Modern censuses: The examples of the integrated census and the rolling census

Ronit Nirel
nirelr@cc.huji.ac.il

May 2011

Census of population

"the operation that produces at regular intervals the official counting (or benchmark) of the population in the territory of a country and in its smallest geographical sub-territories together with information on a selected number of demographic and social characteristics of the total population"

(United Nations Economic Commission for Europe UNECE, 2006, p. 6, no. 19).

Essential features

Universality

A well-defined census population

Simultaneity of information

A reference date for all census data (Census Day)

Individual enumeration:

Accurate data pertaining to individuals with regard to place of residence and other socio-demographic characteristics on Census Day

Alternative approaches to census-taking

Traditional census

Traditional with yearly updates

Rolling census

Register-based census

Combination of register with survey data

Register-based and complete enumeration

Why change the traditional approach?

Availability of massive administrative data

Growing strength of the information technolog

Growing demand for timely and detailed data

Changes in public opinion and decline in survey response rates

Increasing financial burden on national budgets

Outline

IC:

The extended coverage model

Design of the coverage surveys

Example

RC:

Concept and implementation

The Integrated Census (IC)

Replaces the traditional field enumeration by enumeration of the Population Register (PR)

But… the PR is exposed to coverage errors

Uses sample surveys to

(a) estimate coverage errors of the PR

(b) set a reference date

The problem

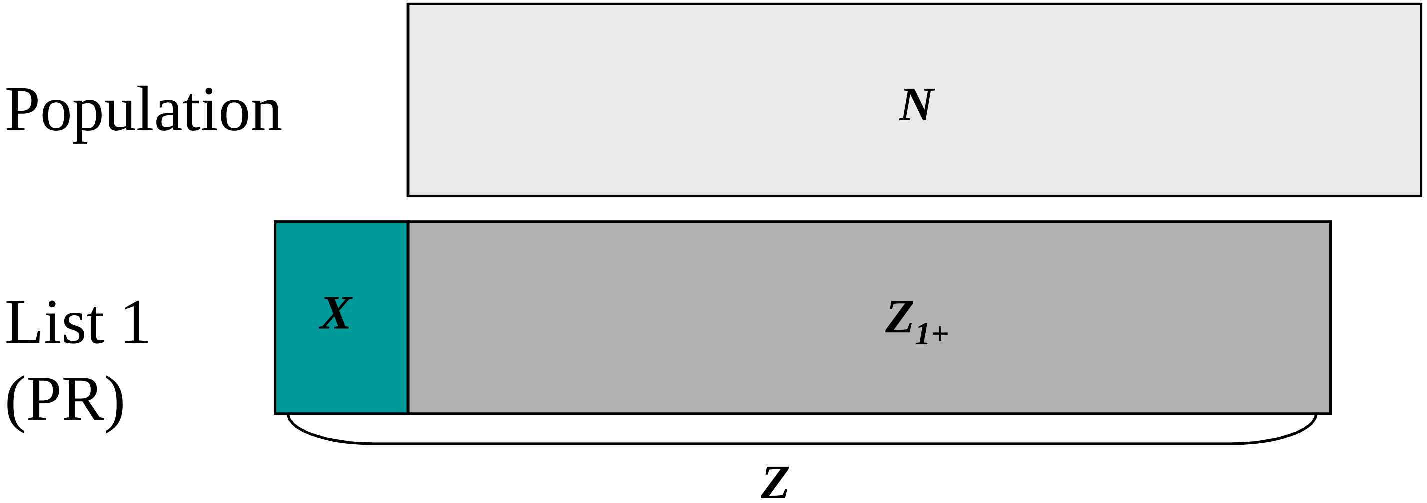
Let $\boldsymbol{N}$ be the number of eligible people in a closed population E in a given geographical area

$\boldsymbol{N}$ is fixed but unknown

Aim: Estimate $\boldsymbol{N}$

Data: A list that fails to capture every individual (undercount) and also counts individuals that do not belong to E (overcount/false captures).

Model overview



Undercount parameter $p_{1+} = EZ_{1+} / N$

Overcount parameter $\lambda = EX / N$

Multinomial model *(DSE, Peterson, 1986)*

Suppose that List 1 is exposed only to **undercount.** Then, consider a capture-recapture experiment.

Capture history of individual k is *multinomial*, that is

$$\mathbf{Z}(k) \sim Mult(\mathbf{p}(k))$$

where $\mathbf{Z}(k) = (Z_{11}(k), Z_{12}(k), Z_{21}(k), Z_{22}(k))$ and $\mathbf{p}(k) = (p_{11}(k), p_{12}(k), p_{21}(k), p_{22}(k))$.

Let $Z_{1+}(k) = Z_{11}(k) + Z_{12}(k)$ and $p_{1+} = p_{11}(k) + p_{12}(k)$

Then $p_{1+}(k) = E(Z_{1+}(k))$ is the capture probability of k in List 1.

Multinomial model (2)

Main assumptions

Homogeneity. Capture probabilities within each list are equal, that is

$$p_{1+}(k) = p_{1+}, \quad p_{+1}(k) = p_{+1} \quad k = 1, \ldots, N$$

Independence. Captures are independent for each k, that is

$$p_{11}(k) = p_{1+}(k) p_{+1}(k) \qquad k = 1, \ldots, N$$

Multinomial model (3)

Under the above assumptions, the MLEs of the capture probability of List 1 and of the population size are given by

$$\hat{p}_{1+} = \frac{Z_{11}}{Z_{+1}} \quad \text{and} \quad \hat{N} = \frac{Z_{1+}}{\hat{p}_{1+}}$$

where $Z_{ab} = \sum_{k=1}^{N} Z_{ab}(k), \;\; a, b = 1, 2, +.$

	List 2			
List 1 (PR)		**In**	**Out**	
	In	Z_{11}	Z_{12}	Z_{1+}
	Out	Z_{21}	Z_{22}	Z_{2+}
		Z_{+1}	Z_{+2}	N

Multinomial-Poison model

(Glickman,Nirel,Ben Hur 2003)

In addition, assume that the number of false captures X in List 1 is Poisson, that is

$$X \sim Poisson(\lambda N),$$

and adding the assumption:

Capture history and false captures are independent, that is $\{\mathbf{Z}(k)\}$ and X are independent

Multinomial-Poison model (2)

Then under the above assumptions, the MLEs of the parameters of interest are given by

$$\hat{p}_{1+} = \frac{Z_{11}}{Z_{+1}}, \ \hat{N} = \frac{Z_{1+}}{\hat{p}_{1+}} \ \text{ and } \ \hat{\lambda} = \frac{X}{\hat{N}}.$$

Denote $Z = Z_{1+} + X$ then $p_{1+} + \lambda = EZ/N.$

Hence the estimate of the population size is

$$\hat{N} = \frac{Z}{\hat{p}_{1+} + \hat{\lambda}}$$

where Z is the (***known***) size of List 1.

Census Weight

Bias and Variance

Taylor approximations yield

$$\mathrm{E}\hat{N} \doteq N + C \quad \text{and} \quad \mathrm{Var}(\hat{N}) \doteq N\ C$$

For a one-stage sample of m clusters out of M

$$C = A + \frac{M-m}{m} B$$

$$A = \frac{(1-p_{1+})(1-p_{+1})}{p_{1+}\,p_{+1}} \qquad B = \frac{1-p_{1+}}{p_{1+}\,p_{+1}} + \frac{\lambda}{p_{1+}+\lambda}\left(\frac{p_{1+}}{p_{1+}+\lambda} - \frac{1-p_{1+}}{p_{1+}}\right)$$

Coverage Surveys

Estimates of the coverage parameters are obtained through two coverage surveys:

The *U-survey* is based on an area sample and estimates the undercount rates

The *O-survey* is based on a sample of people from the PR and estimates the overcount rates

Independence between U-enumeration and the PR data is kept to avoid bias

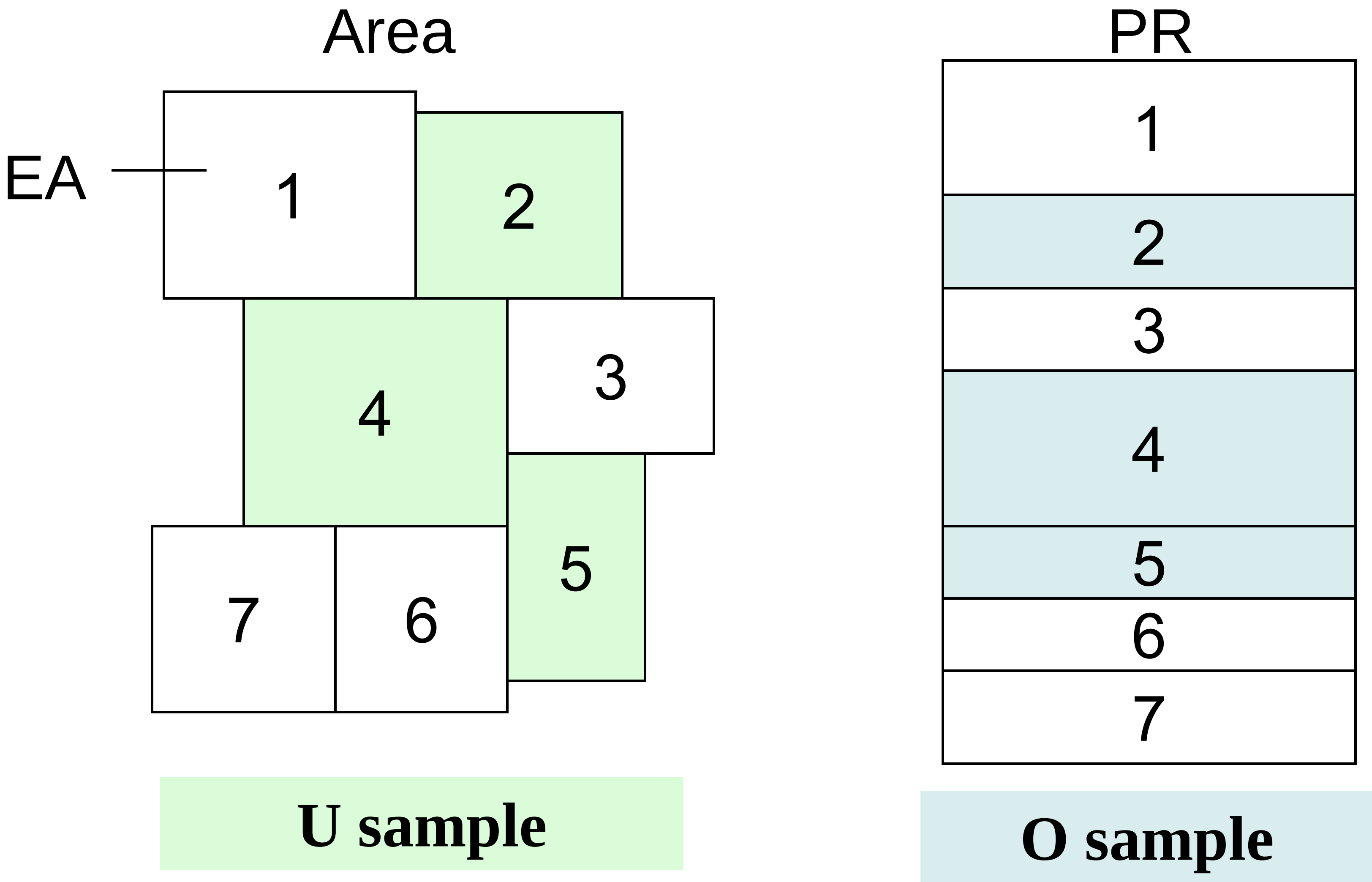
Sample design

Overall sample size. Typically large, e.g. about 1/5 of the population in both samples

Sampling plan. Typically cluster sampling, e.g. select a simple random sample of m enumeration areas (EAs) out of M in each statistical area

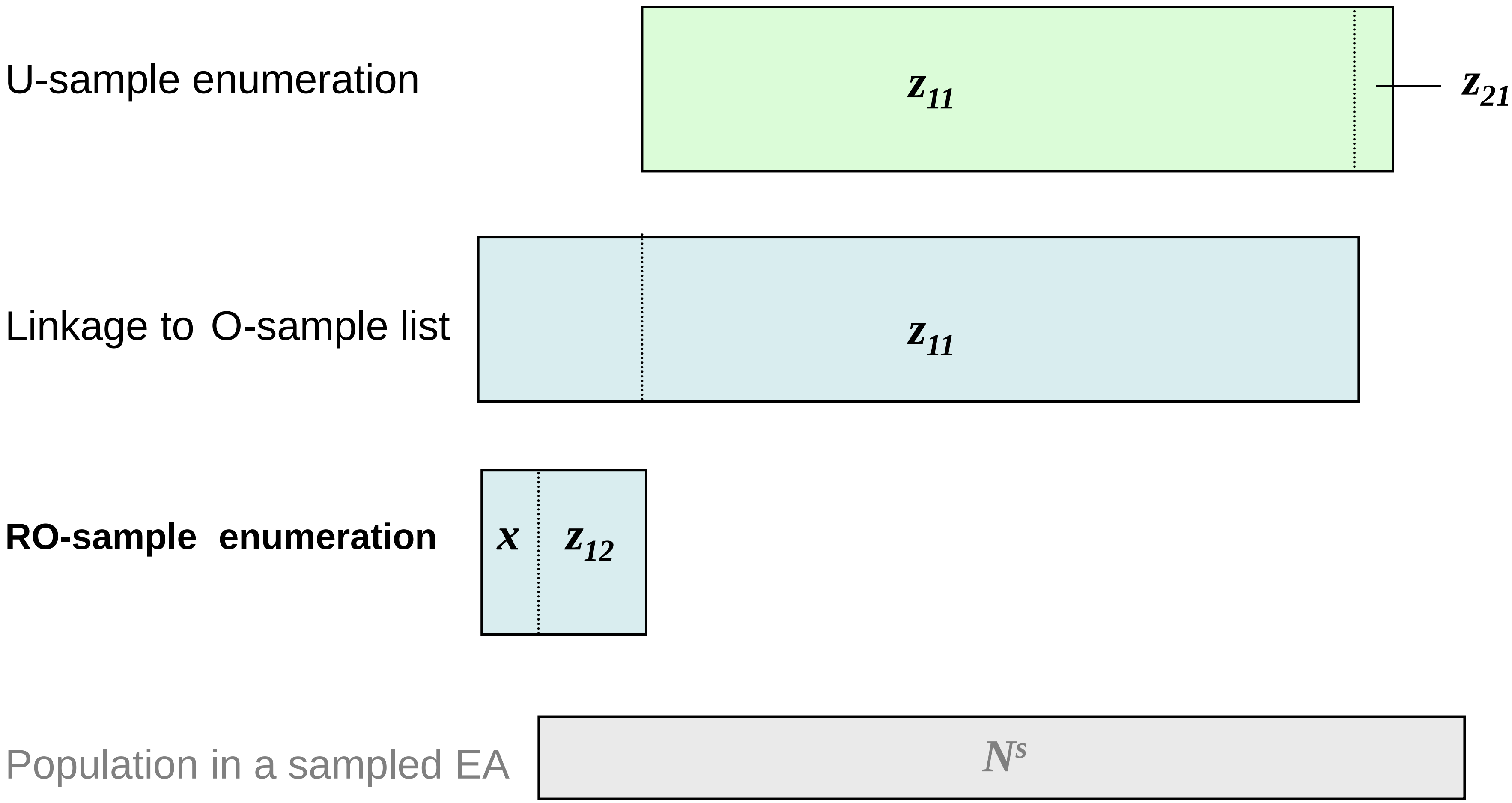
Sample allocation. Based on model and sampling variance

Implementation



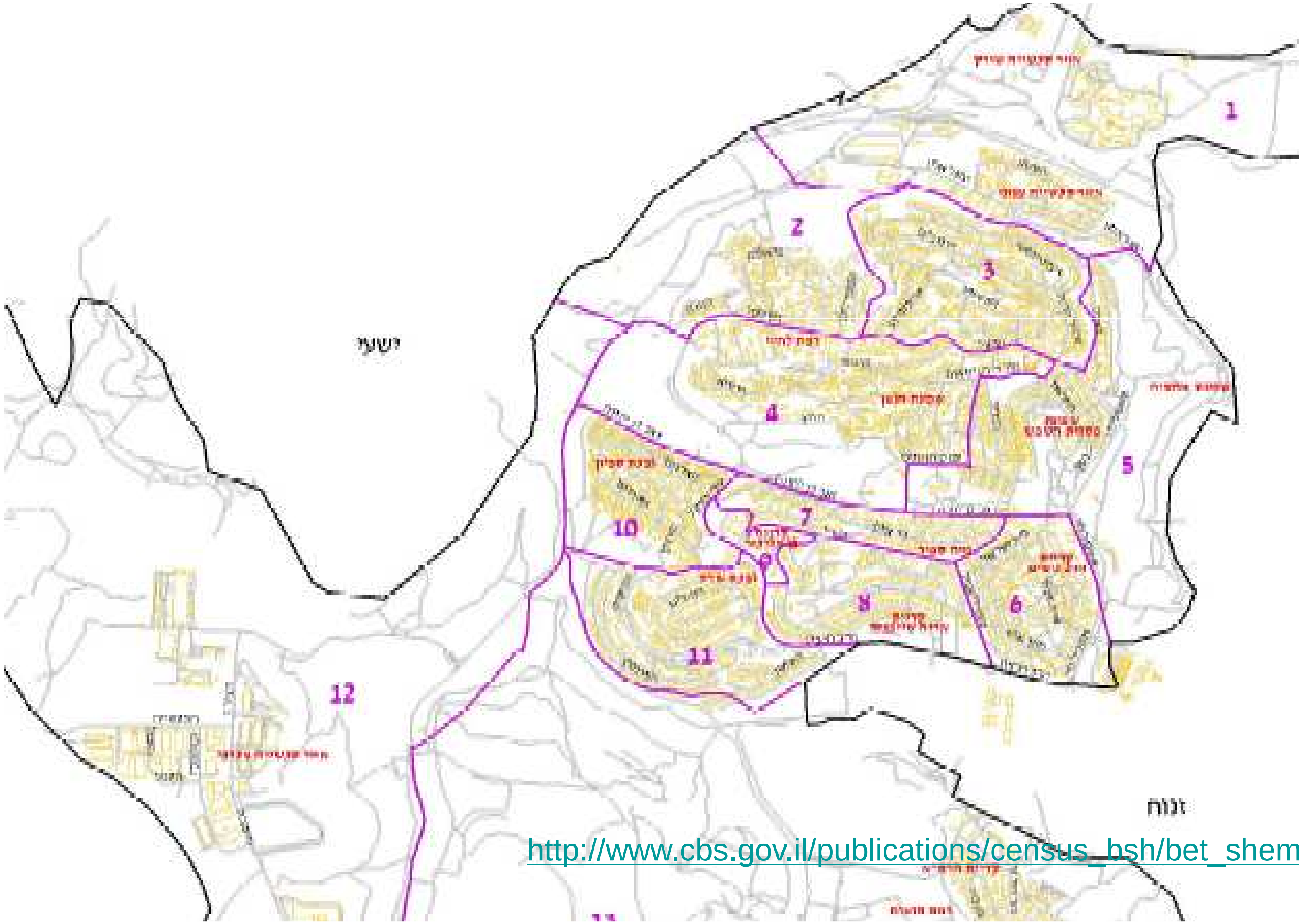
Fieldwork Process

For an EA in the sample:



The first Israeli experiment

(Beit Shemesh, 2002)



http://www.cbs.gov.il/publications/census_bsh/bet_shemesh.pdf

Census weights: N(IC)/N(PR)

	Age Groups			
SA	0-19	20-29	30-39	40+
2	0.97	0.89	0.86	0.90
3	0.91	0.98	0.84	0.85
4	0.90	1.00	0.73	0.96
5	1.18	1.26	1.17	1.02
6	0.91	0.88	0.94	0.84
7	0.85	0.77	0.85	0.83
8	1.03	1.21	0.97	0.98
9	0.96	1.02	0.92	0.90
10	0.98	1.09	0.98	1.04
11	0.99	0.85	0.89	0.83
13A	1.09	0.94	1.00	1.01
13B	1.05	1.00	0.95	0.88

http://www.cbs.gov.il/publications/census_bsh/pdf/text03.pdf

Rolling samples *(Kish, e.g. 1990)*

Jointly select a set of k mutually exclusive (not overlapping) and *representative* periodic samples

Each with a sampling fraction $f = 1/F$

One sample is interviewed at each time period

Accumulation of k periods yields a sample with a fraction k/F

A rolling census

For $k=F$ the entire population is covered over F time periods (e.g., years)

Advantages

- Information on temporal variation: frequent (annual) estimates for national and large domains levels; less frequent estimates for smaller domains
- Uniform expenditure over time

Drawbacks

- Limited information on spatial and demographic variation: no “snapshot” data
- Higher risk of bias: harder to estimate coverage errors

Variants

Cumulated representative sample (CRS) design

Same PSUs in all k samples, with a rolling sample *within* a PSU – inclusion of main units in all samples

Panel design

Overlap between subsequent samples – more efficient for comparisons

French example (since 2004)

Rolling "census" – coverage of k/F=5/7 of the population over five years

A two-stage annual sample:

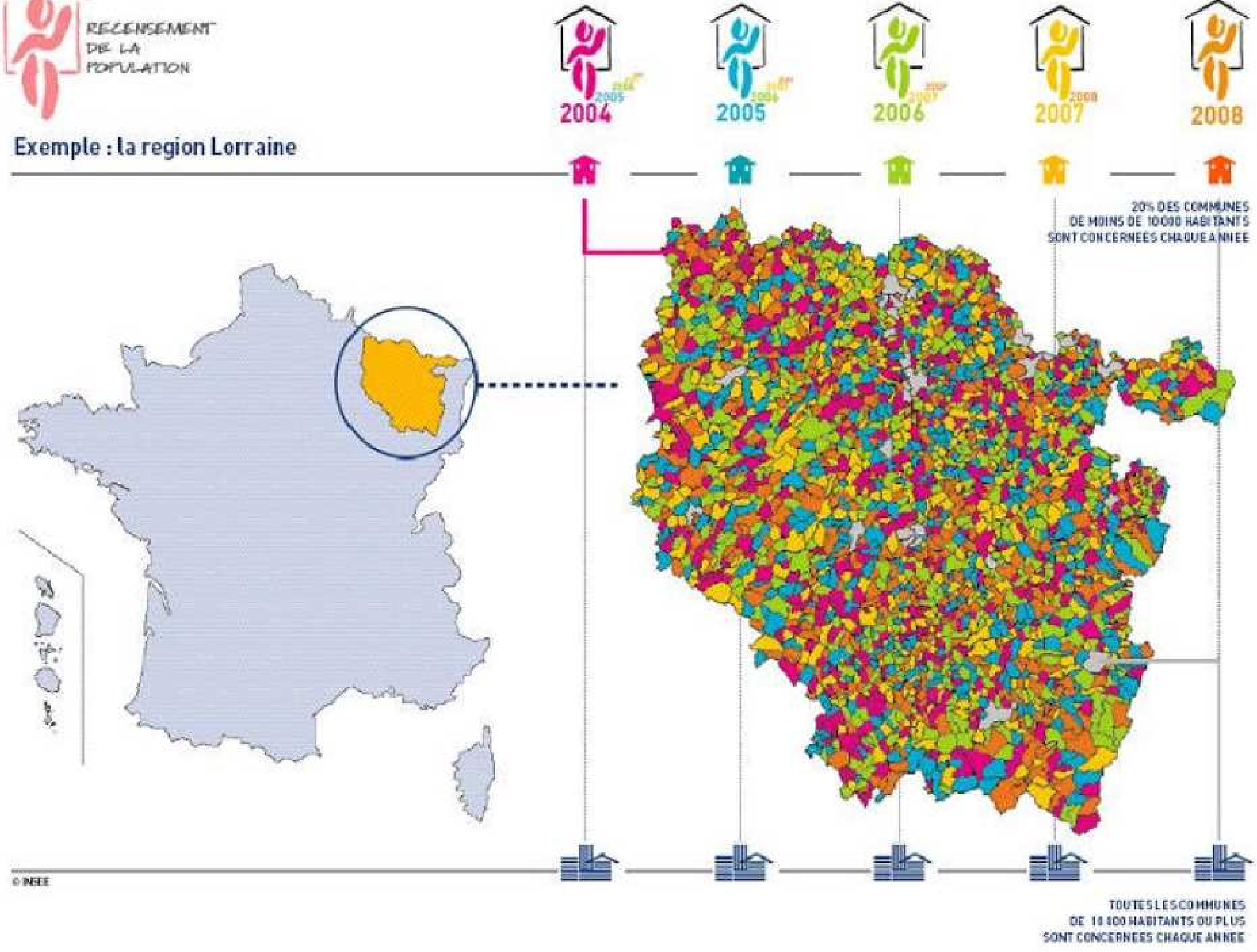
Large communes stratum – CRS design - all communes are included in the sample, and 0.08 of the dwellings are selected for enumeration.

Small communes stratum – Rolling sample - approximately 0.20 of the communes are sampled, and all dwellings in the sampled communes are included in the sample.

$$5(0.5 \cdot 0.08 + 0.5 \cdot 0.2) = 5 \cdot 0.14 = 0.70$$

RECENSEMENT DE LA POPULATION

Exemple : la region Lorraine



Estimation – general (Kish)

Let Y be the outcome of interest

$\hat{Y}_i \quad i = 1, \ldots, F$ the annual estimator

$$\hat{Y}(W) = \sum_{i=1}^{F} W_i \hat{Y}_i$$

$$W = (W_1, \ldots, W_F), \sum_{i=1}^{F} W_i = 1$$

Examples:

$$W_F = 1$$

$$W_i = 1/F$$

$$W_1 \leq \ldots \leq W_F$$

Population estimates for year F-2

For a large commune, let X_i be an auxiliary variable (number of dwellings) during year i,

and $\bar{X} = \sum_{i=F-4}^{F} X_i / 5.$

Let $\hat{Y}_i$ be the expansion count estimator for year i,

$$\bar{\hat{Y}} = \sum_{i=F-4}^{F} \hat{Y}_i / 5.$$

The estimate for year F-2 is the ratio (synthetic) estimate

$$\hat{Y}_{F-2} = \bar{\hat{Y}} \frac{X_{F-2}}{\bar{X}}$$

Population estimates for year F-2 (2)

For a small commune, the estimator depends on the year it was enumerated. Denote by Y_i the population count of a commune that is fully enumerated in year i.

$$\hat{Y}_{F-2}^{F-4} = \hat{Y}_{F-4} \frac{X_{F-2}}{X_{F-4}} \qquad \hat{Y}_{F-2}^{F-3} = \hat{Y}_{F-3} \frac{X_{F-2}}{X_{F-3}} \qquad \hat{Y}_{F-2}^{F-2} = \hat{Y}_{F-2}$$

$$\hat{Y}_{F-2}^{F-1} = \alpha_{F-1} Y_{F-1} + (1 - \alpha_{F-1}) Y_{F-6} \frac{X_{F-2}}{X_{F-6}} \quad \text{and}$$

$$\hat{Y}_{F-2}^{F} = \alpha_F Y_F + (1 - \alpha_F) Y_{F-5} \frac{X_{F-2}}{X_{F-5}},$$

where $0 \leq \alpha_i \leq 1$ $i = F - 1, F$, and is typically no smaller than 0.5.

Estimates for current year

Each annual survey is a representative sample comprising about eight million people.

Hence, usual survey methods (e.g., expansion estimates) enable reliable national and regional estimates for the current year. These estimates are used to calibrate the commune estimates.

תרגיל

תכנון מדגם תאי פקידה עבור מפקד משולב מבוסס על אומדים לשונות המוצגים בשקף 16. בתהליך התכנון "מנחשים" ערכים של הפרמטרים הלא ידועים כדי לחשב את גודל המדגם הדרוש עבור רמת טעות מסוימת.

א. רשמו את גודל מדגם התאים הדרוש m כפונקציה של הטעות היחסית של האומד $a = \sqrt{\mathrm{Var}(\hat{N})} / N$ עבור המקרה $p_{1+} = p_{+1} = p$ $\lambda = 0$

ב. ניתוח רגישות. עבור התרחיש בסעיף א., ישוב בגודל משוער N=50000 אנשים, M=300 תאי פקידה, וטעות יחסית a=0.01 ציירו גרף של גודל המדגם הדרוש m כפונקציה של p, עבור ערכי p בתחום (0.5,1] בקפיצות של 0.05.

ג. מהי מסקנתכם לגבי רגישות התכנון להערכות מוקדמות של p?