Survey Sampling Methods

Moritz Baten, Burak Demiral, Sergej Kaiser

June 20, 2016

1 Critical reading

2 Critical reading

3 Own Analysis of Survey Data

4 Analysis of Survey Data

5 Design of Health Interview Survey

3.A Description of the Belgian Health Survey

The sample for the Health Interview Survey (HIS) 2013 consists of 10.000 units each of which selected from the National Registry throughout the year. Belgium consists of 3 regions and in order to be able to talk about these regions, the sample is stratified into 3; 3.500 of which from Flanders, 3.500 Wallonia and 3.000 Brussels.[1] The regions are further stratified into provinces but this time proportional to their size so that a province won't be over or underrepresented.[2] This stratification ensures a geographical spread and if the provinces are homogenous within themselves it should result in higher precision compared to simple random sampling. In the end this results in total 12 strata.

This 2-step stratification is followed by a multi-stage sampling with three levels in total. First the municipalities within provinces, then households within municipalities and lastly individiuals within households.

In order to select the municipalities systematic sampling method was used in each province. The way this is done in a province is as follows: First municipalities were sorted in a descending order according to their population size and listed with a sum of population of the municipality itself and the cumulative sum of previous municipalities upto that municipality (C). If we want to pick N groups we set an interval I (total population in the province divided by N) according to that number N. And we iterate through the list until we go through N steps and pick N groups, at each step we calculate the threshold $((step_n o - 1) * I + R)$ and pick a group of 50 from the municipality with the closest C that is greater than this threshold. It should be noted that it is possible to select more than one groups from one municipality.

In the 2013 survey 225 groups of 50 were selected from 158 municipalities out of 589. This selection procedure ensures that large municipalities are selected since they are at the beginning of the list and also at least one of the smaller municipalities since they are lumped together at the end of the list. If it is assumed that smaller provinces share similar charecteristics it should be enough to pick at least one of them. Selecting groups of 50 is creates what is called a clustering effect decreasing the precision, similar individuals compared to a simple random sample, but this is still done in order to decrease the amount of effort in the fieldwork.

After selecting the municipalities that the groups of 50 will come from now another sampling is needed to be done in order to determine the households. This is also achieved by systematic sampling, but this selection is done throughout the year in four quarters, because it is likely that households move or people die within a year but also to take into account the temporal representativity, so average number of interviews per group per quarter is then 12.5.

This time households are ordered hierarchically according to 3 categories, first by (1) statistical sector, this is only relevant for municipalities that will give more than 1 group of 50 to the sample[3], then by (2) household size and lastly by (3) age of the reference person.

As before a step size is calculated by dividing the total number of households (N) by the required number of households (n, this is simply calculated by dividing the adjusted mean household size[4] size to 12.5) but this time instead of just picking one household from the list three other consecutive households are also taken from the list. This group of 4 households is called a cluster, this is done in order to have replacement households in case are non-responses from the selected household. The step size further divided by 2 so that there are double the amount clusters that is needed. And again a random number is picked and each at iteration

[1] On top of this there is also 600 additional sampling units for the province of Luxembourg as they payed extra money for this in order to acquire the desired bound of error for their province.

[2] The province Liège was further stratified into two in order to study the German Community.

[3] what is done is that in these municipalities are divided into statistical sectors so that one group comes out of each sector

[4] Calculated as the mean household size but households with more than 4 members are counted as 4 member households as only 4 people are interviewed per household

the step size is added to this random number and the household with the ordering closest to this number and the consecutive 3 is picked. In the end the clusters are listed and the first available household within a cluster is selected for the interview.

This sampling taken from an ordered list ensures that municipalities are well represented with respect to statistical sectors, household sizes and ages of the reference person.

Within a household maximum number of people to interview was capped at 4 in order to avoid familial correlation. This is avoided, because individuals in the same household tend to be similar compared to the individuals in other households, so interviewing one more individual within a household doesn't give much information about the sample in general. Then if a household contains more than 4 individuals there has to be a selection rule to pick 4 members out of the total household. Ideally this selection procedure should be done in total randomness in order to avoid bias, but there are some practical problems with this as it may be difficult to explain the reference person that they will not be interviewed and also the information for the general household questionnaire should come from the reference person, thus for the households with more than 4 members, reference (and the partner if there is one) is always interviewed and the following 3 (or 2 if there is a partner) members are selected by the birthday rule, which is whoever has the closest upcoming birthday to the day of interview is selected and since not everyone has the same selection probability in households with more than 4 member, weights are attached in proportion to inverse of the selection probability in order to mitigate the bias.

3.D Comparison of European Social Survey with General Social Survey (US)

Since the ESS survey is a rather young survey, it seems fitting to compare it to an incumbent with a lot of experience, the GSS from the United States.

The GSS uses a multi-stage area probabiltity sampling. The Primary Sampling Units are the Standard Metropolitan Statistical Areas (SMSAs) or counties in rural areas. Both are stratified by region, age and race before selection. The Secondary Sampling Units (SSU) are block groups (BG) or enumeration districts (ED). The latter describe an area which one interviewer can cover. These BGs and EDs are also stratified to race and income before selection. The Tertiary Sampling Units (TSU) are blocks who are selected proportional to size. In places without block statistics field counting is used.

In addition quotas on sex, age and employment status are used to account for the not-at-homes: District figures from census data are used to keep the sex proportional. Similarly proper portion of employed and unemployed women and proper portion of men over and under 35 were enfored. The argumentation here is that especially young women and men under 35 are difficult to sample otherwise.

Furthermore non-respondent cases are sub-sampled at the end of sampling (about ten weeks before the first release) period to spent more ressources on these difficult cases. This sub-sample is then weighted and added to the final sample.

	GSS	ESS
PSU	counties or metropolitan areas	census sections
stratification	yes: region, race and age	yes: region and town size
SSU	block groups or districts	list of individuals†
stratification	yes: race and income	no
TSU	blocks	–
stratification	no	no
Quota Sampling	yes on sex, age and employment	no, forbidden
oversampling	yes	not recommended, requires authorisation
non-respondent actions	oversampling, non-resp. sub-sample	urge to use more contact attempts

† applies to ESS countries who have a registry, for others fieldwork agencies and random route techniques are used

Table 1: Comparison of ESS and GSS

The ESS uses a two stage sampling procedure. The PSU are selected usually sections of the last census. The PSUs are stratified by region and town-size before selection. Afterwards they are ordered per target population and from this lists then the individuals are sampled (by SRS). There is no stratification on this level.

This procedure is applied if registries are available. In countries where no reliable lists of addresses or households is available like Portugal or Bulgaria PSUs are used but then use fieldwork agencies or random route techniques. Fieldwork agencies create lists of people within an area, while random route techniques use algorithms to randomly select sampling units. However it is not clear how random these random route techniques are. Thus they should be avoided if possible (Lyberg, 2000).

For smaller countries the ESS uses a minimum size of 800 for countries with a population of less than two million. In addition ESS urge the national coordinators to not over-sample anticipated low response rates.

6 Methodology

7 Methodology

7.1 Multistage sampling in survey data

7.2 Missing data in survey data