UNEQUAL PROBABILITY SAMPLING FOR THE EUROPEAN INTERVIEW HEALTH SURVEY IN LATVIA

M. Liberts[1]

[1] Central Statistical Bureau of Latvia
e-mail: martins.liberts@csp.gov.lv

Abstract

A common requirement for a large scale sample survey is to deliver sufficiently precise estimates at population and domain level. Often study domains are with unequal size. Some of study domains can be much smaller then others. Those are contradicting requirements regarding the choice of an optimal sampling design to fulfil those requirements. One of the examples is the latest European Health Interview Survey which has been done in Latvia during 2019/2020. There are quality requirements at population level defined by the corresponding regulation. At the same time there are national requirements defined at domain level. An experimental sampling design with unequal sampling probabilities was proposed and implemented to fulfil those requirements.

Keywords: Unequal probability sampling, European Health Interview Survey.

1 European Health Interview Survey

The European Health Interview Survey 2019 (EHIS-2019) was organised in European Statistical System according to the Commission Regulation (EU) 2018/255 of 19 February 2018 (European Commission, Eurostat, 2018). The implementation of the survey is guided by the Methodological manual (European Union, Eurostat, 2020).

1.1 Precision Requirements of Eurostat

The Annex II of the regulation (European Commission, Eurostat, 2018) defines the precision requirements. This is a citation from the regulation:

1. Precision requirements for all data sets are expressed in standard errors and are defined as continuous functions of the actual estimates and of the size of the statistical population in a country.

2. The estimated standard error of a particular estimate $\hat{SE}(\hat{p})$ shall not be bigger than the following amount:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{f(N)}}$$

3. The function $f(N)$ shall have the form of $f(N) = a\sqrt{N} + b$.

4. The following values for parameters N, a and b shall be used:

- $\hat{p}$: Percentage of population severely limited in usual activities because of health problems (age 15 years or over).
- N: Country population aged 15 years or over residing in private households, in million persons and rounded to 3 decimal digits.
- a: 1200
- b: 2800

1.2 National Survey and Precision Requirements

National survey and precision requirements were defined in the following form:

- Sample size: 11,000 persons.
- Two-stage sampling with geographical clustering should be used to optimise fieldwork cost where two main cost components are travelling expenses and time required for fieldwork operation.
- The main variables of interest:
  - General health self-assessment.
  - Health problems limiting activities.
  - Financial obstacles for receiving health care services.
  - Height and weight.
- The main population domains of interest:
  - Gender split by age groups (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+).
  - Economic activity (employed, unemployed, economically inactive).
  - Education (according ISCED: 0-1, 2, 3-4, 5-8).
  - Household net monthly equalised income quintile groups (five groups).
  - Region (NUTS-3, six regions).

2 Methodology

2.1 Sampling Frame

The sampling frame was created in three sequential steps.

The statistical register of dwellings and persons is a monthly updated statistical register maintained by the Central Statistical Bureau of Latvia. It contains information about all registered persons and inhabited dwellings in Latvia. The administrative data sources are used for updating the register. The main administrative data sources are the Population Register and the Address Register.

The general sampling frame of all registered residents living in private dwellings is a general sampling frame which can be used as an initial data source for any sampling frame for a sample survey where persons are sampled. The general frame is created with a standardized procedure. It is created whenever a sampling frame is necessary for any sample survey where persons are sampled. The data sources for creating the general sampling frame are the statistical register of dwellings and persons, the Address Register, phone lists, samples of other previous surveys, and other data sources.

The EHIS-2019 sampling frame is a frame which was used for sampling of persons for the EHIS-2019. It was created specifically for the needs of the EHIS-2019. For example,

population coverage was reduced to persons aged 15+, persons likely to be *de-facto* non-residents were excluded, extra variables for the needs of the EHIS-2019 were added. The data sources for creating the EHIS-2019 sampling frame were:

- the general sampling frame of all registered residents living in private dwellings (2019-06-27),
- micro data of the population statistics (2016 – 2019),
- administrative data about persons who have received state medical services in 2018,
- yearly income of persons in 2017,
- administrative data about self-employed persons in the first quarter of 2019,
- administrative data about employees and employers in May 2019,
- administrative data about registered unemployed persons in the first two quarters of 2019,
- the highest attained education level on 2019-01-01 from the Population Census data base.

2.2 Sample Size and Sampling Design

The total sample size for the EHIS-2019 was fixed to 11,000 persons. It was required to use two-stage sampling with geographical clustering of sampled persons to optimise fieldwork cost.

2.2.1 Sampling of Survey Areas

There were survey areas (*iecirkņi* in Latvian) available for sampling. The survey areas were created as compact geographical clusters of inhabited private dwellings with a purpose to be used for sample surveys where geographical clustering of sample units is necessary. The size of survey areas is measured by the number of inhabited private dwellings. The survey areas were created with similar size in urban and rural territories (300 for urban and 150 for rural territories). The survey areas were redesigned in 2019. The EHIS-2019 was the first survey to use the 'new' survey areas.

The survey areas were used as the first stage sample units. So, stratification for the first stage sample could be done using geographical information only. Stratification of persons (and sample areas) was done according to the declared living place of persons. There were five strata:

1. Persons with declared living place in Riga (the capital of Latvia).
2. Persons with declared living place in cities under state jurisdiction (eight cities excluding Riga).
3. Persons with declared living place in towns.
4. Persons with declared living place in parishes (rural areas).
5. Persons with cancelled or erroneous declared living place address (for those persons phone number was available to make the first contact over phone or to do data collection over phone using computer assisted telephone interview approach).

Sample allocation by strata was calculated according to the Neyman allocation (Neyman, 1934), where standard deviation was calculated according to the binary variable describing if person had received the state funded medical services in 2018. The sample size of the first four strata was rounded to the closest multiple of six (sample size of persons at the second stage for each PSU). The sample size for the 5th strata was calculated as a reminder (11,000 minus the total sample size of the strata 1–4). Sample allocation is available in Table 1 where:

- **strata**: strata identification
- **N**: frame population size
- **P**: proportion of frame persons who have received the state funded medical services in 2018
- **n_prop**: proportional sample allocation (only as a reference)
- **n_neim**: Neyman optimal sample allocation calculated using **N** and **S**
- **n_SSU**: Neyman optimal sample allocation rounded to the closest multiple of 6 (for the strata 1–4) and sample size in strata 5 is a reminder to the total sample size
- **n_PSU**: number of sampled PSUs
- **f**: sampling fraction

Table 1: Sample allocation by strata

strata	N	P	S	n_prop	n_neim	n_SSU	n_PSU	f
1	523 724	0.753	0.431	3 596	3 738	3738	623	0.007137
2	301 081	0.808	0.394	2 067	1 963	1962	327	0.006517
3	262 583	0.810	0.392	1 803	1 706	1704	284	0.006489
4	505 062	0.771	0.420	3 467	3 512	3516	586	0.006962
5	9 791	0.466	0.499	67	81	80	80	0.008171
Total	**1 602 241**			**11 000**	**11 000**	**11 000**	**1 900**	**0.006865**

Two-stage sampling was used for the first four strata. Single stage sampling was used for the 5th stratum (geographical clustering was not possible for persons with unknown declared living place).

The mentioned survey areas (clusters of persons) were used as the primary sampling units for strata 1-4. Survey areas were sampled in each stratum with systematic sampling with probabilities proportional to area size. The area size was calculated as number of persons available for sampling (there is a negative coordination with samples of other surveys with an aim to reduce the burden of respondents) associated to a respective area. Survey areas have been ordered in each stratum geographically so that contiguous areas in the survey area frame are also geographically close in space.

2.2.2 Sampling of Persons

The secondary sampling units in strata 1–4 and the primary sampling units in stratum 5 were persons. There were six persons sampled in each sampled area for strata 1–4. Sample size for the 5th stratum was 80 persons.

One of the survey requirements for national needs was to optimise the survey to produce reliable survey estimates for several population domains of interest. Those domains were defined as:

- gender and age groups (14 domains),
- economic activity status (3 domains),
- household income (5 domains),
- NUTS-3 regions (6 domains),

- highest achieved education level (4 domains).

It was possible to create those domains in the population frame according to the available external data sources (administrative data were used in most cases; exception is education level were different data sources were used including administrative, sample survey and the last census 2011 data).

Obviously the correspondence of those frame (*de jure*) domains with real (*de facto*) domains differ. For example, gender and age group domains in frame are almost 100 % equal to the real gender and age group domains. There is some level of misclassification errors for all other domains. However, it was assumed that those frame domains are good auxiliary information representing the main domains of interest. So, this information could be used to improve the precision of survey estimates in those target domains.

The population size for those domains differ quite a lot. For example the largest of those domains is persons with secondary education (corresponding to the ISCED 3 or ISCED 4). The size of this domain in the frame was 825,022 making 0.515 share of all frame persons. On the opposite side the smallest domain was unemployed persons. The size of this domain was 41,267 making only 0.026 share of all frame persons.

Assume we are using equal probability sampling. This approach would provide estimates with acceptable precision for large domains. For example, the expected sample size for persons with secondary education would be $11,000 \cdot 0.515 = 5664$, which should be enough to provide estimates with acceptable precision. However, the expected sample size for unemployed persons would be only $11,000 \cdot 0.026 = 283$. Taking non-response and over-coverage into account the expected net-sample size could be close to 155 which would not be enough to provide estimates with acceptable precision.

It would be necessary to over-sample small size domains while large size domains should be under-sampled to keep the total sample size fixed. Such approach would allow to improve the expected precision for estimates in small size domains.

Assume full response and equal variance in all domains, namely

$$S_d^2 = \frac{1}{N_d - 1} \sum_{i \in U_d} (y_i - \bar{y}_d)^2 = S, \text{ for } \forall d,$$

where:

- U_d is subset of target population belonging to domain d, $U_d \subset U$ where U is a set of units belonging to the target population and the size of U is constant.
- N_d is domain d population size,
- y_i is a value of a study variable for a population unit i,
- $\bar{y}_d = \frac{1}{N_d} \sum_{i \in U_d} y_i$.

Assume D non-overlapping domains covering U completely:

$$U_k \cap U_j = \emptyset \text{ for } \forall k, j$$

and

$$\bigcup_{d=1}^{D} U_d = U.$$

The optimal sample allocation in this case would be equal sized sample allocation by domains, namely $n_d = \frac{1}{D} n$, where n is the total sample size. Hence the optimal sampling probabilities would be

$$\pi_{i|d} = \frac{n}{DN_d},$$

where $\pi_{i|d}$ is a sampling probability for a population unit i under assumption $i \in U_d$. Those sampling probabilities provide equal sample size in all domains:

$$\sum_{i \in U_d} \pi_{i|d} = \sum_{i \in U_d} \frac{n}{DN_d} = \frac{n}{D} \text{ for } \forall d.$$

There are 32 target domains which are overlapping. So, this approach cannot be used directly. Those 32 domains can be ordered in five sets of domains where domains from one set are non-overlapping and covering U completely. Those domain sets are:

1. gender and age groups (14 domains),
2. economic activity status (3 domains),
3. household income (5 domains),
4. NUTS-3 regions (6 domains),
5. highest achieved education level (4 domains).

The exception are domains by education which do not cover frame population completely. There are persons with unknown education in a frame. Such domain exists only in a frame (because of missing information). However, such domain does not exist in a target population.

For each of those domain sets an optimal sampling probabilities were calculated as

$$\pi_{i|d_g} = \frac{n}{D_g N_{d_g}},$$

where d_g is a domain d from the domain set g, D_g is a number of domains in a domain set g, $\pi_{i|d_g}$ is a sampling probability for a person i under assumption $i \in U_{d_g}$.

Five sampling probabilities were calculated for each frame person according to each of five domain sets. Obviously those five sampling probabilities differ, so the final sampling probability for each frame person was calculated as an average of those five sampling probabilities:

$$\pi_i = \frac{1}{5} \sum_{g=1}^{5} p_{i|d_g}.$$

Those sampling probabilities π_i would be possible to use directly for a single stage sampling. In case of two stage sampling π_i cannot be used directly for sampling. But π_i were used as a "size measure" for persons to calculate second stage sampling probabilities proportional to π_i. For example:

- The lowest "size measure" was for an employed woman aged 55-64 living in Riga with the secondary education (ISCED 3–4). Those persons represent large size domains, so we want to sample those persons with low sampling probability.
- The highest "size measure" was for an unemployed woman aged 15-24 living in the Vidzeme region without primary education (ISCED 0–1). Those persons represent small size domains, so we want to sample those persons with high sampling probability.

2.2.3 Sampling Algorithm

The systematic sampling method (unequal probabilities, without replacement, fixed sample size) was used for both sampling stages. The function `UPsystematic` from the R (R Core Team, 2021) package `sampling` (Tillé & Matei, 2021) was used to implement the sampling method.

2.3 Weighting

The survey weights were calculated in three steps:

- Design weights.
- Non-response adjustment.
- Calibration of weights.

Design weights were calculated as inverse of the corresponding sampling probabilities.

Response probabilities were estimated using a logit model. Response probabilities were estimated only for the eligible persons (non-eligible persons were excluded from the estimation, persons with unknown eligibility status were assumed to be eligible). Model regressors were constructed using variables as:

- Type of territory (according to the declared living place): Riga (the capital city), cities under state jurisdiction (excluding Riga), towns, parishes,
- NUTS-3 region (according to the declared living place),
- Sex,
- Age group (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+),
- The highest achieved education level (four groups according to the ISCED 2011: 0-1, 2, 3-4, 5-8),
- Usage of the state funded medical services during the year 2018,
- Equalised yearly household income (2017, five quintile groups),
- Economic activity status (employed, unemployed, inactive).

There were 38 variables included in the response logit model. Non-response adjusted weight for each respondent was computed as design weight divided by the estimated response probability.

Only respondents were used in the weight calibration. Calibration variables were constructed using variables as:

- Type of territory (according to the declared living place): Riga (the capital city), cities under state jurisdiction (excluding Riga), towns and parishes,
- NUTS-3 region (according to the declared living place),
- Sex,
- Age group (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+),
- The highest achieved education level (three groups according to the ISCED 2011: 0-2, 3-4, 5-8).

There were 52 variables included in the calibration equation. Linear calibration from the R (R Core Team, 2021) package `surveyweighting` (Breidaks, 2020) was used.

Table 2: Survey outcome

No	Name	Non-weighted	Weighted
0	Sample size	11,000	1,357,679
1	Respondents	6,033	741,635
2	Non-respondents	4,636	574,336
3	Non-eligible	331	41,708
4	Over-coverage rate ([3] / [0])	0.030	0.031
5	Response rate ([1] / [1] + [2])	0.565	0.564

3 Results

3.1 Statistics

The statistics produced using the EHIS-2019 data are available at the Official Statistics Portal of Latvia (`https://stat.gov.lv/en`).

3.2 Non-sampling Errors

The total design weighted response rate was 0.564 and the total design weighted over-coverage rate was 0.031. See more details in Table 2.

3.3 Sampling Errors

The sampling error estimates for the main population parameter estimates are provided in Table 3. The description of the main population parameters:

- HS1: Proportion of persons aged 15+ in good or very good health
- HS2: Proportion of persons aged 15+ with longstanding illness or health problem
- HS3: Proportion of persons aged 15+ that were severely limited in activities people usually do because of health problems for at least past 6 months (this is the parameter used to define the Eurostat precision requirement)
- HO1: Proportion of persons aged 15+ having been hospitalized in the past 12 months
- BMI: Proportion of persons aged 18+ who are obese (BMI equal or above 30, where BMI (body mass index) is calculated as weight in kg divided by height in meters squared)

The Eurostat precision requirements defined at the regulation (European Commission, Eurostat, 2018) were verified. The precision requirement was: *the standard error estimate for the estimated proportion of severely limited in usual activities because of health problems (age 15 years or over, HS3) shall not be bigger than*

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{f\,(N)}}$$

where:

- $f\,(N) = a\sqrt{N}+b$ is the minimum effective sample size necessary to fulfil the corresponding precision requirements.

Table 3: Estimates of the main parameters of interest with respective precision measures

variable	gender	respondents	estimate	SE	confidence interval	deff
HS1	All	5848	0.496	0.004216	0.487 — 0.504	0.418
HS1	Women	3420	0.454	0.005353	0.443 — 0.464	0.375
HS1	Men	2428	0.548	0.006679	0.535 — 0.561	0.475
HS2	All	6025	0.732	0.003984	0.725 — 0.740	0.483
HS2	Women	3495	0.775	0.004845	0.766 — 0.785	0.444
HS2	Men	2530	0.680	0.006468	0.667 — 0.692	0.515
HS3	All	6023	0.091	0.002581	0.086 — 0.096	0.501
HS3	Women	3492	0.106	0.003533	0.099 — 0.113	0.446
HS3	Men	2531	0.072	0.003516	0.065 — 0.079	0.524
HO1	All	6024	0.115	0.003046	0.109 — 0.121	0.558
HO1	Women	3495	0.123	0.004182	0.115 — 0.131	0.545
HO1	Men	2529	0.106	0.004582	0.097 — 0.115	0.616
BMI	All	5528	0.230	0.004014	0.222 — 0.238	0.523
BMI	Women	3265	0.257	0.005278	0.247 — 0.268	0.466
BMI	Men	2263	0.196	0.006105	0.184 — 0.208	0.603

- $\hat{p}$: Percentage of population severely limited in usual activities because of health problems (age 15 years or over). The estimated proportion was 0.091 for Latvia (see the line HS3 "All" in Table 3).
- N: Country population aged 15 years or over residing in private households, in million persons and rounded to 3 decimal digits. It was 1.585 million for Latvia.
- a: 1200
- b: 2800

The threshold value for the estimated standard error is equal to

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{f\,(N)}} = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{a\sqrt{N}+b}} = \sqrt{\frac{0.091\,(1-0.091)}{1200\cdot\sqrt{1.585}+2800}} = 0.004373.$$

We can see the estimated standard error for the respective population parameter estimate was 0.002581 (see the line HS3 "All" in Table 3) which is lower then the threshold value (0.004373). We can conclude that precision requirements defined by the regulation (European Commission, Eurostat, 2018) have been fulfilled for the EHIS-2019 in Latvia.

4 Conclusions

The paper presents an empirical work with an implementing of an unequal probability sampling for the European Interview Health Survey in Latvia. The aim of this approach was to oversample target domains with small population size. It was expected to provide an optimal sampling design to fulfil national and European precision requirements.

The overall precision requirements defined by the regulation (European Commission, Eurostat, 2018) have been satisfied. The precision of the estimates of other main population parameters are good in general.

The precision of domain estimates have not been derived yet (exception is gender). Hopefully some of those results will be available for presenting at the Summer School on Survey Statistics 2021.

Acknowledgement

Author has used some of the material which he has provided to Eurostat for the quality report of the European Interview Health Survey 2019. The EHIS-2019 quality report has not been published yet.

References

Breidaks, J. (2020). surveyweighting: Survey weighting [Computer software manual]. Retrieved from `https://github.com/CSBLatvia/surveyweighting` (R package version 0.7)

European Commission, Eurostat. (2018). *Commission regulation (EU) 2018/255.* Retrieved from `http://data.europa.eu/eli/reg/2018/255/oj`

European Union, Eurostat. (2020). *European health interview survey (EHIS wave 3) — methodological manual (re-edition 2020).* Retrieved from `https://ec.europa.eu/eurostat/` (DOI: 10.2785/135920)

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558–625. Retrieved from `http://www.jstor.org/stable/2342192`

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Tillé, Y., & Matei, A. (2021). sampling: Survey sampling [Computer software manual]. Retrieved from `https://cran.r-project.org/package=sampling` (R package version 2.9)