UNIVERSITY OF Southampton

Southampton Statistical Sciences Research Institute

Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility

Natalie Shlomo
Southampton Statistical Sciences Research Institute
University of Southampton
N.Shlomo@soton.ac.uk

UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Topics for Discussion

- Introduction
- Disclosure Risk Assessment for Sample Microdata
- Some Disclosure Limitation Methods
- Information Loss Measures
- Example
- Discussion

Introduction

• Statistical Agencies release sample microdata from social
surveys under different modes of access:
Safe on-site datalabs
Microdata under contract (MUC)
Public Use File (PUF)

• Future dissemination strategies based on flexible table building software and remote access pose new challenges

• Microdata Review Panels (MRPs) need to make informed
decisions when releasing sample microdata:
- objective disclosure risk assessment
- tolerable risk thresholds
- modes of access

Disclosure Risk Assessment

- Choosing an optimal SDL method is an iterative process:

Manage Disclosure Risk ⇄ Preserve Utility

- In social surveys, we assume that population is unknown (or partially known through margins)

- Disclosure risk scenario:
  - linking microdata to external databases
  - spontaneous recognition, self-identification

- Identifying key variables usually discrete: place of residence, sex, occupation, marital status, ethnicity, age, etc.

Disclosure Risk Assessment

• Disclosure risk assessed on contingency table of sample counts
spanned by identifying key variables and is a function of both
sample and population counts

• Other methods for assessing disclosure risk in microdata

- heuristic based on special uniques (combinations of identifying
key variables that remain unique after aggregation and are
likely to be population uniques)
- probabilistic record linkage

Problems:

Disclosure Risk Assessment

Probabilistic Modelling

- F_k population count and f_k sample count in cell k
- Disclosure risk measures: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$ $\quad \tau_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$
- For unknown population counts, estimate from the conditional distribution of $F_k \mid f_k$

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(F_k = 1 \mid f_k = 1) \qquad \hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}\left(\frac{1}{F_k} \mid f_k = 1\right)$$

- Natural assumption: $F_k \sim Poisson(\lambda_k)$

  Bernoulli sampling: $f_k \mid F_k \sim Bin(F_k, \pi_k)$

It follows that: $f_k \sim Pois(\pi_k \lambda_k)$ and $F_k \mid f_k \sim Poisson(\lambda_k(1-\pi_k))$

where $F_k \mid f_k$ are conditionally independent

Disclosure Risk Assessment

- Skinner and Holmes, 1998, Elamir and Skinner, 2006 use log linear models to estimate parameters $\{\lambda_k\}$

- Sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k \lambda_k$

- Log-linear model for estimating $\{\mu_k\}$ expressed as:

$$\log(\mu_k) = \mathbf{x}_k' \beta$$

where $\mathbf{x}$ design matrix of key variables and their interactions

- MLE's calculated by solving score function:

$$\sum_k [f_k - \exp(\mathrm{x}_k' \beta)]\mathrm{x}_k = 0$$

Disclosure Risk Assessment

- Fitted values calculated by $\hat{u}_k = \exp(\mathbf{x}'_k \hat{\beta})$ and $\hat{\lambda}_k = \dfrac{\hat{u}_k}{\pi_k}$

- Individual risk measures estimated by:

$$\hat{P}(F_k = 1 \mid f_k = 1) = \exp(-\hat{\lambda}_k (1 - \pi_k))$$

$$\hat{E}(\frac{1}{F_k} \mid f_k = 1) = [1 - \exp(-\hat{\lambda}_k (1 - \pi_k))] / [\hat{\lambda}_k (1 - \pi_k)]$$

- Rinott and Shlomo, 2007 develop confidence intervals for
  global risk measures

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

  Example: sum of Bernoulli random variates
  taking a value of 1 with probability: $P(F_k = 1 \mid f_k = 1)$

$$Var(\tau_1 \mid \mathrm{f}) = \sum_k I(f_k = 1) P(F_k = 1 \mid f_k = 1)(1 - P(F_k = 1 \mid f_k = 1))$$

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

Disclosure Risk Assessment

- Skinner and Shlomo, 2008 develop goodness of fit criteria which minimize the bias of risk estimates

- Define: $h(\lambda_k) = P(F_k = 1 \mid f_k = 1)$ for τ_1 and $h(\lambda_k) = E(\frac{1}{F_k} \mid f_k = 1)$ for τ_2

- Consider expression: $B = \sum_k E[I(f_k = 1)][h(\hat{\lambda}_k) - h(\lambda_k)]$

- A Taylor expansion leads to an approximation

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

and the relations: $Ef_k = \pi_k \lambda_k$ and $E[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k] = \pi_k{}^2 E(\lambda_k - \hat{\lambda}_k)^2$ under the null hypothesis of a Poisson fit:

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

For τ_1 :

$$\hat{B}_1 = \sum_k \hat{\lambda}_k \exp(-\hat{\lambda}_k)(1-\pi)\{(f_k - \hat{\mu}_k) + (1-\pi)[(f_k - \hat{\mu}_k)^2 - f_k]/(2\pi)\}$$

Disclosure Risk Assessment

• Method selects the model using a forward search algorithm which minimizes $\hat{B}_i / \sqrt{\hat{v}_i}$ for $\hat{\tau}_i, \quad i = 1,2$ where $\hat{v}_i$ variance of $\hat{B}_i$

Example: Population of 944,793 and sample size 9,448

Key: Area (2), Sex (2), Age (101), Marital Status (6), Ethnicity (17), Economic Activity (10) - 412,080 cells

Model Selection:

Starting solution: simple log-linear model which indicates under-fitting, i.e. minimum error statistics too large and add in higher interaction terms until minimum error statistics indicate fit

Model Search - Simple random sample of size 9,448

True values , $\tilde{\tau}_2 = 159$ $\tilde{\tau}_3 = 355.9$

Area–ar, Sex-s, Age–a, Marital Status–m, Ethnicity–et, and Economic Activity-ec

	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{B}_2 / \sqrt{v_2}$	$\hat{B}_3 / \sqrt{v_3}$
Independence - I	386.6	701.2	48.54	114.19
All 2 way - II	104.9	280.1	-1.57	-2.65
1: I + {a*ec}	243.4	494.3	54.75	59.22
2: 1 + {a*et}	180.1	411.6	3.07	9.82
3: 2 + {a*m}	152.3	343.3	0.88	1.73
4: 3 + {s*ec}	149.2	337.5	0.26	0.92
5a: 4 + {ar*a}	148.5	337.1	-0.01	0.84
5b: 4 + {s*m}	147.7	335.3	0.02	0.66
6b: 5b + {ar*a}	147.0	335.0	-0.24	0.56
6c: 5b + {ar*m}	148.9	337.1	-0.04	0.72
6d: 5b + {m*ec}	146.3	331.4	-0.24	0.03
7c: 6c + {m*ec}	147.5	333.2	-0.34	0.06
7d: 6d + {ar*a}	145.6	331.0	-0.44	-0.03

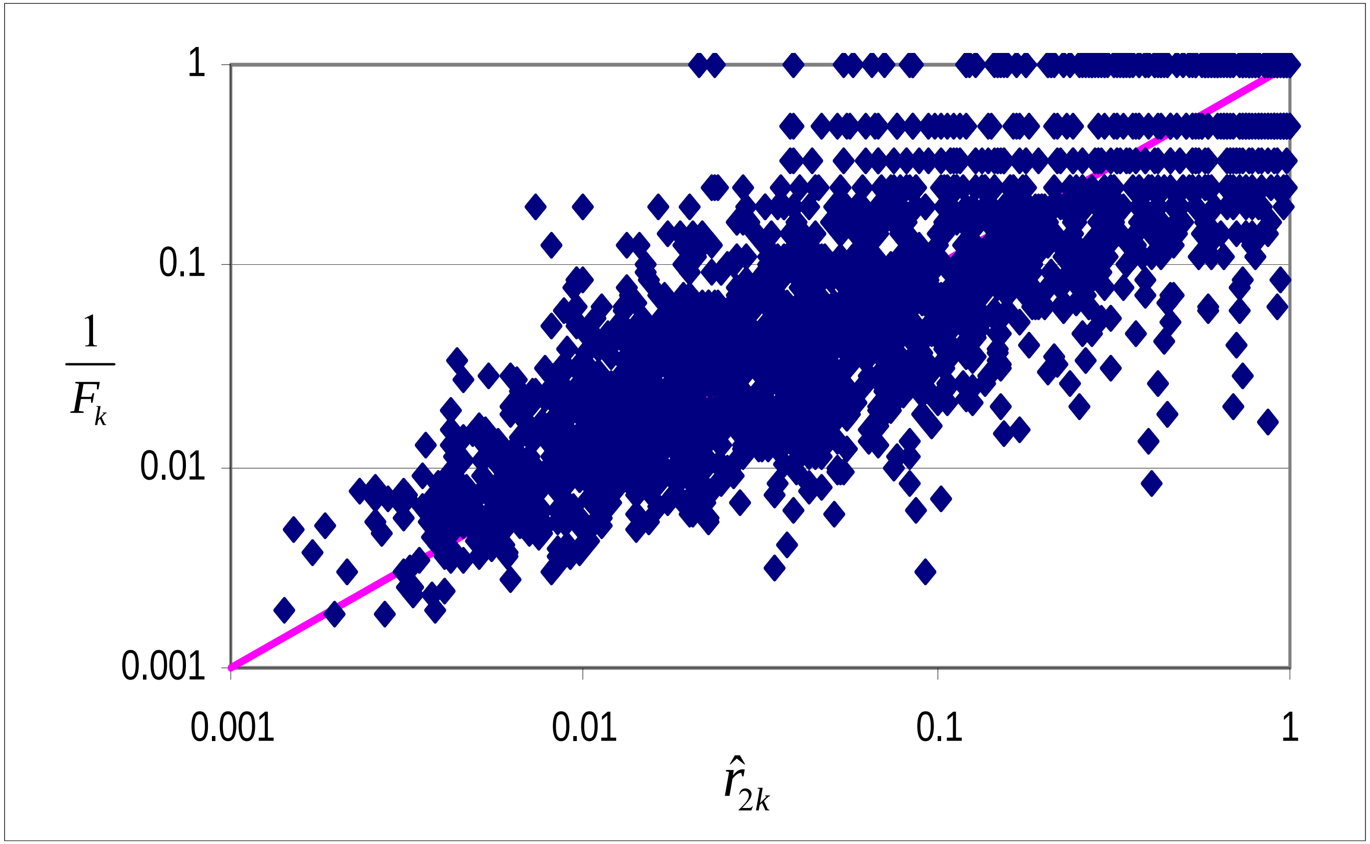
Example

Preferred Model: {a*ec}{a*et}{a*m}(s*ec}{ar*a}

True Global Risk: $\tilde{\tau}_2 = 159$ $\tilde{\tau}_3 = 355.9$

Estimated Global Risk $\hat{\tau}_2 = 148.5$ $\hat{\tau}_3 = 337.1$

Log-scale



Example

Preferred Model: {a*ec}{a*et}{a*m}(s*ec}{ar*a}

True Global Risk: $\tilde{\tau}_2 = 159$ $\tilde{\tau}_3 = 355.9$

Estimated Global Risk $\hat{\tau}_2 = 148.5$ $\hat{\tau}_3 = 337.1$

True Record Level Risk Measures	**Estimated Record Level Risk Measures**			
	0 – 0.1	**0.1 – 0.5**	**0.5 – 1**	**Total**
0 – 0.1	1,391	150	11	1,552
0.1 – 0.5	162	253	76	491
0.5 – 1	26	91	144	261
Total	1,579	494	231	2,304

Disclosure Risk Assessment

- Skinner and Shlomo, 2008 address complex survey designs:

  - Sampling clusters introduces dependencies
    key variables (such as: age, sex, occupation) cut across clusters
    and assumption holds in practice in most household surveys

  - Stratification
    strata id included in key to account for differential inclusion probabilities

$$\sum_k [\hat{F}_k - \exp(x_k'\xi)]x_k = 0$$

  - Incorporate survey weights in risk measure and goodness of fit π_k $\hat{\pi}_k = f_k / \hat{F}_k$ $\hat{F}_k = \sum_{i \in k} w_i$
    criteria using pseudo maximum likelihood estimation
    score function modified to:

changed to: where

Disclosure Risk Assessment

• Model assumes no misclassification errors (including perturbation
  from SDL methods)

•Skinner and Shlomo, 2007 address misclassification errors:
  Let: $M_{kj} = P(\tilde{X} = k \mid X = j)$
  where X cross-classified variables:
  X in population fixed
  $\tilde{X}$ in microdata subject to misclassification (perturbation)

• The per-record disclosure risk measure of a match with a sample
  unique under measurement error:

$$\frac{M_{kk}/(1-\pi M_{kk})}{\sum_j F_j M_{kj}/(1-\pi M_{kj})} \le \frac{1}{F_k}$$

• For small misclassification and small sampling fractions:

$$\frac{M_{kk}}{\sum_j F_j M_{kj}} \qquad \frac{M_{kk}}{\tilde{F}_k}$$

$$\tau_2 = \sum_k I(f_k = 1)\frac{M_{kk}}{F_k}$$

or

$$\hat{\tau}_2 = \sum_k I(\tilde{f}_k = 1) M_{kk} \hat{E}\left(\frac{1}{\tilde{F}_k} \mid \tilde{f}_k\right)$$

SDL for Sample Microdata

- Depending on disclosure risk assessment SDL methods may need to be applied

- Non-perturbative methods limit information released:
  recoding, subsampling, tabulations

- Perturbative methods alter the data:
  rounding, adding noise, misclassification

To minimise information loss:
preserve sufficient statistics and logical consistencies in the data

- Combine and optimize SDL methods

SDL for Sample Microdata

•

• Additive noise on a continuous variable:

- generate noise within small sub-groups such as within percentiles
- correlated noise: define parameter δ, calculate: $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$
- generate noise ε independently for each record with a mean $\mu' = \frac{1-d_1}{d_2}\mu$ and the original variance σ^2

$$z'_i = d_1 \times z_i + d_2 \times \varepsilon_i$$

$$E(z') = d_1 E(z) + d_2 [\frac{1-d_1}{d_2} E(z)] = E(z)$$

and

$$Var(z') = (1-\delta^2) Var(z) + \delta^2 Var(z) = Var(z)$$

For $\delta = 1$ we obtain 'synthetic' data

SDL for Sample Microdata

- Additive noise on continuous variables (multivariate):
  - consider x, y and z where $x + y = z$
  - generate noise within percentiles: $(\varepsilon_x, \varepsilon_y, \varepsilon_z)^T \sim N(\mu', \Sigma)$

    $$\mu'^{\mathrm{T}} = (\mu'_x, \mu'_y, \mu'_z) = (\frac{1-d_1}{d_2}\mu_x, \frac{1-d_1}{d_2}\mu_y, \frac{1-d_1}{d_2}\mu_z)$$

    Σ original covariance matrix

    (generated noise preserves additivity)

  - for each separate variable, eg.: $z'_i = d_1 \times z_i + d_2 \times \varepsilon_{zi}$

    same mean vector and covariance matrix and additivity exactly perturbed

SDL for Sample Microdata

- Microaggregation – records in groups of size k
  each individual in group k has value replaced by group average

- Reduces ‘between’ variance
  - generate additive noise and add to microaggregated averages
  - for multivariate setting and preserving additivity apply linear programming techniques

- Unbiased random rounding
  Let $Floor(x)$ be the largest multiple k of the base b for an entry x
  Define $res(x) = x - Floor(x)$
  x is rounded up to $(Floor(x) + b)$ with probability $\frac{res(x)}{b}$

$(1-\frac{res(x)}{b})$

$$(x - Floor(x)) \times (1 - \frac{res(x)}{b}) + (x - (Floor(x) + b)) \times \frac{res(x)}{b} = 0$$

SDL for Sample Microdata

Selection Strategy:

- With replacement

  Each cell rounded independently in the table, i.e. a random uniform number *u* between 0 and 1 is generated for each cell

  If $u < \frac{res(x)}{b}$ the entry is rounded up, otherwise rounded down

- Without replacement

  Expected number of values to round up calculated based on probabilities, values selected (without replacement) to round up and the remainder rounded down

  Method semi- controls totals (overall and/or rows (or columns)) while maintaining unbiased tables

SDL for Sample Microdata

- PRAM (Post-randomisation method)

- $L \times L$ transition probability matrix P containing conditional probabilities p_{ij} for a categorical variable with L categories:

$$p_{ij} = p(\text{perturbed category is } j \mid \text{original category is } i)$$

- T vector of frequencies

- On each record, category of variable changed or not changed according to P and the result of a draw of a random variate u

- T^* vector of perturbed frequencies

- Unbiased moment estimator of the original data: $\hat{T} = T^* P^{-1}$
    assuming P has an inverse (dominant on the diagonals)

SDL for Sample Microdata

PRAM (Post-randomisation method) - cont.

- Invariant PRAM - Define *P* such that $TP = T$
  (vector of the original frequencies eigenvector of *P*)

- Perturbed data unbiased estimate of the original file

- Expected values of marginal distribution preserved

- Exact marginal distribution preserved using a without replacement selection strategy

- Carry out perturbation within sub-groups (block diagonal transition probability matrix) and compound correlated variables

- Post-editing to correct further inconsistencies

Information Loss Measures

- Utility measured by whether inference can be carried out on perturbed data similar to original data
- Use proxy information loss measures on distributions calculated from microdata:

- Distance Metrics:

$$AAD(D_{orig}, D_{pert}) = \frac{\sum_c | D_{pert}(c) - D_{orig}(c) |}{n_c}$$

where n_c number of cells in

distribution

$$D_{orig}(t) = \sum_c I(c \le t) / n_c$$

Let: empirical distribution

$$KS(D_{orig}, D_{pert}) = \max_j (| D_{pert}(t_j) - D_{orig}(t_j) |)$$

where $\{t_j\}$ values are n_c jointly ordered original and perturbed values

Also, relative difference in means or variances

Information Loss Measures

- Relative difference in Cramer's V for 2-way table: $CV = \sqrt{\dfrac{\chi^2 / n}{\min(R-1),(C-1)}}$

$$RCV(D_{pert}, D_{orig}) = 100 \times \frac{CV(D_{pert}) - CV(D_{orig})}{CV(D_{orig})}$$

- Relative difference in 'Between' Variance:

a target proportion for a cell *c* in row *k*, $P^k_{orig}(c) = \dfrac{D^k_{orig}(c)}{\sum_c D^k_{orig}(c)}$

an overall proportion $P^k_{orig} = \dfrac{\sum_c D^k_{orig}(c)}{\sum_c D_{orig}(c)}$

Between variance: $BV(P^k_{orig}) = \dfrac{1}{n_c - 1} \sum_c (P^k_{orig}(c) - P^k_{orig})^2$

$$BVR(P^k_{pert}, P^k_{orig}) = 100 \times \frac{BV(P^k_{pert}) - BV(P^k_{orig})}{BV(P^k_{orig})}$$

Example

- 1995 Israel Census Sample: N=753,711 with a 1:100 sample, n=7,537
- Key: K=476,850

Locality Code (single codes large localities above 10,000 and single combined code for small localities) (85)
Sex (2)
Age groups (15)
Occupation (11)
Income groups (17)

Compare the following:
A. PRAM versus recoding geographic variable
B. Correlated noise, microaggregation and additive noise, controlled
random rounding to base 10 on income variable

Results

	Original Key	Recoded localities (30 categories)	PRAM (70% on diagonal)
	1025.7	571.5	714.7
Disclosure Risk			
$\hat{\tau}_2$ (test statistic) Sample uniques $\hat{\tau}_2 / n$	1015.5 (1.94) 4005 25.3%	599.9 (1.32) 3376 17.8%	729.5 (1.42) 3479 20.9%
Utility			
AAD across 85 localities	-	7.22	3.88
KS across 85 localities	-	1.53	0.46
RCV for localities *occupation (true=0.1370)	-	-0.33	-0.08
BVR for average income between localities	-	-0.44	-0.09

Results

	Random Noise		Rounding to Base 10		Micro-aggregation	
	Uncorre-lated	Corre-lated	Random	Semi Controlled	Without noise	With noise
AAD across 17 income groups	26.9	22.4	2.4	2.0	4.7	20.3
KS across 17 income groups	0.98	0.90	0.71	0.66	0.11	0.87
Percent relative difference in variance	3.54	0.16	0.00	0.00	-1.47	-0.18
RCV for income groups (17) & occupation (11) (true=0.1736)	0.63	1.15	0.00	-0.11	0.98	-0.29
BVR average income between localities (85) (true=3.08 x 10^9)	2.47	1.21	0.02	-0.01	-0.91	1.11
Percentage of records switching income	10.8%	6.6%	0.4%	0.3%	0.8%	5.3%

Discussion

- Some conclusions from example:

  - can objectively assess disclosure risk through probabilistic models

  - recoding causes significant information loss compared to PRAM
    but is more effective at reducing disclosure risks

  - good practice to combine methods, i.e. recoding and then
    applying perturbative method to remaining high risk cells

  - both recoding and PRAM attenuate the data

  - adding noise has significantly more impact on distortions to
    distributions than random rounding where the “noise” is fixed

Discussion

- Statistical Agencies (MRP) need to:
  - assess disclosure risk objectively
  - set tolerable risk thresholds according to different access modes
  - optimize and combine SDL techniques
  - provide guidelines on how to analyze disclosure controlled datasets

- Future dissemination strategies presents new challenges:
  - synthetic data might be produced for web access before obtaining access to real data
  - need to develop online SDL techniques for flexible table generating software and remote access
  - need methods for auditing query systems

Bridge the Statistical and Computer Science literature on