Problem of Missing Data in Census Who Are the Non-Response Respondents

Jan Hora, Jiří Grim

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

Statistics, 2009, Prague

Outline

1. Standard methods for census results publication
2. Data reproduction by means of a statistical model
   - Basic idea
   - Definition
3. Experimental results - Czech census 2001
   - Missing data
   - Model calculation and it's accuracy
   - Error distribution
4. Demo - Interactive data presentation
5. Conclusions

Standard methods for census results publication

Basic problem: Completeness and accuracy vs. Privacy

- Task 1: Complex and accurate representation of the census result
- Task 2: Preserve anonymity of single persons
- Problem: Even anonymized query form could be identified

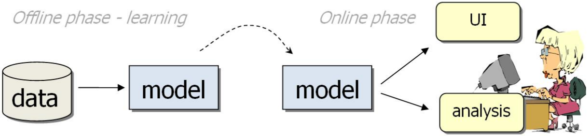
Aggregated data, Tables

perfect accuracy
limited information amount published

Anonymized sets of microdata (1-10% of the original data set)

unlimited possibilities for query formulation
anonymization required, limited distribution, limited accuracy

Data reproduction by means of a statistical model



Statistical model as a knowledge base for probabilistic expert system

- Model is "trained" to represent the original data set
- Model is used instead of the original data
- Statistical properties represented without the original data
- Accuracy decreases with subpopulation size

Statistical model - distribution mixture

Census data .. N-dimensional discrete finite data set $\mathcal{S}$
Query-form .. N-dimensional vector $\mathbf{x}$ of answers

Assumption

$\mathbf{x}$ .. random vector,
$\mathcal{S}$ .. set of iid observations of $\mathbf{x}$,
$P(\mathbf{x})$ can be approximated by a distribution mixture.

$$P(\mathbf{x}) = \sum_{m=1}^{M} w_m F(\mathbf{x}|m) = \sum_{m=1}^{M} w_m \prod_{n=1}^{N} p_n(x_n|m)$$

M	.. component count
w_m	.. weight of the m-th component
$F(\mathbf{x}\|m)$	.. conditional distribution
$p_n(.\|m)$	.. marginal conditional distribution of x_n

Census and Missing data

- $N = 24$ questions, $|\mathcal{S}| = 10230060$ respondents
- 1524240 incomplete records, 2933427 total non-response

	Text of question (name of variable)	Number of values	Non-response in %	Shannon entropy in %
1.	Region of residence	14	0.00	96.88
2.	Type of residence	3	0.00	32.92
3.	Economic activity	10	0.80	67.80
4.	Birth place (relatively)	6	1.95	74.65
5.	Religion	6	0.00	60.57
6.	Occupation type	14	3.89	68.33
7.	Sex	2	0.00	99.95
8.	Marital status	4	0.55	81.01
9.	Education	14	1.11	78.04
10.	Age	9	0.03	96.09
11.	Category of flat	5	0.53	27.81
12.	Bathroom	5	0.59	14.02
13.	Size of flat	7	0.64	80.62
14.	Internet and PC	4	2.85	49.11
15.	Legal relation to flat	9	0.39	72.43
16.	Gas supply	3	0.78	64.54
17.	Number of rooms over 8m^2	7	0.64	80.57
18.	Number of cars in household	4	3.39	71.32
19.	Number of persons in flat	6	0.00	93.79
20.	Vacational property	6	7.45	42.10
21.	Telephone in flat	5	1.80	80.88
22.	Water supply	4	0.35	8.02
23.	Type of heating	6	0.53	74.81
24.	Toilet	6	0.50	16.73

Two ways of handling the missing data

I. Extended model - non-response as a new possible answer

- Simple solution
- Non-responding respondents form specific subpopulations
- ⇒ analysis of their properties

II. Substitution - missing values filling-up

- Model estimated from incomplete data
- Estimated model used for missing data substitution
- Correct substitution cca 73%

Example - Model accuracy

Czech census 2001 - Model with $M = 15000$ components

Used model	Model with substituted values	Extended model with missing values
Mean relative error in %:	**4.07**	**4.10**
Standard deviation of the relative error:	6.33	5.83
Maximum relative error of the model in %:	240.84	250.90
Number of relative errors exceeding 100%:	925	1037
Mean absolute error:	**470**	**459**
Standard deviation of the absolute error:	951	791
Maximum absolute error of the model:	45779	56808
Number of combinations tested:	**3503448**	**3895873**

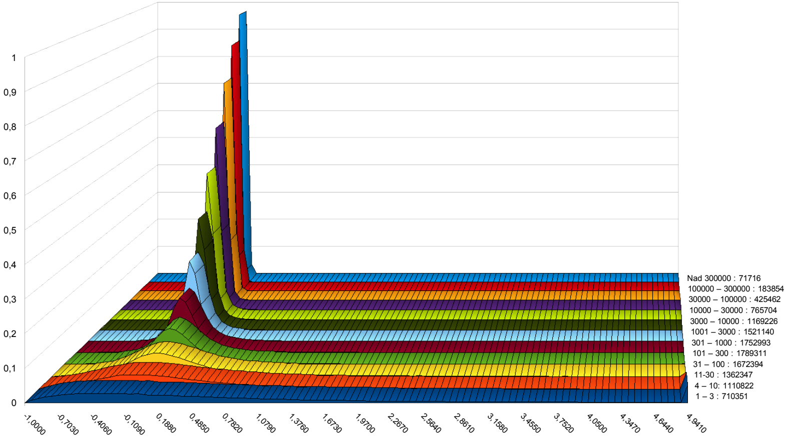
Validation set $\mathcal{A}$: Combinations of up to 4 answers greater than 1612

Mean relative error:

$$\epsilon_R(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} \frac{|\hat{P}(A) - P(A)|}{\hat{P}(A)} \qquad P(A) \text{.. original size, } \hat{P}(A) \text{ .. estimated size}$$

Example - Accuracy depending on the subpopulation size

Relative error distribution according to the subpopulation size



(x-axe: error distribution, z-axe: subpopulation size)

Interactive data presentation

Conclusions

Method characteristics

- Guaranteed anonymity
- Accuracy comparable to microdata sets
- Successfully implemented for Czech Census 2001

Non-Response handling

- Non-Responding respondents analysis
- Filling-up the missing values

Thank You for Attention

Demo application: *http://ro.utia.cas.cz/dem.html*

Jan Hora, *hora@utia.cas.cz*
Jiří Grim, *grim@utia.cas.cz*

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic