UNIWERSYTET EKONOMICZNY W POZNANIU

Kadry dla Gospodarki
Twój sukces się liczy

URZĄD STATYSTYCZNY W POZNANIU

Population size estimation based on multiple lists

Uncertainty analysis for categorical data fusion

Li-Chun Zhang

University of Southampton (L.Zhang@soton.ac.uk)

Statistics Norway (lcz@ssb.no)

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

I. Population size estimation based on multiple lists

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

- Census & pop. coverage survey (PCS)
- Capture-recapture (CR) or dual-system
  - Population size estimate = $(s / n) \cdot x$
  - (x, s, n) = (census, PCS, matched)_count
- Key assumptions
  - Need only one homogeneous capture rate
  - Independence
  - No erroneous (or over-) count!
  - …

- Why allow over-count?
  - *Well, maybe you sit on a helicopter; its' raining; it's windy; it's getting late; you are tired…*
- Settings for applications
  - Ecology
  - Population size estimation beyond census
  - Hard-to-get and dynamic population
  - Multiple screening / tests / edits …
  - …

- CR estimate with over-count in *list*

  $$(s / n) \cdot (x\ \rho)$$

  - Maintain only under-count in “survey”
  - ρ (≤1) adjustment for over-counting in “list”
- A couple of more rates
  – Catch_rate = P(in list | in pop)
  – Error_rate = P(not in pop | in list)
  = 1 – P(in pop | in list)
  = 1 – Hit_rate = 1 - ρ

List - survey data

- two lists: (x_1, x_2, x_{12}) and $x_{1(2)} = x_1 \setminus x_{12}$ and $x_{2(1)} = x_2 \setminus x_{12}$
- with corresponding error rates: $(\theta_1, \theta_2, \theta_{12}, \theta_{1(2)}, \theta_{2(1)},)$
- additional survey: $(n_{12}, n_{1(2)}, n_{2(1)}, n_{(12)})$ and catch rate γ

Assume independence between survey and lists (NB. not btw lists):

$$E(n_{1(2)}) = x_{1(2)}(1-\theta_{1(2)})\gamma$$

$$E(n_{2(1)}) = x_{2(1)}(1-\theta_{2(1)})\gamma$$

$$E(n_{12}) = x_{12}(1-\theta_{12})\gamma$$

$$E(n_{(12)}) = (N - x_{1(2)}(1-\theta_{1(2)}) - x_{2(1)}(1-\theta_{2(1)}) - x_{12}(1-\theta_{12}))\gamma$$

Need one assumption between $(\theta_{12}, \theta_{1(2)}, \theta_{2(1)})$

- Consider the alternative assumptions:
  - P(in pop | in L1 and L2) = P(in pop | in L1 not L2) · P(in pop | in L2 not in L1)
  - P(in pop | in L1 and L2) = P(in pop | in L1) · P(in pop | in L2)
  - P(not in pop | in L1 and L2) = P(not in pop | in L1 not L2) · P(not in pop | in L2 not in L1)

    (NB. Similar to a 3-way log-linear model with structural 0's)
  - P(not in pop | L1 and L2) = P(not in pop | in L1) · P(not in pop | in L2)

    (NB. Can not be expressed as log-linear models at all…)

- Conditional independence states e.g.

$$P(I_{i1} = I_{i2} = 1 | I_{iU} = 1) = P(I_{i2} = 1 | I_{iU} = 1) P(I_{i1} = 1 | I_{iU} = 1)$$

i.e. *conditional joint* probability is the product of *conditional marginal* probabilities

- *Pseudo conditional independence (PCI)* defined as e.g.

$$P(I_{iR} = 1 | I_{i1} = I_{i2} = 1) = P(I_{iR} = 1 | I_{i1} = 1) P(I_{iR} = 1 | I_{i2} = 1)$$

i.e. *joint conditional* probability is the product of *marginal conditional* probabilities

Ω_K : set of all non-empty subsets of $\{1, 2, ..., K\}$, e.g.

$$\Omega_2 = \{\{1,2\},\{1\},\{2\}\} \qquad \Omega_3 = \{\{1,2,3\},\{1,2\},\{1,3\},\{2,3\},\{1\},\{2\},\{3\}\}$$

For $\omega \in \Omega_K$ define CR table counts

$$x_\omega = \sum_i \left(\prod_{k\in\omega} I_{ik}\right) \quad \text{and} \quad x_{\omega(\omega^c)} = \sum_i \left(\prod_{k\in\omega} I_{ik}\right)\left(\prod_{k\in\omega^c}(1-I_{ik})\right)$$

Put saturated log-linear model

$$E(r_\omega \mid x_\omega) = x_\omega \theta_\omega \quad \text{and} \quad \log\theta_\omega = \sum_{a\in\Omega(\omega)} \alpha_a$$

where $\Omega(\omega)$ is the set of non-empty subsets of ω. For K = 2:

$$\log\theta_1 = \alpha_1 \quad \text{and} \quad \log\theta_2 = \alpha_2 \quad \text{and} \quad \log\theta_{12} = \alpha_1 + \alpha_2 + \alpha_{12}$$

The PCI-model, i.e. $\theta_{12} = \theta_1\theta_2$ is given by $\alpha_{12} = 0$. Moreover, define PC-probability

$$\theta_{\omega|a} \overset{\text{def}}{=} \frac{\theta_{\omega\cup a}}{\theta_a} \quad \text{such that the 2-way PCI – model is given by} \quad \theta_{1|2} = \theta_1 \quad \text{and} \quad \theta_{2|1} = \theta_2$$

Hierarchical log-linear models for 3-list CR table (and corresponding log-odds models)

Model restriction	Log-linear model interpretation
-	Saturated model
$\alpha_{123} = 0$	Null 2nd-order PCI-interaction among (L_1, L_2, L_3)
$\alpha_{12} = 0$	PCI between L_1 and L_2
$\alpha_{123} = -\alpha_{12}$	Conditional PCI between L_1 and L_2 given L_3
$\alpha_{12} = \alpha_{123} = 0$	Both PCI and conditional PCI between L_1 and L_2
$\alpha_{12} = \alpha_{13} = \alpha_{123} = 0$	PCI between L_1 and (L_2, L_3)
$\alpha_{12} = \alpha_{13} = \alpha_{23} = \alpha_{123} = 0$	Mutual PCI between L_1, L_2 and L_3

E.g. $\alpha_{123} = -\alpha_{12} \Rightarrow$ *conditional* PCI btw L_1 and L_2 given L_3

$$\eta_{123} = \log\theta_3 + \log\theta_{12|3} = \log\theta_3 + \log\theta_{1|3} + \log\theta_{2|3}$$

$$= \eta_3 + (\eta_{13} - \eta_3) + (\eta_{23} - \eta_3) = \alpha_3 + (\alpha_1 + \alpha_{13}) + (\alpha_2 + \alpha_{23})$$

Assume independent list errors $r_{\omega(\omega^c)} \sim \text{Binomial}(x_{\omega(\omega^c)}, \theta_{\omega(\omega^c)})$ conditional on $x_{\omega(\omega^c)}$. On observing $\mathbf{r} = (r_{\omega(\omega^c)})^T_{\omega \in \Omega_K}$ and $\mathbf{y} = \mathbf{x} - \mathbf{r}$, the log-likelihood is given as

$$\ell(\boldsymbol{\alpha}; \mathbf{r}|\mathbf{x}) = \sum_{\omega \in \Omega_K} \left(r_{\omega(\omega^c)} \log \theta_{\omega(\omega^c)} + y_{\omega(\omega^c)} \log(1 - \theta_{\omega(\omega^c)}) \right) \qquad (3)$$

for $\boldsymbol{\theta} = (\theta_{\omega(\omega^c)})^T_{\omega \in \Omega_K}$ arranged in correspondence with $\boldsymbol{\theta}_\Omega = (\theta_\omega)^T_{\omega \in \Omega_K}$ such that

$$\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\theta}_\Omega \qquad \text{for} \quad \mathbf{C} = \mathbf{Diag}(\mathbf{x}^{-1})\ \boldsymbol{\Gamma}\ \mathbf{Diag}(\mathbf{x}_\Omega)$$

where $\mathbf{x} = \boldsymbol{\Gamma}\mathbf{x}_\Omega$. The score vector and Hessian matrix are given by

$$\frac{\partial \ell}{\partial \boldsymbol{\alpha}} = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}}\right)^T \mathbf{d} = \left(\mathbf{C}\frac{\partial \boldsymbol{\theta}_\Omega}{\partial \boldsymbol{\alpha}}\right)^T \mathbf{d} = \left(\frac{\partial \boldsymbol{\theta}_\Omega}{\partial \boldsymbol{\alpha}}\right)^T \mathbf{C}^T\mathbf{d} = \boldsymbol{\Delta}^T\ \mathbf{Diag}(\mathbf{w})\ \mathbf{C}^T\mathbf{d}$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha}^2} = \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}}\right)^T \mathbf{Diag}(\mathbf{b}) \left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}}\right) + \mathbf{D}$$

for $\mathbf{d} = \frac{\partial \ell}{\partial \boldsymbol{\theta}}$, and $\log \boldsymbol{\theta}_\Omega = \boldsymbol{\Delta}\boldsymbol{\alpha}_\Omega$, and $\mathbf{Diag}(\mathbf{b}) = \frac{\partial \mathbf{d}}{\partial \boldsymbol{\theta}}$, and so on.

$L_1 \rightarrow$

y

$900 - r_{12}$	63
78	

L_2

$+$

r

r_{12}	87
222	

L_2

$=$

x

900	150
300	

L_2

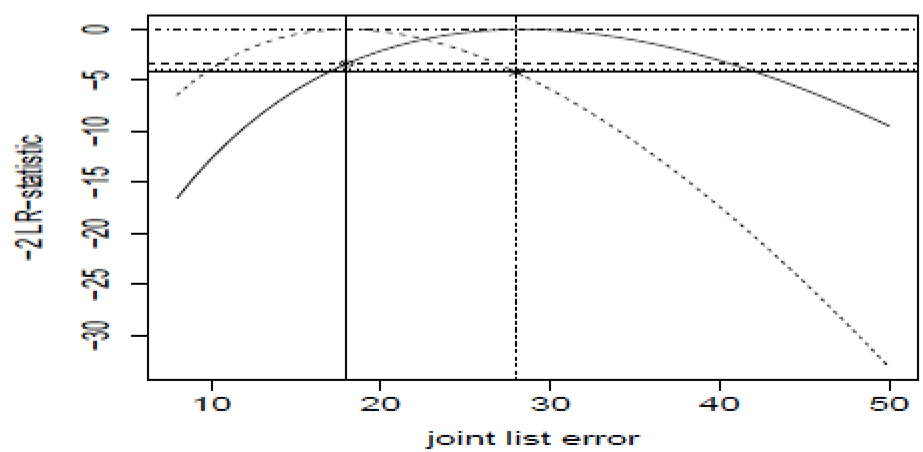


Figure 1: Fitting models with $\alpha_{12} = 0$ for $8 \le r_{12} \le 50$: -2LR of log-linear (curve dashed) and log-odds (curve solid); $r_{12}^0 = 18$ of log-linear and $r_{12}^0 = 28.4$ of log-odds. Critical region of 95% LR test (horizontal dotted).

Illustration cont'd: List-survey data

Estimates $(\hat{N}, \hat{r}_{12})$ for $\hat{r}_{12} = E(r_{12}|n_{12};\hat{\psi})$ assuming $\alpha_{12} = 0$. LLR-selection marked †.

	(N, r_{12}, n)			
$(\gamma_L, \gamma_s) = (0.95, 0.95)$	$(1087, 8, 1032)$	$(1077, 18, 1023)$	$(1066, 28, 1012)$	$(1056, 38, 1003)$
Log-linear $(\hat{N}, \hat{r}_{12})$	$(1074, 18.2)^\dagger$	$(1077, 18.0)^\dagger$	$(1078, 17.7)$	$(1081, 17.5)$
Log-odds $(\hat{N}, \hat{r}_{12})$	$(1061, 29.4)$	$(1063, 28.9)$	$(1065, 28.3)^\dagger$	$(1068, 27.8)^\dagger$
	(N, r_{12}, n)			
$(\gamma_L, \gamma_s) = (0.80, 0.95)$	$(1291, 8, 1226)$	$(1279, 18, 1215)$	$(1266, 28, 1202)$	$(1254, 38, 1191)$
Log-linear $(\hat{N}, \hat{r}_{12})$	$(1276, 18.2)^\dagger$	$(1279, 18.0)^\dagger$	$(1281, 17.7)$	$(1283, 17.5)$
Log-odds $(\hat{N}, \hat{r}_{12})$	$(1260, 29.4)$	$(1263, 28.9)$	$(1265, 28.3)^\dagger$	$(1268, 27.8)^\dagger$
	(N, r_{12}, n)			
$(\gamma_L, \gamma_s) = (0.90, 0.98)$	$(1148, 8, 1125)$	$(1137, 18, 1114)$	$(1126, 28, 1104)$	$(1114, 38, 1092)$
Log-linear $(\hat{N}, \hat{r}_{12})$	$(1135, 18.2)^\dagger$	$(1137, 18.0)^\dagger$	$(1139, 17.7)$	$(1141, 17.5)$
Log-odds $(\hat{N}, \hat{r}_{12})$	$(1125, 26.0)^*$	$(1123, 28.9)$	$(1125, 28.4)^\dagger$	$(1127, 27.8)^\dagger$

Note: Fitted log-likelihood equal to maximum achievable log-likelihood except marked *

Set-up: $(x_{1(2)}, x_{2(1)}, x_{12}, r_{1(2)}, r_{2(1)}) \xrightarrow{r_{12}} (y_{12}, y_L) \xrightarrow{\gamma_L} N \xrightarrow{\gamma_s} (n_{1(2)}, n_{2(1)}, n_{12}, n_{(L)})$.

- Beyond Census and PCS, without CPR
- Multiple lists from admin sources (in all EU)
  - GP, Income Tax, Electoral, School Enrolment, and ?
  - $K \geq 3$ will probably do (NB. Beware of partial coverage)
- Setting up DSE using multiple lists
  - Independence btw PCS and lists (not btw lists)
  - Instead of $(x \cdot \rho)$ use x0 by combining all the lists
    - Much cheaper than census
    - x0 should have negligible over-count – how?
    - x0 does not have to be representative
    - Development focus: list updating over time

II. Uncertainty analysis for categorical data fusion

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

- Many statistical problems characterised by lack of observations of interest
- Example: non-response

observed: $f(x,y,r=1) = f(x,r=1)f(y|x,r=1)$

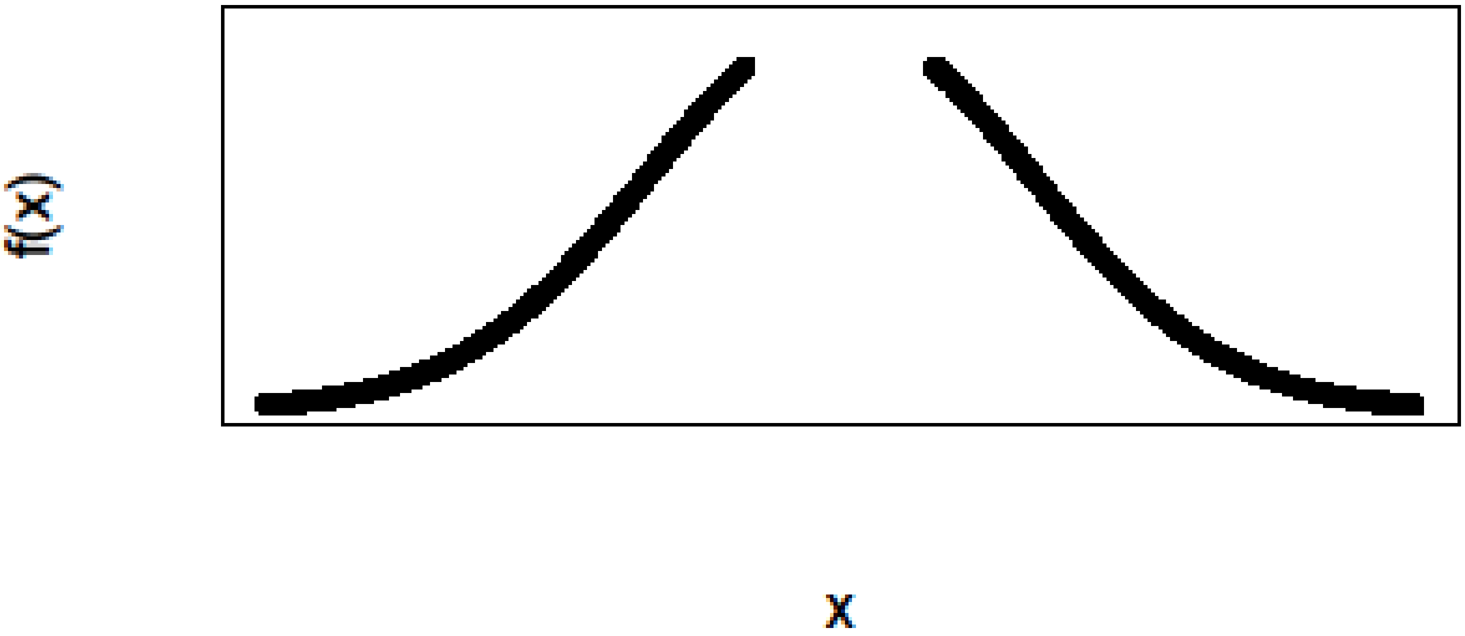
$$f(x,-,r=0) = f(x,r=0)$$

unobserved: $f(y|x,r=0)$

where $f(x,r=0) = \int f(x,y,r=0)dy$

$$= \int f(x,r=0)f(y|x,r=0)dy$$

Example: Censoring



Identification vs. sampling uncertainty

Identification uncertainty if "inference even from an infinite number of observations is suspect" (Koopmans, 1949)

Education	Didn't Vote	Did Vote	Total	Proportion
Low	(154)	(855)	1039	p
High	(28)	(676)	704	1-p
Total	182	1561	1743	
Proportion	q	1-q	*(Source: Election Survey 2005)*	

Theoretical **Fréchet bound**: max(0, p+q-1) ≤ P(Low, Didn't vote) ≤ min(p,q)
NB. p and q may be known or based on infinite number of observations
Estimated **Fréchet bound**: 0 ≤ P(Low, Didn't vote) ≤ 182/1743

- Generic setting
  - Two categorical variables (Y1, Y2)
  - P(Y1, Y2) unobserved; P(Y1) & P(Y2) observed
  - Possible auxiliary information X
- Application areas
  - Data fusion or statistical matching
  - Ecological inference
  - Market research
  - …
  - Micro data dissemination: information vs. confidentiality

Uncertainty space: the set of joint distributions that are compatible with the identified marginal distributions

Point-wise measure of uncertainty around $\theta_{ij} = P(Y_1 = i, Y_2 = j)$:

$$\max(0, \phi_i + \phi_j - 1) = L_{ij} \le \theta_{ij} \le U_{ij} = \min(\phi_i, \phi_j)$$

$$\Delta_{ij} \overset{def}{=} U_{ij} - L_{ij} \overset{lemma}{=} \min(\phi_i, 1-\phi_i, \phi_j, 1-\phi_j)$$

$$\text{for} \quad \phi_i = P(Y_1 = i) = \sum_j \theta_{ij} \quad \text{and} \quad \phi_j = P(Y_2 = j) = \sum_i \theta_{ij}$$

Overall measure (Conti et al., 2012):

$$\Delta = \sum_{i,j} w_{ij} \Delta_{ij}$$

NB. w_{ij} may vary by choice; $w_{ij} = \theta_{ij}$ unavailable

Education	Didn’t Vote	Did Vote	Total	Proportion
Low	(154)	(855)	1039	p
High	(28)	(676)	704	1-p
Total	182	1561	1743	
Proportion	q	1-q		

Estimated identification uncertainty (i.e. measure of uncertainty space):

$$\hat{\Delta} = \hat{\Delta}_{ij} \overset{\text{binary y's}}{\equiv} \min(\hat{p}, 1-\hat{p}, \hat{q}, 1-\hat{q}) = 182/1743 = 0.104$$

Sampling uncertainty:

$$E(\hat{\Delta}) \quad \text{and} \quad V(\hat{\Delta}) \quad \text{and} \quad \text{so on}$$

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Uncertaint y given auxiliary $X = 1,\ldots,K$:

Point - wise measure around $\lambda_{\mathrm{ij}}^{\mathrm{k}} = P(Y_1 = i, Y_2 = j \mid X = k)$:

$$\max(\lambda_{\mathrm{i}}^{\mathrm{k}} + \lambda_{\mathrm{j}}^{\mathrm{k}} - 1, 0) = L_{ij}^{k} \le \lambda_{\mathrm{ij}}^{\mathrm{k}} \le U_{ij}^{k} = \min(\lambda_{\mathrm{i}}^{\mathrm{k}}, \lambda_{\mathrm{j}}^{\mathrm{k}})$$

$$\Delta_{ij}^{k} \overset{def}{=} U_{ij}^{k} - L_{ij}^{k}$$

$$\overline{\Delta}_{\mathrm{ij}} \overset{def}{=} E_X(\Delta_{ij}^{k}) = \sum_{k} \phi_k \Delta_{ij}^{k}$$

for $\lambda_i^k = P(Y_1 = i \mid X = k)$ and $\lambda_j^k = P(Y_2 = j \mid X = k)$ and $\phi_k = P(X = k)$

By Jensen' s inequality :

$$\overline{\Delta}_{\mathrm{ij}} \le \Delta_{ij}$$

Relative efficiency (RE) of auxiliary informatio n (X) :

$$\mathrm{RE} = \overline{\Delta} / \Delta \qquad \text{where} \qquad \overline{\Delta} = \sum_{i,j} w_{ij} \overline{\Delta}_{ij}$$

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Proxy data: Education-in-register, Turnout from Electoral Office

Education	Didn't Vote	Did Vote	Total
Low	210	920	1130
High	44	569	613
Total	254	1489	1743

Important distinction btw proxy and other auxiliary variables:
- proxy: similar definition & same support
- when available, proxy variables are usually the most power auxiliary
- affects not only uncertainty but also fusion techniques

Available Proxy	Uncertainty Bound (Lower, Upper)	Width (SE)	RE
-	(0.000, 0.104)	0.104 (0.0073)	1
Register Education	(0.018, 0.104)	0.086 (0.0065)	0.83
Election Turnout	(0.065, 0.104)	0.039 (0.0044)	0.38
Both	(0.074, 0.095)	0.021 (0.0034)	0.20

- Data fusion: identification uncertainty with SE is 0.021 +/- 2*0.0034
- True-data 95% confidence interval has estimated length 0.027
- If Education and Voting in two separate surveys: no need for additional survey in order to collect the joint observations

NB. Plausible pseudo point-estimates available from *proxy* data fusion

Uniwersytet Ekonomiczny w Poznaniu

Kadry dla Gospodarki
Twój sukces się liczy

Urząd Statystyczny w Poznaniu

Theoretical identification bounds and width given auxiliary data (all numbers in 10^4)

Joint auxiliary data	$P(E,E)=6411$ Lower	Upper	Width	$P(E,U)=71$ Lower	Upper	Width	$P(E,N)=685$ Lower	Upper	Width
(Z_1,Z_2)	6038	6738	702	19	149	130	357	1065	708
(Z_1,Z_2,Z_t,Y_t)	6287	6718	431	26	137	111	376	811	435
Y_t	6241	7164	923	0	151	151	0	857	857
Joint auxiliary data	$P(U,E)=98$ Lower	Upper	Width	$P(U,U)=9$ Lower	Upper	Width	$P(U,N)=60$ Lower	Upper	Width
(Z_1,Z_2)	19	151	132	0	94	94	9	146	137
(Z_1,Z_2,Z_t,Y_t)	40	148	108	2	86	84	11	120	109
Y_t	0	167	167	0	138	138	0	167	167
Joint auxiliary data	$P(N,E)=710$ Lower	Upper	Width	$P(N,U)=71$ Lower	Upper	Width	$P(N,N)=1885$ Lower	Upper	Width
(Z_1,Z_2)	381	1086	705	0	131	131	1505	2216	711
(Z_1,Z_2,Z_t,Y_t)	400	835	435	2	119	117	1754	2199	445
Y_t	0	880	880	0	151	151	1721	2630	909

Simulated: survey (Y_1, Y_t, Y_2) in 2011, 2012 and 2013, respectively; register (Z_1, Z_t, Z_2), correspondingly

UNIWERSYTET EKONOMICZNY W POZNANIU

Kadry dla Gospodarki
Twój sukces się liczy

URZĄD STATYSTYCZNY W POZNANIU

Thank you!

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

UNIWERSYTET EKONOMICZNY W POZNANIU

Kadry dla Gospodarki
Twój sukces się liczy

URZĄD STATYSTYCZNY W POZNANIU

Population size estimation based on multiple lists

Uncertainty analysis for categorical data fusion

Li-Chun Zhang

University of Southampton (L.Zhang@soton.ac.uk)

Statistics Norway (lcz@ssb.no)

Projekt Kadry dla Gospodarki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY