Population size estimation

Danutė Krapavickaitė

Vilnius Gediminas Technical University
Statistics Lithuania

Workshop of Baltic-Nordic-Ukrainian Network
on Survey Statistics
August 21-24, 2018, Jelgava, Latvia

Outline

Presentation includes following items:

- Introduction
- The dual system model
- Maximum likelihood estimator for a population size
- Loglinear model for population size estimation
- Loglinear model + 1 covariate
- Loglinear model + 2 covariates
- Final remarks

Introduction

Nordic countries & Austria already had the register-based censuses.

Baltic counties will have the administrative data-based censuses in 2021 firstly.

The methods needed for that should be intensively studied and developed.

Under census regulation the quality report is obligatory.

One of the quality aspects is under-coverage of the census data.

Population size should be estimated for that.

The dual system model

Two lists for population U elements available: list A and list B.
Each person in the population U is either
in or *not in* list A as well as either
in or *not in* list B.

This model leads to the following 2x2 Table of counts,
assuming no matching errors, fictitious cases and other erroneous inclusions:

Table 1. Population counts

		List B		
		In (1)	Out (0)	Total
List A	In (1)	N_{11}	N_{10}	N_{1+}
	Out (0)	N_{01}	$\mathbf{N_{00}}$	N_{0+}
	Total	N_{+1}	N_{+0}	$\mathbf{N}$

Aim: to estimate $\mathbf{N_{00}}$, $\mathbf{N}$

Survey data

The population $\mathcal{U}$ can be

- population of immigrants (legal and illegal), included into population register and police list;
- the set of a road accidents, included into police list and hospital list;
- population of a country included into the population census list and post-enumeration survey list;
- population of the natural resources.

Probabilistic notations

Let us denote the probability to be included into the 2x2 Table cell (i,j) by $\pi_{ij}^{(l)}$, $i,j = 1,0$, $l = 1,2,...,N$, the probability to be included into the 2x2 Table cell (i,j).

Table 2. Inclusion probabilities into a Table

		List B		
		In (1)	Out (0)	Total
List A	In (1)	$\pi_{11}^{(l)}$	$\pi_{10}^{(l)}$	$\pi_{1+}^{(l)}$
	Out (0)	$\pi_{01}^{(l)}$	$\pi_{\mathbf{00}}{}^{(l)}$	$\pi_{0+}^{(l)}$
	Total	$\pi_{+1}^{(l)}$	$\pi_{+0}^{(l)}$	

Assumptions

(a) The event for the lth individual to be included into a list A is **independent** of the event of its inclusion into a in list B: $\pi_{ij}^{(l)} = \pi_{i+}^{(l)} \cdot \pi_{+j}^{(l)}$, $(i, j = 1, 0)$, $\pi_{i+}^{(l)}$ and $\pi_{+j}^{(l)}$ are marginal probabilities;

(b) The marginal probabilities $\pi_{i+}^{(l)}$ and $\pi_{+j}^{(l)}$ are **homogenious** across individuals l: $\pi_{i+}^{(l)} = \pi_{i+}$, $\pi_{+j}^{(l)} = \pi_{+j}$ *at least for one of the lists*;

(c) The population U is **closed** (no changes due to death or birth during the study period);

(d) It is possible **to link** elements of the lists A and B **perfectly**.

Sample

Due to inclusion probabilities for the population elements less than 1, no exact cell counts for subpopulations enumerated are known, but only the random values of a sample:

Table 3. Sample counts

		List B		
		In (1)	Out (0)	Total
List A	In (1)	n_{11}	n_{10}	n_{1+}
	Out (0)	n_{01}	$\mathbf{n_{00}}$	n_{0+}
	Total	n_{+1}	n_{+0}	$\mathbf{n}$

The sample cell value $n_{00} = 0$ and the total size n is not known.
Aim: to estimate them.

Maximum likelihood estimator

Under assumptions (a)-(d), the multinomial distribution is applied to the cell counts in the Table 3, and the likelihood function is given by

$$L(n, \pi_{1+}, \pi_{+1}) =$$
$$= \frac{n!}{n_{11}!n_{10}!n_{01}!n_{00}!}\pi_{1+}^{n_{1+}} \cdot \pi_{+1}^{n_{+1}} \cdot (1-\pi_{1+})^{n-n_{1+}} \cdot (1-\pi_{+1})^{n-n_{+1}}.$$

The maximum likelihood estimators for N, π_{1+} and π_{+1} are

$$\tilde{N} = \tilde{n} = \left[\frac{n_{1+}n_{+1}}{n_{11}}\right],$$
$$\tilde{\pi}_{1+} = n_{11}/n_{+1}, \quad \tilde{\pi}_{+1} = n_{11}/n_{1+}.$$

Estimator of variance

Estimator of variance for fixed sample size n due to Sekar and Deming (1949):

$$\widehat{Var}(\tilde{n}) = (n_{1+} \cdot n_{+1} \cdot n_{10} \cdot n_{01})/n_{11}^3.$$

based on the Petersen estimator (1896).

Petersen introduced the method to estimate a number of fish in a lake.
The initial fish sample of size n_{1+} is captured (list A), marked and released to go back into the lake.

The second fish sample of size n_{+1} is captured.
The number of the marked fish which are recaptured in the second sample, n_{11}, is used to estimate the size of the lake fish population N.

The method is called by *capture-recapture*.

Post-stratidication

Assumption (b) (homogeneity of inclusion probabilities) hardly holds for human population.

One of the ways to overcome it is **post-stratification** of the population and consider assumption (b) to be valid for post-strata.

Post-enumeration surveys

Population censuses are aimed to make complete enumeration of the population and to find N.

But they have often omissions, duplications, false enumerations, miss-classifications.
It means, the net *census undercount* is available.

The population is post-stratified by age, sex, region.
The dual system method is used for post-strata.

In a traditional census, the census list is considered as list A. A *post-enumeration survey* (PES) is carried out. Probability sample is drawn in a post-stratum, it is considered as a list B.
Post-stratum population size is estimated.

In an administrative data-based census, two registers play a role of list A and list B.
Alternative: census enumeration list and one of the registers.

Example, USA 1980. Availability of erroneous inclusions

When list B is *sampled* in PES, the values N_{11}, N_{+1} in the estimator $\tilde{N} = [N_{1+}N_{+1}/N_{11}]$ are not known, because list *B is not enumerated completely*. The marginal count n_{1+} differs from N_{1+} because of *false enumerations* and PES data available only for sample.

PES was based on 2 samples:
Current population survey 1980 (CPS) sample;
list E – enumeration data, from which erroneous inclusions were deleted.
The count N_{11} was estimated by matching the CPS sample to the list E and estimated using CPS weights.
N_{1+} estimated using CPS weights.
N_{+1} estimated by subtracting an estimate of erroneous inclusions (size of list E) from the census count.
The estimator for N in post-stratum takes the form

$$\hat{N} = (\hat{N}_{+1} \cdot \hat{N}_{1+})/\hat{N}_{11}, \quad (\text{Rao}, 2003).$$

Violation of the assumption (a) (independency of the element inclusion into the lists)

Model is constructed to estimate population size.

Known approaches:

- to use covariates in the model, whose levels have heterogeneous inclusion probabilities for both lists and loglinear models.
- to use the third list, three-way contingency tables and loglinear models.
- to use latent variables to take heterogeneity of inclusion probabilities into account.

We present here *the first approach* developed by *van der Heijden et al*.

Assumptions are valid

Let us consider the numbers n_{ij} as random values, their expectations denote by $m_{ij} = En_{ij}$, $i, j = 1, 0$.

The problem. We need to estimate $m_{00} = En_{00}$.
The estimator obtained may be used to estimate the total number of the population elements N or the size of its subpopulation.

Case 1. Let us assume all four *assumptions are valid*.
1^{st} solution. The cell count m_{00} can be estimated using Sekar and Deming (1949) estimator for a fixed sample size $n = n_{11} + n_{10} + n_{01}$: $\hat{m}_{00} = n_{10}n_{01}/n_{11}$ with the variance estimator $\widehat{Var}(\hat{m}_{00}) = n_{1+}n_{+1}n_{10}n_{01}/n_{11}^3$.

It can be used also to estimate the variance for the population size N estimator $\hat{N} = n_{+1} + n_{10} + \hat{m}_{00}$:
$\widehat{Var}(\hat{N}) = \widehat{Var}(\hat{m}_{00})$.
Normal approximation for $\hat{N}$ can be used to estimate confidence interval for N.

2^{nd} solution

In the case of independency (1st assumption),
$\pi_{ij} = P(A = i, B = j) = \pi_{i+} \cdot \pi_{+j}$ and the cell count expectation m_{00} can be estimated using a loglinear model:

$$\log(m_{ij}) = \lambda + \lambda_i^{(A)} + \lambda_j^{(B)}.$$

This method is preferable because it allows to estimate the more general log-linear models as well.
Each term on the right-hand side of the $\log(m_{ij})$ means the contribution of the corresponding factor (inclusion to the list A or inclusion to the list B) to the value of $log(m_{ij})$ like in the case of ANOVA.
Restrictions $\lambda_0^{(A)} = \lambda_0^{(B)} = 0$ are used to get unique solution.
The loglinear model above is estimated by the maximum likelihood method under the assumption that the n_{ij} follow the Poisson distribution.
After estimation of the model parameters we obtain
$\hat{m}_{00} = \exp(\hat{\lambda})$, $\hat{m}_{11} = \exp(\hat{\lambda} + \hat{\lambda}_1^{(A)} + \hat{\lambda}_1^{(B)})$,
$\hat{m}_{10} = \exp(\hat{\lambda} + \hat{\lambda}_1^{(A)})$, $\hat{m}_{01} = \exp(\hat{\lambda} + \hat{\lambda}_1^{(B)})$,

Case 2. One covariate

Usually the first *independency assumption (a) is violated*, inclusion of the an element into both lists is not independent in the human populations.
This assumption *cannot be verified from the data*.
Therefore it is assumed that there is a covariate X *available in both of the lists*,
and conditionally on the known value of the covariate, inclusion of the an element into both of the lists becomes independent.
Let the levels of a covariate X be indexed by 1, 0. It means that the first assumption is loosened and replaced by the *conditional independency* under the known value x of the covariate X:

$$\begin{aligned} \pi_{ijx} &= P(A=i, B=j, X=x) \\ &= P(A=i|X=x)P(B=j|X=x)P(X=x) \\ &= \pi_{i|x}\pi_{j|x}\pi_{x}. \end{aligned}$$

n_{ijx} – cell count for A=i, B=j, $X=x$.

Model for one covariate

Table 4. Expected values of the observed counts $m_{ijx} = En_{ijx}$, $i, j = 1, 0$:

	$X = 1$		$X = 0$	
	B=1	B=0	B=1	B=0
A=1	m_{111}	m_{101}	m_{110}	m_{000}
A=0	m_{011}	$\mathbf{m}_{001}$	m_{010}	$\mathbf{m}_{000}$

The expected counts m_{001} and m_{000} are unknown and should be estimated, while the corresponding observed counts are
$n_{001} = n_{000} = 0$
The loglinear model for m_{ijx} in this case is

$$log(m_{ijx}) = \lambda + \lambda_i^{(A)} + \lambda_j^{(B)} + \lambda_x^{(X)} + \lambda_{ix}^{(AX)} + \lambda_{jx}^{(BX)}$$

Restrictions, solution

Restrictions:

$$\lambda_0^{(A)} = \lambda_0^{(B)} = \lambda_0^{(X)} = \lambda_{01}^{(AX)} = \lambda_{01}^{(BX)} = 0$$

Because of independence between A and B conditional on X,
$\lambda_{ij}^{AB} = \lambda_{ijx}^{(ABX)} = 0$.
The loglinear model is estimated under assumption:
the counts n_{ijx} follow a Poisson distribution.
After estimation of the model parameters we obtain the solution needed:
$\hat{m}_{001} = \exp(\hat{\lambda} + \hat{\lambda}_1^{(X)})$,
$\hat{m}_{00} = \exp(\hat{\lambda})$.

Case 3. Two covariates

Two covariates are used to weaken the first independency assumption: X_1 with the values in A (not available in B), X_2 with the values in the list B (not available in A).

Table 5. Expected values of the observed counts $m_{ijkl} = En_{ijkl}$, $i, j = 1, 0$ for two lists and two partially observed covariates

		$B=1$		$B=0$	
		$X_2=1$	$X_2=0$	$X_2=1$	$X_2=0$
	$X_1=1$	m_{1111}	m_{1110}	m_{1011}	m_{1010}
A=1	$X_1=0$	m_{1101}	m_{1100}	m_{1001}	m_{1000}
	$X_1=1$	m_{0111}	m_{0110}	$\mathbf{m_{0011}}$	$\mathbf{m_{0010}}$
A=0	$X_1=0$	m_{0101}	m_{0100}	$\mathbf{m_{0001}}$	$\mathbf{m_{0000}}$

Counts m_{0111}, m_{0101} are not observed,only the sum $m_{0111} + m_{0101}$, because X_1 is not known for elements, for which $A = 0$.
Similarly there are observed only sums
$m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$, $m_{1001} + m_{1000}$.
There are no observed values for m_{0011}, m_{0001}, m_{0010}, m_{0000}.

Conditional independency and model

The assumption of independency of inclusion of the population element into two lists is replaced by the weaker assumption of conditional independency:

$$\begin{aligned}\pi_{ijkl} &= P(A=i, B=j, X_1=k, X_2=l) \\ &= P(A=i|X_1=k)P(B=j|X_2=l)P(X_1=k, X_2=l) \\ &= \pi_{i|k}\pi_{j|k}\pi_{kl}.\end{aligned}$$

The loglinear model:

$$log(m_{ij}) = \lambda + \lambda_i^{(A)} + \lambda_j^{(B)} + \lambda_k^{(X_1)} + \lambda_l^{(X_2)} + \lambda_{il}^{(AX_1)} + \lambda_{jk}^{(BX_2)} + \lambda_{kl}^{(X_1X_2)}.$$

Restrictions.
EM algorithm is used to obtain maximum likelihood estimates.
Model parameters can be assessed using R package `cat`.
Bootstrap is used to estimate confidence intervals for N.

Example, two covariates

Table 6. Created data set, $n = 27$. The observed counts

		$B = 1$		$B = 0$
		$X_2 = 1$	$X_2 = 0$	X_2 missing
	$X_1 = 1$	3	1	6
A=1	$X_1 = 0$	2	4	4
A=0	X_1 missing	4	3	–

Table 7. The fitted frequencies

		$B = 1$		$B = 0$	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
	$X_1 = 1$	2.0	2.0	3.0	3.0
A=1	$X_1 = 0$	3.0	3.0	2.0	2.0
	$X_1 = 1$	1.6	1.2	2.4	1.8
A=0	$X_1 = 0$	2.4	1.8	1.6	1.2

$\hat{N} = 34,\ P(N \in (25.3; 48)) = 0.95)$

Final remarks

Other research directions:

- three lists and loglinear models (Netherlands);
- Bayesian methods used to estimate distributions of the model parameters (New Zealand);
- latent variables to take heterogeneity of the inclusion probabilities into account (ISTAT);
- linkage errors for two lists available (Netherlands&ISTAT).

Methods presented here can be used in practice.

References

1. Rao J. N. K. (2003), *Small Area Estimation*. Hoboken: John Wiley & Sons.

2. Gerritse S., van der Heijden P. G. M., and Bakker B. F. M. (2015), Sensitivity of Population Size Estimation for Violating Parameter Assumptions in Log-linear Models. *Journal of Official Statistics*, **31**(3), pp. 357-379.

3. Heijden, P. G. M. van der, Whittaker, J., Cruyff, M. J. L. F., Bakker, B. F. M., and Van der Vliet, H.N. (2012), People Born in the Middle East but Residing in the Netherlands: Invariant Population Size Eestimates and the Role of Active and Passive Covariates. *The Annals of Applied Statistics* **6**, pp. 831–852. `https://eprints.soton.ac.uk.344644`
(accessed on 1 August 2018)

4. Heijden, P. G. M. van der, Smith P. A., Cruyff M., and Bakker B. (2018) An Overview ... *JOS* **34**(1), pp. 239-263.

Thank you for your attention!