KU LEUVEN

FACULTY OF SCIENCE

Exam Assignment

Sampling Theory

Laura Juliana GUERRERO VELASQUEZ
laurajuliana.guerrerovelasquez@student.kuleuven.be
Daniel Gerardo GIL SANCHEZ
daniel.gilsanchez@student.kuleuven.be

Prof. Geert Molenberghs

Academic year 2017-2018

Exam Assignment

Introduction

The importance of collecting information to know and understand a population is of primary interest in medical, social, economic sciences and many others. It would be optimal if researchers could have information about the entire population of interest, but for several reasons as size, costs and logistics, it becomes an impossible task. Here comes the relevance of sampling theory, it allows to draw a representative sample of the population in order to infer the behavior of the entire population, when doing it appropriately.

In this paperwork, several aspects of sampling theory are discussed to give an overview of how it works and relevant details are shown when performing it. In the following, a discussion is made about responsive designs, a methodology that aims to solve problems of non-response while decreasing the cost of the survey and increasing their efficiency. Then, a random sample is selected from the Belgian Health Interview Survey (BHIS) sample frame and analyzed for a variable of interest, Body Mass Index (BMI). Finally, a discussion is made around changing a key feature in the sampling design of the BHIS and a review on how it handled missing data in previous years.

Responsive Designs

Survey designers encounter many problems during fieldwork, but one in particular draws special attention which involves the refusal of sample units to respond to the survey. Over time, rejection is becoming more common due to the lack of time of the respondents as well as trust issues related to survey research in terms of confidentiality when giving information about their opinions, habits or personal characteristics.

In most surveys, this non-response is not a general behavior that occurs randomly in the sample, but it could actually be identified in specific subgroups of the target population, e.g., high income households. The possibility to identify profiles in the population that are more likely to refuse a questionnaire, makes relevant to assess these non-response problems, otherwise particular sub-populations estimates will be systematically biased. For this reason, non-response rate is a quality indicator of the survey estimates, where higher values increase the chance of getting biased results. Hence, during fieldwork efforts are made to decrease non-response rate because on top of affecting final results, it directly impacts the costs and efficiency of the entire survey.

As a consequence, the need of having quality information, at the lowest cost possible while being efficient is the gold-rule in survey design. So a survey statistician should take into account all these aspects to be able to define optimal settings to collect information.

In this sense, a methodology called **Responsive design** is proposed to solve non-response issues while exploiting the growth of computer-assisted data collection devices to make the survey design an adaptive process. Two articles are reviewed to show how it is defined and how it works (Groves and Heeringa, 2006; Tourangeau et al., 2017). So, a general description of the method is presented, followed by a discussion of the pros and cons that represent this method along with suggestions for further work.

The general idea behind a responsive design is to make use of the information provided by computer-assisted data collection tools to reduce the effects of non-response, identifying initial failures on the design and fixed them in subsequent phases of the survey with the goal of improving the rates of response. As stated by Groves and Heeringa (2006), responsive survey designs consist in

(a) preidentify a set of design features potentially affecting costs and errors of survey estimates

(b) identify a set of indicators of the cost and error properties of those features and monitor those indicators in initial phases of data collection

(c) alter the features of the survey in subsequent phases based on cost-error trade-off decision rules and

(d) combine data from the separate design phases into a single estimator

A special feature of this kind of designs is that is organized about design phases. A design phase is described by specific and not variant characteristics of the process such as the sampling frame, mode of data collection, sample design, recruitment protocols and general fieldwork conditions. In this sense, once the first phase of the survey ends, the survey statistician can change aspects of the survey design to improve the non-response rate, the error measurement, the accuracy on the variable(s) of interest, costs and efficiency of the fieldwork.

A subsequent phase involves the proposed changes and it is monitored again to identify the relevance of new phases and further survey design changes. The proposed changes include more design features than merely the sample design, they can go from the mode of data collection, reduction of questions in the questionnaire, increasing on the number of contact moments to include or increase incentives.

Now, an important question is how to determine when a phase ends. A key notion for responsive designs is that each set of design features, e.g., same sample design, mode of collection, brings with it a maximum level of quality for a given cost. When the phase achieves stability of an estimate, it can be said that is ”fully matured” or that it has reached its phase capacity. One example of phase capacity is the stability of a key survey estimate as a function of callbacks that are made to acquire an interview. It is, of course, desirable to reach stability as early as possible, in the sense that when this is achieved, the evaluation can be conducted and new decisions can be me made in the following phases.

Once the fieldwork is complete, the data collected in all phases of the study are combined to construct estimates of the quantities of interest (Tourangeau et al., 2017). This latter part of the analysis is conducted with the objective of improving the quality of the estimates but, the fact that different conditions are used in subsequent phases, could also increase bias in the final estimates depending on what kind of changes were involved.

To tackle this problem, Groves and Heeringa (2006) suggest the use of a randomized experiment in the initial design phase, where several aspects of the design are taken into consideration and compared to decide which are the optimal settings to develop the survey in order to get maximal response at the lowest cost and/or highest efficiency. In this perspective, randomized experiments are preferred because it allows studying different alternatives on respondents with similar characteristics on survey variables. Therefore, the decision of which could be the best alternative to implement in the subsequent phase with a larger partition of the sample is controlled by possible confounding factors.

Here is where the use of computer-based data collection tools gain importance, since these devices allow to have access to objective information about the process of the survey in real time, often referred as `process data` or `paradata`. Examples of this kind of data could be the number of contacts made before an effective interview, the duration of the interviews, the time and the day it was done. In this context, Groves and Heeringa (2006) also suggested implementing in the collection tool, extra information provided by the interviewer about the fieldwork such as time of traveling and administrative activities, but also to include aspects to assess the grade of reluctance of the respondents, such as whether the householder delivered a negative statement about the survey request or whether the householder noted that it was a bad time to talk. These variables can be used later as a proxy of difficulty and costs involved in an effective interview.

In a broader context, paradata is a useful source of objective information to infer the propensity of a sample unit to respond. Hence, it can be used to develop a model that predicts the propensities for the respondents to accept an interview. As a result, respondents can be classified according to their predicted propensities to respond, where lower values will represent higher efforts in the contact and adaptation of the design features, according to the implied difficulty derived from the covariates considered in the model.

Analyzing process data in real time, allows researchers to obtain indicators regarding effort expended to obtain effective interviews. This can be noted as phase capacity, and the researcher could determine thresholds where further fieldwork efforts will not represent a substantial decrease in non-response rate. For example, a study can identify the number of maximal calls an interviewer should made to accomplish an effective contact, after that value it does not worth to continue expending resources.

Furthermore, as stated in Tourangeau et al. (2017), responsive designs could be used for different objectives such as to minimize variance, minimize non-response bias, maximizing response rate, reduce measurement error and reduces costs.

From a critical point of view, the theory behind sampling and survey design encounters many problems during fieldwork. The truth is that in practice many of the assumptions made by designs are hard to hold, and the responsive design seems as a good alternative to make informed decisions about how to solve many issues during the collection process.

The analysis of paradata is an interesting option that can represent significant improvements in the efficiency of the collection process, reducing at the same time the costs of a survey. However, changing design aspects based on this information should be done carefully and protocols should include clear guidelines on the possible changes and motivation behind them before the start of the study. In this sense, an ethical component of the definition of what should be a good practice of responsive design should be established in order to avoid researchers to make changes on the designs given the observed values in variables of interest that will manipulate them as convenient.

On the other hand, none of the articles present clear information on how the sampling design could be affected by this technique. Ideally, the sample should be drawn with sufficient replacements, where the propensities represent relevant approximations to the appropriate number of replacements in order to fulfill the sample with complete statistical rigor. In addition, the way second and third phase respondents are found should not represent additional bias or error in the estimates.

It is very clear that the main purpose of responsive designs is to deal with non-response in real time while fieldwork is still running. Both articles agree on the fact that many efforts have been made to develop responsive designs, but almost none to see how it affects the estimations. Tourangeau et al. (2017) states that there is a possibility that responsive design itself may introduce bias, hence further work should be done regarding this aspect. Nonetheless, it is understandable how is not yet quantifiable the accuracy effects of this technique, because every survey contemplates different scenarios according to its sampling and target population. In some, only incentives are changed while in others the mode of collection, the questionnaire and the incentives could change. In this sense, being able to measure how much bias or error is introduced would be different from one survey to another.

As it was described before, a special feature of responsive design is that is organized in phases. There are projects where the fieldwork does not last more than a week or even less, so the fact that the collection process is divided into phases seems unrealistic in such situations. It is very useful, though, in cases where the fieldwork lasts long enough to be able to divide it in phases. In this respect, the existence of process data in all surveys, no matter if the responsive design can be done or not, will always be a good idea because it brings new information about the fieldwork and could be used for further analysis.

Finally, many surveys have similar or identical target populations and faced similar fieldwork issues, specially in national statistics offices, which is why the use of previous survey process information should be used to design new studies. Using prior paradata can give a clear overview of what should be expected in the fieldwork, and what considerations should be taken into the sampling design and logistics of the survey before starting the collection process.

Sampling Selection from Belgian Health Interview Survey data set

A sample selection is made using a BHIS data set, where considerations are made to display key aspects of sampling theory. In general, sample can be selected in many different ways, so what is studied here is just an idea considering the data set as the population of interest.

First, the *sample size* is defined according to 15% of the total population. The *target population* are the 8564 registers available in the data set, which is simultaneously the *sampling frame*, optimal in the sense all units in the population have selection probability greater than zero. Thus, the sample size is 1285.

The second step consists in deciding how the sample is selected. Given that the sample frame has different variables that are useful for the determination of a representative sample, a stratified sample is chosen to select individuals. Two goals can be achieved by using this methodology: increase precision and obtain inferences about the strata, but sadly the selection of one goal or another has differential implications for sample size calculation within each stratum.

Since the survey is health related, it is of interest to obtain precise estimates for Belgium and for each province. The region is not considered as stratum because each province belongs to only one region, so it is assumed that if it is possible to obtain precise estimates for each province, it is also possible to aggregate those estimates and obtain precise estimates for each region. In addition, is also of interest to obtain inference about the health condition discriminated by sex, therefore is also considered as stratum.

Additionally, it is well-known that variables as age, income level and educational level are related to health status. However, it is not possible to consider them as strata because the number of possible combinations between them will explode. For this reason, they are considered in the selection process within each stratum in the following way:

- There are 12 provinces in the sample frame which combined with sex results in 24 strata.
- Within each stratum, the sample frame is ordered by age, income level and educational level following the hierarchic serpentine scheme.

- This scheme consists in sorting the sample frame by age in ascending order. Then within the first age found, income level is sorted in ascending order. Within the second age found, income level is sorted in descending order. This process continues to alternate between ascending and descending sorting throughout all ages clusters found in the sample frame. Then educational level is sorted within age and income level, again between descending and ascending order.

- This scheme minimizes the change from one observation to the next with respect to age, income and educational level values, thus making nearby observations more similar.

- Having done the sorting process, systematic sampling is conducted within each stratum to be sure that all possible combinations of these variables are present in the sample. Note that it is possible to obtain a similar sample with simple random sampling but it will be determined just by chance (very unlikely).

As it was mentioned before, the selection of one goal of stratification or another has implications in the sample size calculation within each stratum. Because the interest of this survey is to obtain precise estimates at the national level as well as strata level, a proportional compromise allocation is chosen. So the sample size for each stratum is calculated as:

$$n_h = n\frac{N_h^k}{\sum_{h=1}^{H} N_h^k}$$

Where n is the total sample size, n_h is the sample size of stratum h, N_h is the population size of stratum h, and k is set to 0.5; considering when k is equal to zero the focus is on the precision at the stratum level, whereas when it is equal to one is at the population level.

It is important to notice that there are four individuals in the sample frame with no province (three of them are in Brussels and one is in Flanders) and no information in the variables of interest. Therefore, they are not considered in the sampling process and the sample size considered is 1284.

Finally, to select this sample the procedure *surveyselect* in the statistical software SAS is used. This is the code used:

```
*        Stratified sample with proportional compromise allocation;
proc tabulate data=sframe out=sex_province;
    class sex province;
    table province,sex;
run;

*   Sample size in each stratum;
*   sum_N_h is the denominator;
data sex_province;
        set sex_province;
        N_h = N ** 0.5;
        sum_N_h = 424.5402099; * Sum of N_h;
        _NSIZE_ = ROUND(1284*(N_h/sum_N_h),1);
run;
```

```
*   Sample selection;
proc surveyselect data=sframe out=sample_prop_comp
    method=sys seed=12345 sort=serp outsort=sframe_sort sampsize=sex_province;
    strata province sex;
    control age7 fa3 edu3;
run;
```

This code contains all of the essential information:

- Proc tabulate computes the population size for each stratum as combination of province and sex (24 in total).
- Data sex_province is a second dataset that contains the sample size for each stratum using the output from the procedure mentioned above. Here the total sample size is considered.
- In proc surveyselect, data=sframe and out=sample_prop_comp are the input and output datasets.
- method = sys option specifies the choice for systematic sampling (By definition is without replacement).
- seed=12345 option initiates the random number generator. This ensures that the same sample is obtained every time the code is used.
- sort=serp option specifies hierarchic serpentine sorting within each strata.
- outsort=sframe_sort option specifies a new dataset with the sample frame sorted by the control variables.
- sampsize=sex_province option specifies the sample size for each stratum, as defined before.
- strata province sex option indicates strata in the sample design
- control age7 fa3 edu3 option indicates what variables are considered to sort the sample frame within each stratum.

As a result, the sample obtained presents similar distribution in each variable considered in the design as in the considered total population, see Table 1. Note how the distribution in each province in some cases is not similar to the population because of the compromise allocation. However, the inference will be more precise compared to a proportional allocation, where the overall distribution of the sample would be almost the same as the population. In regard to age, income level and educational level, the distribution is pretty similar that the population, which means that systematic sampling is always a good idea.

Variable	**Category**	**Population**		**Sample**	
		Freq.	**Prop.**	**Freq.**	**Prop.**
Region	**Brussels**	2568	30	216	16.81

Variable	Category	Population		Sample	
		Freq.	Prop.	Freq.	Prop.
	Flanders	2986	34.88	519	40.39
	Walloonia	3006	35.12	550	42.8
Province	Antwerpen	796	9.3	121	9.42
	Vlaams Brabant	506	5.91	96	7.47
	Limburg	386	4.51	84	6.54
	Oost Vlaandaren	678	7.92	111	8.64
	West Vlaanderen	620	7.24	107	8.33
	Brabant Wallon	281	3.28	72	5.6
	Hainaut	1075	12.56	141	10.97
	Liege	774	9.04	119	9.26
	Luxembourg	238	2.78	66	5.14
	Namur	392	4.58	85	6.61
	Brussels	2568	30	216	16.81
	Eupen	246	2.87	67	5.21
Sex	Male	4139	48.35	633	49.26
	Female	4421	51.65	652	50.74
Age	15-24	1149	13.42	176	13.7
	25-34	1644	19.21	243	18.91
	35-44	1615	18.87	244	18.99
	45-54	1296	15.14	196	15.25
	55-64	1094	12.78	166	12.92
	65-74	1078	12.59	160	12.45
	75+	684	7.99	100	7.78
Income level	<30000	4325	50.53	642	49.96
	30000-40000	2700	31.54	415	32.3
	40000+	1131	13.21	165	12.84
	Missing	404	4.72	63	4.9
Educational level	<=Primary	2976	34.77	441	34.32
	Secondary	2425	28.33	367	28.56
	Higher	2805	32.77	412	32.06
	Missing	354	4.14	65	5.06

Table 1: Descriptive statistics

Analysis of selected sample

In addition, the selected sample is analyzed where a model is built using as response variable a trichotomous version of the Body Mass Index (BMI), as follows

$$BMI_c = \begin{cases} 1 & \text{if } BMI \leq 20 \\ 2 & \text{if } 20 \leq BMI \leq 25 \\ 3 & \text{if } BMI > 25 \end{cases}$$

Here, three scenarios are considered: a model that ignores both the sampling design and structure of data; a model that considers the sampling design and a model that considers the structure of the data.

In SAS, different procedures should be used to analyze these three scenarios, to estimate the effect of covariates on the response variable. The first approach consists in ignoring the fact that the data is obtained via probability sampling and assume independence between individuals, this is done via `proc logistic`. The second approach considers all sampling design characteristics in the estimation of the model, this is done via `proc surveylogistic`. A third way to tackle this problem is through Generalized Estimating Equations, where the correlation between individuals within the same houselhold is taken into account, this is done via `proc genmod`.

Body mass index (BMI) has been directly related with sex, age, smoking patterns, education and income, so these variables are considered in the model.

It is important to mention that the same procedures can be used, even though the response variable is not binary. In the case of `proc logistic` and `proc surveylogistic`, the procedure recognizes that the response has three categories and performs a cumulative model. It fits a common slopes cumulative model, which is a parallel lines regression model based on the cumulative probabilities of the response categories rather than on individual probabilities, i.e., it assumes that the estimates for each covariate are the same for every category in the response variable. In `proc genmod`, the same model is conducted but the distribution is now multinomial and the link function used is cumulative logit. The reference category in the response variable is BMI smaller or equal than 20. In this way, the estimates of these models are comparable.

All variables considered in the model are categorical. Age has seven categories, education level and income level have three categories each and sex and smoke are dichotomous variables. In variables that have three categories or more, reference-coding is used taking the last category as reference. So in Age the reference category is people with 75 years or more; in educational level the reference category is higher and in income level the reference category is incomes higher than 40000. Binary variables are treated as dummy variables.

The first model considered is a main effects model with all variables mentioned, where neither income level nor smoking are significant. The same happens in the three scenarios mentioned above.

By removing these variables, age, educational level and sex are significant to the model, so a comparison between the estimates of the three approaches is shown, see Table 2.

Parameter	Logistic		SurveyLogistic		GEE	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept 1	-2.627	0.31	-2.326	0.31	-2.326	0.311
Intercept 2	0.1	0.298	0.332	0.301	0.332	0.301
Female	0.796	0.118	0.803	0.121	0.803	0.121
15-24	1.122	0.266	0.808	0.28	0.808	0.281
25-34	-0.334	0.251	-0.462	0.271	-0.462	0.273
35-44	-0.464	0.251	-0.607	0.27	-0.607	0.27
45-54	-0.803	0.259	-1.002	0.274	-1.002	0.272
55-64	-1.419	0.269	-1.47	0.288	-1.47	0.288
65-74	-1.033	0.267	-1.355	0.293	-1.355	0.293
<=Primary	-0.572	0.147	-0.656	0.156	-0.656	0.158
Secondary	-0.381	0.146	-0.414	0.151	-0.414	0.15

Table 2: Comparison of estimates

When comparing the estimates, the logistic approach differs in comparison with surveylogistic and gge, where results are identical. In terms of standard errors, is also clear how not considering design aspects when analyzing this kind of data underestimates them. As a consequence, this example shows the importance of considering the appropriate structure for the analysis of sampling data, because if it is ignored, it could lead to wrong results and inferences.

Modification Sampling Design BHIS

The BHIS sampling is a *Stratified Clustered Multi-Stage design*. In detail, the Belgian population is stratified by Region (Flanders, Wallonia, Brussels) and Provinces within the region; where no sampling is made since all strata is selected. The first stage, consist in doing a systematic sampling of the towns within provinces, where a list of them ordered by their size is used. The second stage, involves the selection of households (HH) in the selected towns, again using systematic sampling in a list ordered by statistical sector, the size of the HH and age of the reference person. The last stage considers the selection of at most four members of the HH.

Now, it is considered the implications of modifying the sampling design by not considering provinces in the stratification, but regions and a subdivision of the provinces (see Table 3). The total and within each region sample size remains as in the previous design: 10000 in total, 3500 for Flanders, 3500 for Wallonia and 3000 for Brussels.

Flemish Province	**Care sectors**
West-Vlaanderen	Brugge, Oostende, Roeselare and Kotrijk
Oost-Vlaanderen	Gent, Sint-Niklaas and Aalst
Vlaams-Brabant	West and Leuven
Antwerpen	Antwerpen, Mechelen and Turhout
Limburg	Hasselt and Genk
Walloon Province	**Arrondissement**
Hainaut	Tournai, Charleroi, Thuin, Mons and Soignies
Lige	Huy, Lige, Verviers, German Community
Luxembourg	Arlon and Neufchateau
Namur	Namur and Dinant
Brabant-Wallon	Nivelles

Table 3: Proposed stratification at province level for the BHIS

It is possible to say that there is still stratification at the region and province level in an indirect way. The major difference is that at the province level, there are now more strata (going from 11 provinces to 29 care-sector/arrondissement). The gain of considering a sample at the care-sector/arrondissement level is that now it is possible to give inferences for each of those and still give a precise estimate at province level by aggregating these sectors, since each of them belongs to one and only one province.

If it is assumed that sample sizes are defined according to the population size of each subdivision (care-sector/arrondissement), then the provinces will have a similar proportion with respect to the overall sample size as before. Table 4 presents the proportion of the sample size of each province for the new approach and the two previous designs. There, it can be seen how similar are the resulting proportions.

Flemish Province	**New Proportion**	**2001**	**2004**
West-Vlaanderen	0.183	0.190	0.190
Oost-Vlaanderen	0.229	0.229	0.229
Vlaams-Brabant	0.173	0.171	0.171
Antwerpen	0.281	0.277	0.277
Limburg	0.134	0.133	0.134
Walloon Province	**New Proportion**	**2001**	**2004**
Hainaut	0.372	0.393	0.381
Lige	0.305	0.290	0.305
Luxembourg	0.077	0.075	0.075
Namur	0.136	0.135	0.133
Brabant-Wallon	0.109	0.107	0.106

Table 4: Proportion of sample size by Province

Given that this design is stratified by region and care-sector/arrondissement, no random selection is made here. It is now assumed that further stages of the sampling remain as in the original design, this means that in each care-sector/arrondissement there would be a systematic sampling to select towns. Within towns a selection of households (HH) is based on a systematic sampling using the National register as before and inside the households, individuals are selected according to the established methodology. So, it is again a three stage sampling, where primary sampling units are towns, secondary sampling units are HH and tertiary sampling units are individuals.

Since total population size for the provinces remains the same when taking into account the care-sector/arrondissement division, it is assumed that towns within each division belongs to the same province as in the original design. However, town size is not relative to the province anymore, but to the care-sector/arrondissement it belongs to. In this sense, the entity size could influence the selection of a town or the number of times it is selected in comparison with the original design.

Additionally, in both sampling designs of the survey, 2001 and 2004, an oversampling for the German community was considered. In the new approach, there is an arrondissement for them, so according to its population size it could only have approximately 75 surveys, whereas in the other designs had 300.

As in the original design, it is expected that the efficiency of estimations at National and Regional level are sufficient, but it could be too small for estimation purposes at entity level (Demarest et al., 2001). In a similar fashion, in this design, weights should be considered to make correct inferences to the national level, but now these should consider the size of the entity relative to region-population.

For this reason, the weights in this design would change in the sense that different strata are defined. Given that the weights in the BHIS take into account the selection probability, province, and region of the participant and that the care-sector/arrondissement are smaller geographical areas than the provinces, the probability of select a specific individual change and so its corresponding weight. In particular, every participant in the survey had an assigned weight that considered the variables province, age, sex, household size and quarter in which the interview took place Charafeddine et al. (2013). These weights depended on the total number of interviews and the size of the province. So their corresponding weights should be changed relative to the size of the subdivision of the province.

It is important to mention that in the particular case of the German community, if oversampling is conducted, then a correction of these weights is needed as well.

Missing Data in the BHIS

The BHIS considers two levels of non-response: at unit non-response (a selected participant does not respond to the entire survey) and at item non-response. At the first level, there could be missingness of an entire household or a household member, while at item level three different types of missing values are considered: not applicable, no answer and does not know. In regard to analysis, missing values are not considered and a complete case analysis is used (Charafeddine et al., 2013), which means that any individual that has a missing value is excluded from the analysis.

As for the design, the survey considers three aspects to avoid missing values and reduce the non-response rates (Van Der Heyden et al., 2014)

1. Efforts are made in fieldwork to maximize the probability of response. Namely, prenotifications are sent to the selected HH, incentives are used to motivate participation, intensive training and follow-up of the interviewers. In detail, the interviewers had to register every time a household is contacted as well as information about the date, time, contact mode and result of the contact. If a contact was not successful in the first attempt, the interviewer would have to make at least four extra contacts in different weeks and time days.
2. After five failed attempts of contact, a house could be classified as *non contactable* and then could be replaced by a substitute household, this measure is defined as *field substitution*. The same list of the systematic sampling to select households within municipalities is used to select the substitutes. For each selected household, three additional consecutive households are selected, which guarantees that the substitutes will be located in the same statistical sector, and share the same characteristics regarding the size of the house and the age of the reference person as the initial selected unit.
3. Finally, in order to correct for bias due to non-response and the unequal selection probability of households and individuals given the regional stratification (which does not depend on size), the calculation of post-stratification weights is implemented by comparing the distribution of the sample regarding age, gender, household size and province to the total population, available from the National Register.

Demarest et al. (2017) state that the reason for using field substitution in the BHIS is to ensure that the predefined number of interviews and the composition of the sample in terms of sex, age group and household size is met at the end of the fieldwork.

On the other hand, in terms of individual non-response, intra-households substitution is not done. Van Der Heyden et al. (2014) found that the impact of this kind of refusal is minimum, since for the 2004 BHIS the 97.7% of household members agreed to participate.

References

Charafeddine, R., Demarest, S., Drieskens, S., Gisle, L., Tafforeau, J., and Van der Heyden, J. (2013). Health interview survey 2013: Research protocol.

Demarest, S., Molenberghs, G., Van der Heyden, J., Gisle, L., Van Oyen, H., de Waleffe, S., and Van Hal, G. (2017). Sample substitution can be an acceptable data-collection strategy: the case of the belgian health interview survey. *International Journal of Public Health*, 62(8):949–957.

Demarest, S., Tafforeau, J., Van Oyen, H., Bruckers, L., Molenberghs, G., Tibaldi, F., and Van Steen, K. (2001). Health interview survey 2001: Protocol for the sampling design.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):439–457.

Tourangeau, R., Michael Brick, J., Lohr, S., and Li, J. (2017). Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society*, 180(1):203–223.

Van Der Heyden, J., Demarest, S., Van Herck, K., De Bacquer, D., Tafforeau, J., and Van Oyen, H. (2014). Association between variables used in the field substitution and post-stratification adjustment in the belgian health interview survey and non-response. *International Journal of Public Health*, 59(1):197–206.