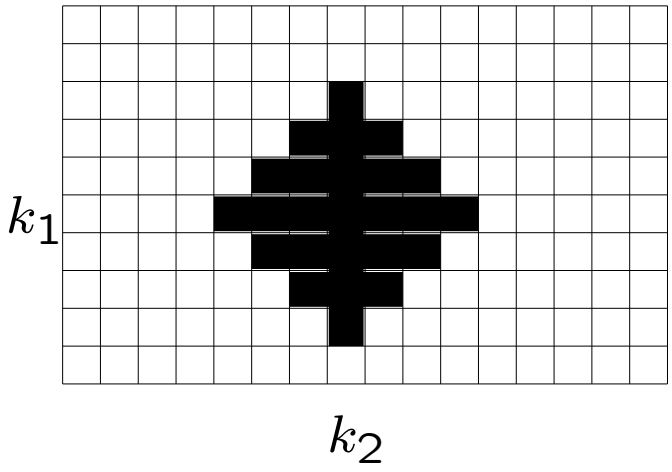
Log linear models use a **neighborhood** of cells of cell k to infer on cell k (γ_k).

Independence Neighborhoods, $k = (i, j)$:



Local smoothers for large sparse (ordinal) tables, e.g., Bishop-Fienberg-Holland (1975), Simonoff(1998±): use local neighborhoods to fit a simple smooth function to f_k or to estimate γ_k by some model after smoothing, [and then (Abberger, 2002) test for Independence, etc.]

Proposed Neighborhoods (work with Natalie Shlomo):



Fix cell k and let $k' \in M$, neighborhood of cell k.

$f_{k'} \sim NB(\alpha_k, p_k = \frac{1}{1+N\pi_k\beta_k})$ (♣)

Likelihood of data in M of cell k: $L = \prod_{k' \in M} P(f_{k'})$.

Assume

$\mathsf{E} f_{k'} = \theta_0 + \theta_1(k'_1 - k_1) + \vartheta_1(k'_2 - k_2) + \ldots + \theta_t(k'_1 - k_1)^t + \vartheta_t(k'_2 - k_2)^t.$

Compute MLE, and estimate $\mathsf{E} f_k$ by $\exp(\widehat{\theta}_0)$.

Example 1 Population : extract from the 1995 Israeli Census. $N = 37,586$, $n = 3,759$, $K = 11,648$. Attributes (with number of levels in parentheses): Sex(2) * Age Groups (32) * Income Groups(14) * Years of Study (13). Sex fixed in neighborhoods.

$$M = \{k' : k'_1 = k_1, \max_{i \geq 2} |k'_i - k_i| \leq c\},$$

with $c = 2$, and since we vary three variables, each over a range of five values, we have $|M| = 125$.

Model	τ_1	τ_2
True Values	187	452.0
Argus	137.2	346.4
Log Linear Model: Independence	217.3	518.0
Log Linear Model: 2-Way Interactions	167.2	432.8
NB Smoothing $t = 2$ $\|M\| = 125$	181.9	461.3

Final comments:

The argument on Risk definitions and measures will go on.

In the sample-population setup, Risk measure estimation is a non standard statistical question, requiring suitable methods.