European Health Interview Survey in Latvia – Challenges and Opportunities

Mārtiņš Liberts

Central Statistical Bureau of Latvia

25 August 2015

European Health Interview Survey

- Reliable data on:
  - health status
  - health care
  - health determinants
- from all EU member states
- on regular bases

The First Wave of EHIS

- Gentlemen's agreement
- 17 EU member states:
  - Austria, Belgium, Bulgaria, Cyprus, Czech republic, Estonia, France, Germany, Greece, Hungary, Latvia, Malta, Poland, Romania, Slovakia, Slovenia, Spain
- 2006–2009

The Second Wave of EHIS

- Regulation (EC) No **1338/2008** of the European Parliament and of the Council of 16 December 2008 on Community statistics on public health and health and safety at work
- **every five years** statistics on health status, access and use of healthcare and health determinants
- Commission Regulation (EU) No **141/2013** of 19 February 2013 implementing Regulation (EC) No **1338/2008**
- European Health Interview Survey Manual

EHIS in Latvia

- The first wave was done in 2008
- The second wave was done in 2014

Population Frame

- The population frame was made as a list of individuals
- Data sources:
  - Statistical Dwelling Register – base for the frame (list of individuals)
  - Population statistics – for over-coverage reduction
  - National Health Service – indication if individual has used national health services
  - Population Census 2011 – additional contact information

Precision Requirements

- Precision requirements are defined with minimum effective sample size for estimating the population parameter "percentage of people severely limited in daily activities"
- it is 4555 for Latvia (defined for each country in the implementation regulation)
- sample size should be at least 4555 if simple random sampling of individuals is used (full response and frame without over–coverage errors).

Precision Requirements

- Minimum effective sample size can be expressed in terms of the coefficient of variation (CV)
- Input:
  - Minimum effective sample size: $n_e = 4555$ (defined by the regulation)
  - Population size: $N = 1711928$ (residents of private households in age 15+) computed from the population frame
  - Parameter of interest: $\hat{p} = 0.106$ estimated from The European Health and Social Integration Survey (EHSIS) 2012

Precision Requirements

- Expected population variation:

$$s^2 = \frac{N}{N-1}\hat{p}\left(1-\hat{p}\right) = 0.0950332$$

- Corresponding variation, standard error and CV:

$$\operatorname{var}\left(\hat{p}\right) = \frac{1-\frac{n_e}{N}}{n_e}s^2 = 0.0000208$$

$$\operatorname{se}\left(\hat{p}\right) = \sqrt{\operatorname{var}\left(\hat{p}\right)} = 0.00456$$

$$\operatorname{cv}\left(\hat{p}\right) = \frac{\operatorname{se}\left(\hat{p}\right)}{\hat{p}} = 0.043 = 4.3\%$$

CAPI / CATI

- The contact phone number was known for 57% of individuals in the population frame
- The population frame was divided into two groups (master strata):
  - CAPI – population part surveyed by personal interviews
  - CATI – population part surveyed by phone

Sampling design

- Sample size: 11 340
- Different sampling designs were chosen:
  - CAPI – two stage sampling to balance the cost and precision (sample size 6252)
  - CATI – one stage sampling to increase precision (sample size 5088)

Sampling design – CAPI

- The 1st stage:
  - Stratification by the degree of urbanisation (four strata)
  - Systematic πps sampling of census counting areas
  - Ordering by geographical location in each stratum
  - Sample size: 1042 PSUs
- The 2nd stage:
  - Sample size: six persons in each PSU
  - Systematic sampling of individuals in each sampled PSU
  - Ordering by NHS, gender, age, random number

Sampling design – CATI

- The 1st stage:
  - Stratification by NHS and seven age groups (14 strata)
  - Systematic sampling of individuals
  - Ordering by region and gender in each stratum

Expected precision

- Expected precision for the estimate of population parameter:
  - population size (population frame),
  - population variation (EHSIS data),
  - response rate (EHSIS data)
  - design effect (EHSIS data: deff $= 2$ for CAPI, deff $= 1$ for CATI)
  - sample allocation

$$\underset{\text{exp}}{\text{Var}}\left(\hat{p}\right) = \sum_{h=1}^{H} N_h^2 \frac{1 - \frac{n_h r_h}{N_h}}{n_h r_h} s_h^2 \underset{h}{\text{deff}}\left(\hat{p}\right)$$

- It was 3.6% (precision requirements were 4.3%)

Weighting

- There are two sources for auxiliary information:
  - Population frame – many auxiliary variables
  - Population statistics – more precise population counts
- The aim was to use both available sources of auxiliary information in weighting of EHIS

Weighting

- Auxiliary information – population counts by:
  - **Gender**
  - **Age groups**
  - **Region**
  - Education level
  - Usage of public health services
  - Household size
  - Economic activity (employed, unemployed, ...)
  - Degree of urbanisation (DEGURBA)

Two step weighting

- The 1st step:
  - The population frame was calibrated to the population statistics
- The 2nd step:
  - Additional calibration variables were introduced
  - Calibration totals were computed as weighted sums of calibration variables from the population frame

Two step weighting

- Auxiliary vector $\boldsymbol{x}_i = \left(\boldsymbol{x}_i^A, \boldsymbol{x}_i^B\right)'$
- Frame totals:
  - $\boldsymbol{X}^A = \sum_U \boldsymbol{x}_i^A$
  - $\boldsymbol{X}^B = \sum_U \boldsymbol{x}_i^B$
- Additional totals $\tilde{\boldsymbol{X}}^A$ are available
- Totals $\tilde{\boldsymbol{X}}^A$ are more precise if compared to $\boldsymbol{X}^A$

Frame calibration

- Frame calibration to known population totals $\tilde{\boldsymbol{X}}^A$
- Input for calibration:
  - Design weights: $d_i^F = 1$
  - Calibration variables: $\boldsymbol{x}_i^A$
  - Totals: $\tilde{\boldsymbol{X}}^A$
- Result of frame calibration:
  - Frame calibration weights g_i^F
  - Properties: $\sum_U \boldsymbol{x}_i^A d_i^F g_i^F = \sum_U \boldsymbol{x}_i^A g_i^F = \tilde{\boldsymbol{X}}^A$
- $\tilde{\boldsymbol{X}}^B = \sum_U \boldsymbol{x}_i^B g_i^F$

Calibration of respondents

- Respondents can be linked to the population frame through the ID number
- Input for calibration:
  - Design weights with non-response correction: $d_i^R = \frac{1}{\pi_i p_i^R}$
  - Calibration variables: $\boldsymbol{x}_i = \left(\boldsymbol{x}_i^A, \boldsymbol{x}_i^B\right)'$
  - Totals: $\tilde{\boldsymbol{X}} = \left(\tilde{\boldsymbol{X}}^A, \tilde{\boldsymbol{X}}^B\right)' = \sum_U \boldsymbol{x}_i g_i^F$
- Result of calibration:
  - Calibration weights g_i^R
  - Properties: $\sum_{s_R} \boldsymbol{x}_i d_i^R g_i^R = \tilde{\boldsymbol{X}}$

Design Effect

- Two components:
  - Sampling effect:

$$\underset{\text{sam}}{\text{deff}}\left(\hat{\Theta}\right) = \frac{\text{Var}\left(\hat{\Theta}\middle|\text{current, HT}\right)}{\text{Var}\left(\hat{\Theta}\middle|\text{SRS, HT}\right)}$$

  - Sampling effect:

$$\underset{\text{est}}{\text{deff}}\left(\hat{\Theta}\right) = \frac{\text{Var}\left(\hat{\Theta}\middle|\text{current, cal}\right)}{\text{Var}\left(\hat{\Theta}\middle|\text{current, HT}\right)}$$

Design Effect

- Design effect:

$$\begin{aligned}
\operatorname{deff}\left(\hat{\Theta}\right) &= \underset{\text{sam}}{\operatorname{deff}}\left(\hat{\Theta}\right) \cdot \underset{\text{est}}{\operatorname{deff}}\left(\hat{\Theta}\right) \\
&= \frac{\operatorname{Var}\left(\hat{\Theta}\middle|\text{current, HT}\right)}{\operatorname{Var}\left(\hat{\Theta}\middle|\text{SRS, HT}\right)} \cdot \frac{\operatorname{Var}\left(\hat{\Theta}\middle|\text{current, cal}\right)}{\operatorname{Var}\left(\hat{\Theta}\middle|\text{current, HT}\right)} \\
&= \frac{\operatorname{Var}\left(\hat{\Theta}\middle|\text{current, cal}\right)}{\operatorname{Var}\left(\hat{\Theta}\middle|\text{SRS, HT}\right)}
\end{aligned}$$

Results

- The main population parameter “percentage of people severely limited in daily activities” was estimated as 0.104 (it was 0.106 in EHSIS 2012)
- Coefficient of variation for the estimate is 3.0 % (calibration effect is taken into account)
- The design effect is 0.717 where:
  - The sampling effect is 0.817
  - The estimation effect is 0.877

Results

Total sample size is 11340:

- Over-coverage cases: 378 (3%)
- Eligible cases: 10962 (97%)
  - Non-response cases: 3885 (35%)
  - Response cases: 7077 (65%)
    - Item non-response cases: 8
    - Item response cases: 7069

Effective sample size is 9858 (requirement is 4555)

$$n_{\text{eff}} = \frac{n_{\text{resp}}}{\text{deff}\,(\Theta)}$$

Concludions

- Combination of one stage and two stage sampling designs:
  - Improvement of precision – the sampling effect is 0.817
  - Increase in interview travelling costs (820 cases were transferred from CATI to CAPI)

Concludions

- Combination of one stage and two stage sampling designs:
  - Improvement of precision – the sampling effect is 0.817
  - Increase in interview travelling costs (820 cases were transferred from CATI to CAPI)
- Two step weighting:
  - All available auxiliary information (individual level and population level) has been used
  - The estimation effect is 0.877

Concludions

- Combination of one stage and two stage sampling designs:
  - Improvement of precision – the sampling effect is 0.817
  - Increase in interview travelling costs (820 cases were transferred from CATI to CAPI)
- Two step weighting:
  - All available auxiliary information (individual level and population level) has been used
  - The estimation effect is 0.877
- The precision requirements are fulfilled (effective sample size is 9858)

Thank you!

Data Collection Modes

Mode	NA	CAPI	CATI	CAWI
CAPI	46	5679	163	406
CATI	0	820	3743	483