Németh, Renáta

Working Paper

Sampling design of health surveys: Household as a sampling unit

LIS Working Paper Series, No. 358

Provided in Cooperation with:
Luxembourg Income Study (LIS)

Suggested Citation: Németh, Renáta (2003) : Sampling design of health surveys: Household as a sampling unit, LIS Working Paper Series, No. 358, Luxembourg Income Study (LIS), Luxembourg

This Version is available at:
https://hdl.handle.net/10419/95404

Luxembourg Income Study Working Paper Series

Working Paper No. 358

Sampling Design of Health Surveys: Household as a Sampling Unit

Renáta Németh

September 2003

Luxembourg Income Study (LIS), asbl

Sampling Design of Health Surveys

- Household as a sampling unit –

Renáta Németh

Hungary

Abstract: The problem of drawing a person from a household often occurs at the final stage of a health survey design. In case of such designs, selection probabilities pertaining to the households are usually more or less equal. Sampling procedures that are used to select a person within households intend to be also quasi-random, i.e. conditional selection probabilities pertaining to persons living in the same household are equal. We found that, contrary to the widely held opinion, the above design is not capable of providing representativity by gender and age. We suggest a modification of the Kish grid, which is one of the widely used quasi-random procedures. The modified grid is more appropriate for selecting a representative sample. Since the performance of the grid depends on the household structure within the target population, its modification varies country by country. Those countries are considered in which the Kish grid is regularly used in surveys.

1. Introduction

Health surveys have a special role in obtaining health related and health behaviour related information which are not collectible through the existing registries, either because they fall out from the framework of standard health care, or because they can only be obtained directly from the people. The most typical of this sort of information is people's own perception of their health, and information related to a person's functionality and lifestyle. The surveys are also ideal to assess the public opinion on health policies and the quality of health care.

A sample survey may be defined as a study involving a subset (sample) selected from a larger population. Characteristics of interest are measured on each of the sampled individuals. From this information, extrapolations can be made concerning the entire population. The validity and reliability of these extrapolations depend on how the sample was chosen.

2. Reasons for sampling households

It is often best to draw the sample in two stages. These are designs in which primary sampling units are selected at the first stage, and secondary sampling units are selected at the second stage within each unit selected previously. Sampling designs considered in this paper are those in which households are selected at first, and then one adult member of each selected household is chosen.

When does the need for two-stage sampling arise, rather than selecting the respondents directly from the population? Lists of adults, from which the sample can be taken, are often not available. For example, the electoral register is usually a good quality database of addresses, but a poor quality database of individual adults. The register has many errors because of non-registration and population mobility. In practice, the register is used to construct a sample of flats or households, and the sample of adults is obtained at a second stage in some other way.

Another method involving respondent selection within household is called area sampling. It is used when the target population is located in a geographical region, such as a city. A frame for studying a population of a city may in the first stage consist of a list of districts, followed by a list of streets, followed by a list of blocks, then a list of households. And again, at the final stage, a sample of respondents is obtained from the sample of households.

The problem of translating a sample of the households into a sample of adult persons often arises in telephone surveys as well. Households are usually contacted by random digit dialling.

There is no need of selection if the respondent is uniquely defined e.g. as the head of the household. Suppose the household contains more than one member of the desired population. One may decide to include in the sample every member within the household. This may be a statistically inefficient procedure, unless one of these two conditions holds:

- There is seldom more than one member of the population in the household.
- If intra-class correlation within the household of the variables measured is of negligible size. Otherwise, the distribution is characterized by some homogeneity. Usually, the homogeneity of households is greater than in the case when individuals were assigned to them at random. Since homogeneity within sample clusters increases the variance of estimations, the sampler wants to reduce it in this case by selecting only one member per household (see Kish, 1965).

These conditions generally do not hold in surveys. Hence, there is a need for a procedure of selection that will translate a sample of households into a sample of the adult population. It is desired to make not more than one interview in every household. On the other hand, an interview in every sample household is desired in order to avoid futile calls on households without interviews. Finally, the procedure should be applied and checked without great difficulty. The simplest procedure could be applied is the uncontrolled selection in which the interview is conducted with those who open the door or answer the phone. A serious problem comes up in this case. The resulting sample will be made up of those persons more likely to be available at the time interviewer calls or who are most willing to be interviewed. Experiences show that they are made up of women and older adults.

3. The Kish grid

The Kish grid gives a procedure of selection[1]. The expression "Kish grid" comes from the name of Leslie Kish, the Hungarian born American statistician. Kish was one of the world's leading experts on survey sampling.

When creating the grid Kish intended to select persons within the household with equal probability. On the other hand the use of the grid can be checked easily contrary to e.g. a decision depending on the toss of a coin.

When applying the Kish grid, the interviewer at the first step uses a simple procedure for ordering the members of the household. A cover sheet is assigned to each sample household. It contains a form for listing the adult occupants (see Table 1), and a table of selection (see Table 2).

Table 1

Form for listing the adult occupants

Relationship	Sex	Age	No.	Selection
Head	M		2	
Wife	F	40	5	
Head's father	M		1	
Son	M		3	
Daughter	F		6	
Wife's aunt	F	44	4	✔

Source: Kish; 1965.

The interviewer lists each adult on one of the lines of the form. Each is identified in the first column by his/her relationship to the head of the household. In the next two columns, the interviewer records the sex and, if needed, the age of each adult. Then the interviewer assigns a serial number to each adult. First, the males are numbered in order of decreasing age, followed by the females in the same order. Then the interviewer consults the selection table. This table tells him the number of the adult to be interviewed. In the example, there are six adults in the household and selection table D tells to select adult number 4 (see Table 2).

[1] It is often used in health surveys. Some examples: *World Health Survey* 2002, WHO; *ICPE study on mental health*, Canada, 1997; *Oregon Health Behavior Surveys*, 1998, State of Oregon, Health Division; Health Survey for England, 1997; *Scottish Health Survey*, 1995, The Scottish Office Department of Health; *Risk Factors for Sleep Bruxism in the General Population*, telephone survey, Italy, 2000; *Hungarian Health Survey*, 1994, Central Statistical Office, Hungary*; Health Survey in Veresegyház* 1998, SOTE – TÁRKI, Hungary.

Table 2

One of the eight selection tables

Selection Table D	
If the number of adults in household is:	Select adult numbered:
1	1
2	2
3	2
4	3
5	4
6 or more	4

Source: Kish; 1965.

Selection table D is only one from the 8 types (see Table 3). One of the 8 tables (A to F) is printed on each cover sheet. The cover sheets are prepared to contain the 8 types of selection tables in the correct proportion, e.g. table A is assigned to one-sixth of the sample addresses. The aim is to reach equal selection probabilities within household without the necessity of printing many more forms. Table 4 shows the selection probabilities. It can be seen that the chances of selection are exact for all adults in households with 1, 2, 3, 4 and 6 adults. As numbers above six are disallowed, there are some adults who are not represented. On the other hand, there is an overrepresentation of number five in the households with five adults.

Table 3

Summary of eight selection tables

Proportion of assigned tables	Table number	If the number of adults in household is:					
		1	2	3	4	5	6 or more
		Select adult numbered:					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3	5
1/12	E2	1	2	3	4	5	5
1/6	F	1	2	3	4	5	6

Source: Kish; 1965.

Table 4

Summary of selection probabilities

Adult numbered	If the number of adults in household is:					
	1	2	3	4	5	6 or more
1	1	1/2	1/3	1/4	1/6	1/6
2		1/2	1/3	1/4	1/6	1/6
3			1/3	1/4	1/4	1/6
4				1/4	1/6	1/6
5					1/4	1/6
6						1/6
7 or more						0

It may be noted that the procedure was modified several times by many researchers. Kish himself suggests modifying the tables for special reasons. In paper-and-pencil interview the interviewer uses the grid as described above. In computer-assisted telephone or personal interviews, the tables are randomly assigned to the households by

the computer in their prescribed proportions. The researchers stick to ordering the persons by sex and age, though they have the technical background for generating random number. By using random numbers, it would be possible to select a person from the set of the previously identified adults. Although nobody expresses it explicitly, they consider the sample to be *representative* by sex and age with the use of the original Kish grid. This representativity would much less be expected if the applied procedure was e.g. identifying the adults by first name, then selecting one of them by generating a random number.

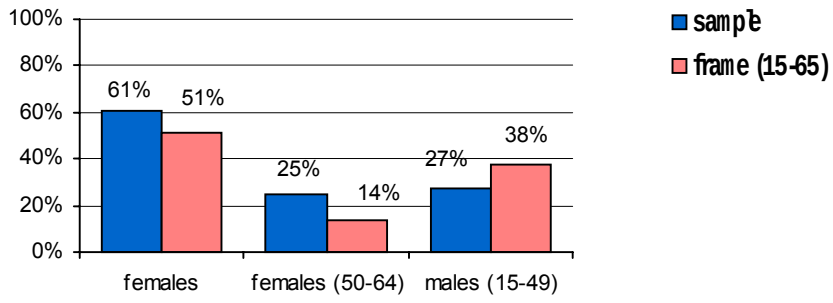
4. ”Representativity”

It was mentioned above that representativity is a desirable character of a sample. It refers to the similarity between the sample and the population in some characteristics of interest. Why is it desirable to reproduce the distribution of certain population characteristics in the sample? Suppose there is a high positive correlation between the characteristic to be estimated and a different one. The more representative the sample is by the latter one, the more reliable the estimation of the former one will be. (The *reliability* of an estimator is evaluated on the basis of its variance.)

It is a standard practice to evaluate the sample by its representativity in order to support the validity of the extrapolations or estimations. We attempted to take into account the accessible literature about samples obtained by using the Kish grid. When evaluating the representativity of their samples, Hungarian researchers often refer to the undersampling of males and overrepresentation of elderly people (see ISSP Család II, 1994, Táblaképek az egészségről, 2000, Egészségi Állapotfelvétel, 1994). The next two examples demonstrate this finding, presenting the result of two Hungarian health surveys. Figure 1 and 2 show that the sample differs from the sampling frame in sex and age distributions. Women are oversampled, especially elderly women. Further, males appear to be underrepresented. The same problems are reported by researchers from other countries.

Figure 1

KSH94: frame and sample

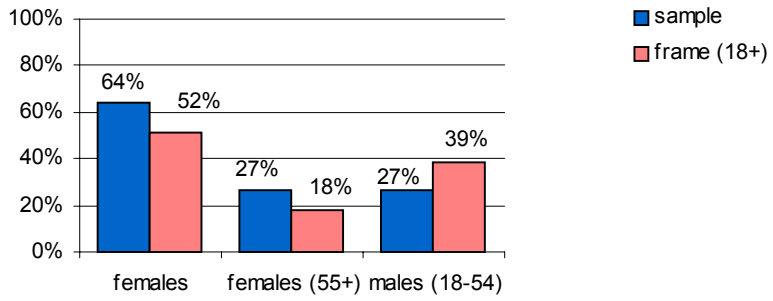


According to the comments this deviation stems from problems that occur when putting the interview into practice, e.g. males are undersampled because they are more difficult to find at home, and are less willing to participate. Later some theoretical evidence will be given that explains the representation problems without considering these assumptions.

It is important to mention that according to Kish’s own words, he used the variables sex and age only for ordering the household members. He did not aim explicitly to reproduce the sex and age distributions. At the same time, however, he expected the sample to be representative. In the article published first about the grid, Kish checked the distribution of the respondents, and explained the males’ underrepresentation with practical problems mentioned above: they are more difficult to find at home etc. [Kish, 1949]. Although he emphasized the fact that the grid is for random selection within household, he was the first not to make a distinction between randomness and representativity.

Figure 2

Veresegyház98: frame and sample



4.1 The cause of non-representativity - assumption

When households are selected with equal probabilities, and the selection probabilities within household are equal, then the chance of selection of a single adult becomes inversely proportional to the number of adults in the household. Hence overall selection probabilities are not equal.

If the selection probability is the function of the household size, and the household size is not independent from the members' demographic characteristics, then the sampling design itself is the source of the representation problems. The sample would not be representative even if perfectly random household sample and 100 percent response rate could be obtained. Kish found that the samples obtained by using the grid show close agreement with the population data on important demographic characteristics. Kish developed the grid in the 1950's USA. He emphasized the relatively low variance of the selection probabilities. That was because of the high concentration within a small range of household sizes: over 70 percent of households contained two adults (see Table 5).

Our results so far show that representativity is the function of the current household structure, and the grid's performance depends on where and when it is used. It is worth making a comparison between the current Hungarian household structure and the one observed by Kish. In today's Hungary, 26 percent of the households are one-person household that is 2 times greater than the one examined by Kish. This difference for itself is so significant that the question arises whether to accept the grid without modification.

Table 5

Household structure, USA, 1957

Number of adults in the household	1	2	3	4	5	6 or more
Proportion	14,6	73,0	9,0	2,8	0,4	0,2

Source: Kish; 1965.

4.2 The cause of non-representativity - proof

To put assumptions into a concrete form, the exact connection between the grid's performance and the population household structure need be determined. The required information on current Hungarian population is not available. That is the reason why we worked with a sample from a large national household survey[2]. The data contain information on each member of the sample households, so it can be used as a population for further sampling. In the following it will be referred as "*pseudopopulation*". Table 6 shows age and sex distribution in the pseudopopulation.

[2] Computations are based on datasets of the Luxembourg Income Study (LIS). The LIS database is a collection of household income surveys. *Microdatabase, (1994-2000); harmonization of original surveys conducted by the Luxembourg Income Study, asbl. Luxembourg, periodic updating.*

Table 6

Pseudopopulation, age and sex distribution (n=4248)

Age	Sex		
	Male	Female	Total
18–39	19,17	18,82	37,99
40–59	15,94	18,78	34,72
60+	10,58	16,72	27,30
Total	45,69	54,31	100,00

The use of the grid can be tested with the help of this pseudopopulation, concerning the age and sex distribution in the samples. The *expected* sex and age proportions of the sample can be formulated as follows. Let p_{kl} denote the selection probability of the adult l living in a household of size k ($k = 1 \ldots 6$, $l = 1 \ldots k$), supposing the household is already selected. As households are sampled with equal probabilities, the chance of choosing a household of size k equals to the proportion of these households. Let H_k denote this value. The expected sex and age-group joint distribution can be given by a 3×2 matrix, denoted by a. $a[11]$ is the proportion of young males, $a[21]$ is the proportion of middle-aged males etc., $a[32]$ is the proportion of elderly females.

Information on households' composition is also needed: after selecting the person number l within a household of size k, what is the probability of his or her being a male or female, of his or her being young or middle-aged or elderly. Let a_{kl} be 3×2 matrix ($k = 1 \ldots 6$, $l = 1 \ldots k$) In the above way, $a_{kl}[11]$ denotes the proportion of young males among the persons numbered l living in a household of size k, $a_{kl}[21]$ is the proportion of middle-aged males etc.

The expected age and sex joint distribution is the function of the other parameters (see Equation 1). H_k, a, and a_{kl} are known input parameters. They come from the information about pseudopopulation.

$$a[ij] = \sum_{k=1..6} H_k \left(\sum_{l=1..k} p_{kl} a_{kl}[ij] \right) \quad i = 1, 2, 3 \; j = 1, 2. \qquad /1/$$

Substituting the known parameters, the expected distribution shown in Table 7 is obtained.

Table 7

Expected age and sex distribution

Age	Sex		
	Male	Female	Total
18–39	16,24	17,61	*33,86*
40–59	14,79	18,06	*32,86*
60+	11,16	22,12	*33,28*
Total	*42,20*	*57,80*	*100,00*

It is seen that the expected sample differs from the population in sex and age distributions. Firstly, the elderly people are oversampled, especially the women. (It is worth mentioning, that in the current population of Hungary, great proportion of the one-person households consists of an older female occupant, and the quarter of all households is one-person-household, so it can be concluded that it is more likely to select an elderly female in this way than by simple random sampling.) Secondly, males appeared to be underrepresented. Our experiences are similar to those obtained from real surveys.

5. Modification of the Kish grid

5.1 Hungary

In this section a modification of the Kish grid is presented. Our intention was to receive a representative or at least more representative expected sample. The grid was modified by changing the selection tables. This modification method is not unprecedented in the literature: Kish himself already suggested modifying the tables when needed.

All the sampling features are fixed, i.e. the following conditions hold:

- each household has the same chance of selection
- one and only one interview per household is made
- the selection tables are based on a list of the household members
- this ordering is made by sex and age
- the population to be surveyed is the previously mentioned pseudopopulation
- 12 selection tables are used (Obviously, the more tables are used, the finer probabilities can be achieved, that is the closer agreement between the sample and the population can be obtained. This is why the number of the tables is limited.)
- the same rules are applied to households with 6 ore more members.

The problem is to make selection tables those yield a sample giving close agreement with the pseudopopulation data. The modification can be simplified: instead of determination of the tables, it is enough to determine the selection probabilities.

Our aim was to obtain a representative expected sample, which is as close to the distribution given by table 6 as possible. Let A denote the 3×2 matrix describing the sex and age joint distribution in the pseudopopulation, $A[11]$ equals to the young males proportion etc. Using the notation of Equation 1 the problem is as follows. H_k's and a_{ij}'s are given parameters, and a is to be determined as the functions of p_{kl}'s, so as to reproduce A. Equation 2 is to be solved.

$$\sum_{i=1,2,3\ j=1,2} \left| a[ij] - A[ij] \right| = 0, \qquad /2/$$

with constraints:

$$\begin{aligned} &\sum_{j=1\ldots i} p_{ij} = 1 \quad \forall i \\ &p_{ij} > 0 \quad \forall i, j \\ &p_{ij} = k_{ij}/12 \quad \forall i, j, \text{where } k_{ij} \text{ integer.} \end{aligned} \qquad /3/$$

The constraints make the solution to meet the conditions: one and only one person per households is needed, and 12 tables are used that means probabilities are given in 1/12. The model is a nonlinear equation, with inequality and integer constraints. Microsoft Excel Solver package was used to solve the equation. The problem has no solution.

This raises the question whether there would be a solution if the limitation of the number of the tables did not hold. Using more tables has a practical disadvantage: it implies increasing costs. Moreover, the sample size itself is an upper limit of the tables. Apart from this, the theoretical problem is worth considering. In this case, the integer constraint is to be omitted from /3/. The problem does not have a solution in this way either.

It is impossible to obtain a perfectly representative sample. Let us modify the problem instead: which selection table yields a sample that is the closest to the pseudopopulation. A distance function is needed to find the closest solution that minimizes the distance function. Two functions were used according to two different approaches. The first one is similar to the Pearson-chi-square. Equation 4 shows function f to be minimized.

$$f(a) := \sum_{i=1,2,3\ j=1,2} (a[ij] - A[ij])^2 / A[ij]. \qquad /4/$$

The idea of using the other distance function comes from weighting which is widely used in survey statistics. Weights are generally used to improve the precision of the estimate. Poststratification is a weighting method that produces a sample in which each stratum is represented in its proper proportion. In our case, strata are defined as the 6 cells of the sex and age group crosstable. Poststratification weight for a given person in a given stratum is defined as the proportion of the population stratum divided by the proportion of the sample stratum. The disadvantage of using the poststratification is that in some cases it increases the variance of the estimation. Increase in variance is a monotonic function of the sum of the weights squared. This implies the following approach: to find the selection table that yields a sample with the minimal sum of poststratification weights squared. Equation 5 shows function *g* to be minimized.

$$g(a) := \sum_{i=1,2,3\ j=1,2} A[ij]^2 / a[ij] = (1/n) \sum_{k=1\ldots n} W_k^2 , \qquad /5/$$

where n is the sample size.

As it was mentioned when finding the solution of the equation with absolute values, the constraints can be determined in two different ways. If they include the integer constraint, then the use of 12 tables are assumed. Otherwise, the number of the tables is not limited; therefore selection probabilities can be any real numbers between zero and one. Combining the two dimensions four problems appear: let us find the minimum value of function *f* or *g*, with or without the integer constraint.

A model, in which the objective function or any of the constraints is not a linear function of the variables, is called a nonlinear programming (NLP) problem. In our case, inequality and integer constraints are added to the model. The Weierstrass's theorem states that a real valued continuous function on a closed bounded set assumes a maximum and a minimum value. In our case the conditions of Weierstrass's theorem meets, but the determination of the minimum is not a simple mathematical problem. Apart from special cases the nonlinear optimization problems have numerical solutions. Microsoft Excel Solver package was used to find the minimums.

Table 8 contains the results.

Table 8

Optimization results

Original Kish-grid

	Function value when substituting Kish grid	Expected sex/age distribution (Matrix *a*)		p_{ij}					
	f: 0,024119	16,24	17,61	p_{21}	1/2	p_{31}	1/3	p_{41}	1/4
	g: 1,020764057	14,79	18,06	p_{22}	1/2	p_{32}	1/3	p_{42}	1/4
		11,16	22,12			p_{33}	1/3	p_{43}	1/4
								p_{44}	1/4
				p_{51}	1/6	p_{61}	1/6		
				p_{52}	1/6	p_{62}	1/6		
				p_{53}	1/4	p_{63}	1/6		
				p_{54}	1/6	p_{64}	1/6		
				p_{55}	1/4	p_{65}	1/6		
						p_{66}	1/6		

Optimization of function *f*

Constraints	Optimum value	Expected sex/age distribution (Matrix *a*)		p_{ij}					
$\Sigma_{j=1..3} p_{ij}=1\ \forall i$	0,012624653	17,96	17,63	p_{21}	2/3	p_{31}	1/12	p_{41}	3/12
$p_{ij}>0\ \forall i, j$		14,45	17,73	p_{22}	1/3	p_{32}	1/12	p_{42}	6/12
$p_{ij}=k_{ij}/12\ \forall i, j,$		12,14	20,09			p_{33}	10/12	p_{43}	1/12
where k_{ij} integer								p_{44}	2/12
				p_{51}	1/12	p_{61}	1/12		
				p_{52}	7/12	p_{62}	7/12		
				p_{53}	1/12	p_{63}	1/12		

				p_{54}	1/12	p_{64}	1/12		
				p_{55}	2/12	p_{65}	1/12		
						p_{66}	1/12		
$\Sigma_{j=1..k} p_{ij}=1\ \forall i$, $p_{ij}>0.01\ \forall i, j$	0,01075269	18,11	17,80	p_{21}	0,7029	p_{31}	0,0100	p_{41}	0,3225
		14,55	17,65	p_{22}	0,2971	p_{32}	0,0100	p_{42}	0,4885
		12,32	19,57			p_{33}	0,9800	p_{43}	0,0100
								p_{44}	0,1790
				p_{51}	0,0100	p_{61}	0,0100		
				p_{52}	0,9600	p_{62}	0,9500		
				p_{53}	0,0100	p_{63}	0,0100		
				p_{54}	0,0100	p_{64}	0,0100		
				p_{55}	0,0100	p_{65}	0,0100		
						p_{66}	0,0100		

Optimization of function *g*

Constraints	Optimum value	Expected sex/age distribution (Matrix *a*)		p_{ij}					
$\Sigma_{j=1..k} p_{ij}=1\ \forall i$ $p_{ij}>0,\ \forall i, j$ $p_{ij}=k_{ij}/12\ \forall i, j$, where k_{ij} integer	1,011464011	17,88	17,63	p_{21}	2/3	p_{31}	1/12	p_{41}	3/12
		14,52	17,73	p_{22}	1/3	p_{32}	1/12	p_{42}	6/12
		12,16	20,09			p_{33}	10/12	p_{43}	1/12
								p_{44}	2/12
				p_{51}	2/12	p_{61}	1/12		
				p_{52}	6/12	p_{62}	7/12		
				p_{53}	1/12	p_{63}	1/12		
				p_{54}	1/12	p_{64}	1/12		
				p_{55}	2/12	p_{65}	1/12		
						p_{66}	1/12		
$\Sigma_{j=1..k} p_{ij}=1\ \forall i$ $p_{ij}>0.01,\ \forall i, j$	1,009849293	18,06	17,75	p_{21}	0,7002	p_{31}	0,0100	p_{41}	0,3486
		14,66	17,63	p_{22}	0,2998	p_{32}	0,0100	p_{42}	0,4850
		12,30	19,60			p_{33}	0,9800	p_{43}	0,0100
								p_{44}	0,1564
				p_{51}	0,0100	p_{61}	0,0100		
				p_{52}	0,9600	p_{62}	0,9500		
				p_{53}	0,0100	p_{63}	0,0100		
				p_{54}	0,0100	p_{64}	0,0100		
				p_{55}	0,0100	p_{65}	0,0100		
						p_{66}	0,0100		

Some expected trends can be observed in all the four cases. For example $p_{21} \sim 2/3$, that affects against the males' underrepresentation that was found when using Kish grid (since p_{21} is the selection probability of the first adult in a two-persons-household, and the first one tends to be male because of the ordering procedure).

The optimal sex and age group distributions (matrix *a*) compared to the one belonging to the Kish grid shows that we managed to improve the young people and the females agreement with the population data, while other cells show some change for the worse.

The four solutions do not differ from each other, either regarding matrix *a* or p_{ij}'s. This means it is not worth using more than 12 tables. Moreover, the return value of function *g* at the optimum place of function *f* is very close to the real optimum value of *g* - and vice versa, i.e. the optimal tables are close to each other whether measured by *f* or measured by *g*. It can be said that the optimal methods perform well from both points of view.

Table 9 presents the modified selection table obtained by function *f* with the integer constraint.

Table 9

Modified Kish-tables

Proportion of assigned tables	Table number	If the number of adults in household is:					
		1	2	3	4	5	6 or more
		Select adult numbered:					
1/12	1.	1	1	1	1	1	1
1/12	2.	1	1	2	1	2	2
1/12	3.	1	1	3	1	2	2
1/12	4.	1	1	3	2	2	2
1/12	5.	1	1	3	2	2	2
1/12	6.	1	1	3	2	2	2
1/12	7.	1	1	3	2	2	2
1/12	8.	1	1	3	2	2	2
1/12	9.	1	2	3	2	3	3
1/12	10.	1	2	3	3	4	4
1/12	11.	1	2	3	4	5	5
1/12	12.	1	2	3	4	5	6

5.2 Other countries

Since the performance of the grid depends on the household structure of the target population, its modification varies country by country. In the following those countries are considered in which the Kish grid is used in health surveys or in other surveys. The national datasets are from the database of the Luxembourg Income Study (LIS)[3]. The optimal solutions were obtained by optimizing function *f* with the integer constraint. Table 10 presents the results. Countries in the table are sorted by D1 that is the distance between the pseudopopulation and the sample obtained by using the Kish grid. As before distance between two distributions was measured by function *f*. It can be seen that the performance of the grid is the worst in Italy, and the best in Canada. The pseudopopulation and the sample are about four times as far from each other in Italy than in Canada. Hungary is among the worst three countries. The United Kingdom, where health surveys are usually carried out by using the grid, is among the best ones.

The forth column presents the distance between the pseudopopulation and the expected sample got by using the optimal solution (D2). The list of the countries sorted by D1 or sorted by D2 can be compared with each other. E.g. it can be seen that Russia moved from a middle position into the last one, i.e. the improvement in its case was more significant than in the case of other countries.

Efficiency of the modification can be evaluated with the help of the last two columns. The absolute difference between the original distance and the optimal distance is shown in the penultimate column. It can be seen, that usually the worse of the performance of the original grid, the greater absolute improvement can be achieved. The last column presents the percentage difference between the original and the optimal distance, i.e. the relative improvement obtained through optimization. The distance from the pseudopopulation was decreased by 20-70%, thus the grid is successfully modifiable in each country. A significant improvement was achieved in Italy, in Hungary, in Slovenia, in Austria, in Russia, in Estonia and in the USA.

Table 10

Optimization results

Country	*Year of the survey*	*Distance between the pseudopopulation and the Kish grid sample (D1)*	*Distance between the pseudopopulation and the optimal solution (D2)*	*Difference between the original and the optimal one (D1-D2)*	*% Difference between the original and the optimal one (100(D1-D2)/D1)*
Italy	1995	0,0278	0,0147	0,013	47,4
Czech Republic	1996	0,0267	0,0211	0,006	20,9
Hungary	1999	0,0241	0,0126	0,011	47,7
Poland	1999	0,0235	0,0157	0,008	33,0

[3] *Microdatabase, (1994-2000); harmonization of original surveys conducted by the Luxembourg Income Study, asbl. Luxembourg, periodic updating.*

Slovenia	1999	0,0222	0,0106	0,012	52,3
Germany	1994	0,0217	0,0163	0,005	24,8
Ireland	1996	0,0186	0,0112	0,007	39,7
Belgium	1997	0,0175	0,0138	0,004	21,2
Austria	1995	0,0171	0,0096	0,008	44,0
Russia	1995	0,0162	0,0050	0,011	69,3
France	1994	0,0139	0,0105	0,003	24,0
Netherlands	1994	0,0133	0,0102	0,003	23,3
Estonia	2000	0,0132	0,0070	0,006	47,1
Norway	1995	0,0132	0,0098	0,003	25,3
United Kingdom	1999	0,0127	0,0100	0,003	21,5
Australia	1994	0,0114	0,0087	0,003	23,7
United States	2000	0,0091	0,0054	0,004	41,1
Finland	2000	0,0080	0,0059	0,002	26,5
Canada	1998	0,0072	0,0055	0,002	23,9

5.3 Summary

The main results of our work are as follows.

- The samples obtained by using the Kish tables differ from the population in sex and age group distributions. It was proved that this is caused by the sampling method and not by practical problems. The literature does not prove this connection, nor does it normally mention it.
- The grid is successfully modifiable if our aim is to adjust the sample to the population.
- The problem treated is of international significance. The trends observed in Hungarian household structure are global trends. Size of households is currently decreasing, and the proportion of the single persons is on the rise.

There are further problems to be considered. The scope of the analysis was limited to the representativity by sex and age. It should be useful to take into account the distributions of other characteristics when using the grid. At the same time, the distributions of other characteristics need checking when using the modified tables. Obviously, improving the sex and age adjustment does not mean that the sample shows agreement to the population with respect to other variables.

Change in selection probabilities implied by the modification needs further consideration. The variability of the probabilities can result in an increase of the estimation variance.

The optimal solution was derived from the pseudopopulation. It would be worth developing the study with the real population as a starting point in order to support the generalization of the results.

6. References

D. Binson, J.A. Canchola, and J.A. Catania. Random selection in a telephone survey: A comparison of the kish, next-birthday, and last-birthday methods. Journal of Official Statistics, 16:53–59, 2000.

R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg, editors. Telephone Survey Methodology. John Wiley and Sons, Inc., New York, 1988.

ISSP. ISSP 1994, Család II. TÁRKI, 1994. http://www.tarki.hu.

J.M. Kennedy. A comparison of telephone survey respondent selection procedures, 1993. http://www.indiana.edu/csr/aapor93.html.

L. Kish. A procedure for objective respondent selection within the household. Journal of the American Statistical Association, pages 380–387, 1949.

L. Kish. Survey Sampling. John Wiley and Sons, Inc., New York, 1965.

Egészségi Állapotfelvétel 1994 - Életmód, kockázati tényezők. Központi Statisztikai Hivatal.

Magyar Statisztikai Évkönyv 1993. Központi Statisztikai Hivatal, Budapest, 1994.

Mikrocenzus, 1996 A népesség és a lakások jellemzői. Központi Statisztikai Hivatal, Budapest, 1996a.

Mikrocenzus, 1996 Főbb eredmények. Központi Statisztikai Hivatal, Budapest, 1996b.

Magyar Statisztikai Évkönyv 1999. Központi Statisztikai Hivatal, Budapest, 2000.

P.J. Lavrakas. Telephone survey methods: Sampling, selection and supervision. Applied Social Research Methods Series, 7, 1993.

P.S. Levy and S. Lemeshow. Sampling of Populations. John Wiley and Sons, Inc., New York, 1999.

Táblaképek az egészségről - A veresegyházi példa. MTA Szociológiai Kutatóintézet - Fekete Sas Kiadó, 2000.

R.W. Oldendick, G.G. Bishop, S.B. Sorenson, and A.J. Tuchfarber. A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. Journal of Official Statistics, 4:307–318, 1988.

T. Piazza. Respondent selection for CATI/CAPI (Equivalent to the Kish Method). http://srcweb.berkeley.edu:4229/res/rsel.html.