Bulletin de méthodologie sociologique
Bulletin of sociological methodology

83 | 2004
July

Representativeness Problems Inherent In Address-Based Sampling And A Modification Of The Leslie Kish Grid

Renáta Németh

Electronic version
URL: http://journals.openedition.org/bms/248
ISSN: 2070-2779

Publisher
Association internationale de méthodologie sociologique

Printed version
Date of publication: 30 July 2004
Number of pages: 43-60
ISSN: 0759-1063

Electronic reference
Renáta Németh, « Representativeness Problems Inherent In Address-Based Sampling And A Modification Of The Leslie Kish Grid », *Bulletin de méthodologie sociologique* [Online], 83 | 2004, Online since 09 July 2008, connection on 19 April 2019. URL : http://journals.openedition.org/bms/248

This text was automatically generated on 19 April 2019.

© BMS

Representativeness Problems Inherent In Address-Based Sampling And A Modification Of The Leslie Kish Grid

Renáta Németh

Introduction

1 The Leslie Kish grid or its modified versions are often used in everyday survey practice all over the world[1]. The grid, like the last birthday or next birthday method, is capable of selecting an adult randomly within the household. When applying it, households are selected at the first stage, and then the interviewer lists all residents over 18 at the address and randomly selects one of them. In second section below, those factors are discussed that may justify the use of this address-based two-stage sampling.

2 The paper[2] is centred on the representativeness of samples obtained by the Kish grid. The usual way of applying the grid is described in the thrid section. According to our hypothesis expressed in the fourth section, the representativeness problems can be derived from the design itself; i.e., considering practical problems (for example, systematic non-response) is unnecessary to explain them. The hypothesis is supported by samples obtained in Hungarian surveys. Beyond this particular example, a mathematical derivation suitable for proving the hypothesis is showed in the second part of section four. In the proof, the composition of the theoretically expected sample is determined. It is also showed that the key is not responsible for the problem, since the same phenomenon occurs necessarily in the case of all address-based samples. On the other hand, the grid is a manageable case, because the age and sex composition of the sample can be controlled when applying it. Taking advantage of this fact, a modification of the grid is carried out in the fifth section.

Reasons For Sampling Households

3 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 previously selected unit. Sampling designs considered in this paper are address-based sample designs in which households are selected at first, and then one adult member of each selected household is chosen.

4 When does the need for *two-stage sampling* arise, instead of using population register sample that selects 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 containing many errors due to non-registration and population mobility, is usually a good quality database of addresses, but a poor quality database of individual adults. 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.

5 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.

6 The problem of translating a sample of households into a sample of adult persons often also arises in telephone surveys when households are usually contacted by random-digit dialling.

7 There is no need for selecting an individual if the respondent is uniquely defined, such as the head of the household. Suppose the household contains more than one member of the desired population, and one may decide to include in the sample every member in 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 within-household intra-class correlation of the measured variables is of negligible size.

8 Otherwise, the distribution is characterized by some homogeneity. Usually, within-household homogeneity is greater than in the case when individuals were assigned to the households at random. Since homogeneity within sample clusters increases the estimation variances, these can easily be reduced by selecting only one member per household (see Kish, 1965).

9 The two conditions listed above generally do not hold in surveys. Hence, there is a need for a selection procedure that will translate a sample of households into a sample of the adult population, taking into consideration the following criteria: Primarily, it is desired to make not more than one interview in every household. Secondarily, an interview in every sample household is desired to avoid futile calls on households without interviews. Finally, the procedure should be applied and be checked without great difficulty.

10 The simplest procedure which could be applied here is the uncontrolled selection in which the interview is conducted with those who open the door or answer the telephone. However, a serious problem arises in this case. The resulting sample will be made up of those persons more likely to be available at the time interviewers call or who are most

willing to be interviewed. Experience shows that these respondents are largely women and older adults.

The Kish Grid

11 The Kish grid (also known as Kish tables) provides a selection procedure. The expression "Kish grid" comes from the name of Leslie Kish, the Hungarian born American statistician who was one of the world's leading experts on survey sampling.

12 When creating the grid, Kish intended to select persons within the household with equal probability. Moreover, he recommended the grid since its proper use can be checked easily, compared to other methods such as a decision depending on tossing a coin.

13 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 (Source: Kish, 1965.)

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	

Table 2: One of the eight selection tables (Source: Kish, 1965.)

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

14 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 by decreasing age, followed by the females in the same order. Then the interviewer consults the selection table. This table tells him the number of adults to be interviewed. In the example, there are six adults in the household and selection table D indicates the selection of adult number 4 (see Table 2).

15 Selection table D is only one from eight types (see Table 3). One of the eight tables (A to F) is printed on each cover sheet. The cover sheets are prepared in such a way that they contain the eight types of selection tables in the correct proportion; for exaple, 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 identical 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 (Source: Kish, 1965.)

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

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

16 It may be noted that the procedure has been modified several times by many researchers. Kish himself suggests modifying the tables for special reasons. In paper-and-pencil interviews, 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 prescribed proportions. The researchers stick to ordering the persons by sex and age, though they have the technical background for generating random numbers. By using random numbers, it would be possible to select a person from the set of the previously identified adults. Although nobody states so explicitly, they consider the sample to be representative by sex and age with the use of the original Kish grid. This representativeness would be expected much less if the applied procedure was, for example, identifying the adults by first name, then selecting one of them by generating a random number.

“Representativeness”

17 Described above as a desirable property of a sample, representativeness refers to the similarity between the sample and the population for 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 for 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).

18 It is a standard practice to evaluate the sample according to its representativeness to justify the validity of the extrapolations or estimations. We attempted to take into account the accessible literature on samples obtained using the Kish grid. When evaluating the representativeness of their samples, Hungarian researchers often refer to the undersampling of males and overrepresentation of elderly people (ISSP Család II. 1994, Táblaképek az egészségről 2000, Egészségi Állapotfelvétel 1994). The next two examples demonstrate this finding, presenting the results of two Hungarian health surveys. Table 5 and 6 show that the sample differs from the sampling frame in sex and age distributions: women, especially elderly women, are oversampled, while young males

appear to be underrepresented. The same problems are reported by researchers in other countries.

Table 5: KSH94: frame and sample

	Females	Females (55+)	Males (18-54)
Sample	60,50%	24,50%	27,40%
Frame (18+)	50,90%	13,70%	37,70%

Table 6: Veresegyház98: frame and sample

	Females	Females (55+)	Males (18-54)
Sample	63,80%	26,90%	26,80%
Frame (18+)	51,70%	18,20%	38,80%

19 According to the researchers' comments, this deviation stems from problems that occur when putting the interview into practice: for example, 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.

20 It is important to mention that according to Kish, he used the variables sex and age only for ordering the household members. He did not explicitly intend to reproduce the sex and age distributions. At the same time, however, he expected the sample to be representative. In the first article published on the grid, Kish checked the distribution of the respondents and explained male underrepresentation by referring to 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 representativeness.

The cause of non-representativeness – assumption

21 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.

22 If the selection probability is a function of household size, and household size is not independent of members' demographic characteristics, then the sampling design itself is the source of representation problems. In this case, the sample would not be representative even if a perfectly random household sample and a 100 percent response

rate could be obtained. As for Kish's results, he found that samples obtained by using the grid show close agreement with population data on important demographic characteristics, and he emphasised the relatively low variance of the selection probabilities. His results followed from the fact that the grid was developed in the USA, in the1950s, when household structure showed a high concentration within a small range of household sizes: over 70 percent of households contained two adults. (see Table 7).

Table 7: Household structure, USA, 1957 (Source: Kish, 1965.)

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

23 Our results so far show that representativeness is a function of 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. 26 percent of the households are one-person households in 2001 in Hungary (census data, Hungarian Central Statistical Bureau); that is twice the figure when Kish examined the situation. This difference in itself is so significant that the question arises whether or not one can accept the grid without modification.

The cause of non-representativeness – proof

24 To put these assumptions in a concrete form, the exact connection between the grid's performance and the population household structure needs to be determined. As the required information on the current Hungarian population is not available, we worked with a sample from a large national household survey[3]. The data contain information on the household of each member of the sample, so it can be used as a population for further sampling. In the following, it will be referred as the "pseudopopulation". Table 8 shows age and sex distribution in the pseudopopulation.

Table 8: 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

25 The grid's performance can be tested with the help of this pseudopopulation concerning the age and sex distributions 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 ... 6, l = 1 ... 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 joint distribution of expected sex and age groups 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.

26 Information on the composition of households is also needed: namely the probability of a person number l within a household of size k being male or female, young, middle aged or elderly is required. Let a_{kl} be a 3×2 matrix (k = 1 ... 6, l = 1 ... 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.

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

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

28 Substituting the known parameters, the expected distribution shown in Table 9 is obtained.

Table 9: 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

29 It can be seen that the expected sample differs from the population in sex and age distributions. Firstly, elderly people, especially women, are oversampled.

30 It is worth mentioning, that in the current population of Hungary, a large proportion of one-person households consist of an older female occupant, and a quarter of all households are one-person households, so it can be concluded that it is more likely to select an elderly female in this way than by simple random sampling.

31 Secondly, males appeared to be underrepresented. Our experiences are similar to those obtained from real surveys.

Modification of the Kish Grid

Hungary

32 In this section a modification of the Kish grid is presented. Our intention was to generate a representative or, at least, a more representative expected sample with respect to sex and age. The grid was modified by changing the selection tables. This modification method is not unprecedented in the literature: Kish himself had suggested modifying the tables when needed.

33 There are some aspects worth mentioning at this point. The scope of our present analysis is limited to representativeness according to sex and age. It may, in the future, be useful to take into account the distribution of other characteristics when using the grid. At the same time, the distribution of other characteristics may need checking when using the modified tables. Obviously, improving the sex and age adjustment does not mean that the sample shows agreement with the population with respect to other variables. Change in selection probabilities implied by the modification needs further consideration as well. The variability of the probabilities can result in an increase of the design-weight-based estimation variance.

34 When modifying the grid, all sampling features are fixed; this is, 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 used, the finer probabilities can be, and closer agreement between the sample and the population can be obtained, but, for practical reasons, the number of tables has to be limited)
selection rules of households with 6 members are applied to bigger households

35 The problem is to construct selection tables that yield a sample with close agreement to the pseudopopulation data. The modification can be simplified: instead of determining the tables, it is enough to determine the selection probabilities.

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

$$\sum_{j=1\ldots j} p_{ij} = 1 \quad \forall i$$

$$p_{ij} > 0 \quad \forall i, j \qquad /3/$$

$$p_{ij} = k_{ij}/12 \quad \forall i, j, \text{ where } k_{ij} \text{ integer.}$$

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

38 This raises the question whether or not there is a solution if the limitation on the number of tables did not hold. Apart from the fact that more tables implies increased costs, and that the number of tables is limited by the sample size itself, the theoretical problem is

still 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.

39 Therefore, it is impossible to obtain a perfectly representative sample. Let us pose the following question instead: which selection table yields a sample that is the closest possible to the pseudopopulation. A distance function has to be defined to find the closest solution; the one that minimizes the distance function. Two functions were used, corresponding to two different approaches. The first one is similar to the Pearson chi-square. Equation 4 shows function f to be minimized.

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

40 The idea of using the other distance function comes from weighting, a widely applied method in survey statistics generally used to improve the precision of the estimates. Poststratification is a weighting method that produces a sample in which each stratum is represented in its appropriate proportion. In our case, strata are defined as the six cells of the sex and age group cross-table. 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 poststratification is that in some cases it increases the estimation variances. Increase in variance is a monotonic function of the sum of squared weights. This implies the following approach: to find the selection table that yields a sample with the minimal sum of squared poststratification weights. Equation 5 shows function g to be minimized.

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

where n is the sample size.

41 As mentioned when solving 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 is 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 are to be solved: let us find the minimum value of function f or g, with or without the integer constraint.

42 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 theorem states that a real-valued continuous function on a closed bounded set assumes a maximum and a minimum value. Although the conditions of the Weierstrass theorem, in our case, do hold, determining the minimum is still not a simple mathematical problem. Apart from special cases, nonlinear optimization problems have numerical solutions. The Microsoft Excel Solver package was used to find the minimums. Table 10 contains the results.

Table 10: 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}						
$\Sigma_{j=1..6} p_{ij}=1\ \forall i$, $p_{ij}>0\ \forall i, j$ $p_{ij}=k_{ij}/12\ \forall i, j$, where k_{ij} integer	0,012624653	17,96 17,63	p_{21}	2/3	p_{31}	1/12	p_{41}	3/12	
		14,45 17,73	p_{22}	1/3	p_{32}	1/12	p_{42}	6/12	
		12,14 20,09			p_{33}	10/12	p_{43}	1/12	
							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			
$\Sigma_{j=1..6} p_{ij}=1\ \forall i$, $p_{ij}>0.01\ \forall i, j$	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			

Table 10 (suite)

			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}						
$\Sigma_{j=1..6} p_{ij}=1\ \forall i$, $p_{ij}>0,\ \forall i, j$ $p_{ij}=k_{ij}/12\ \forall i, j$, 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			
$\Sigma_{j=1..6} p_{ij}=1\ \forall i$, $p_{ij}>0.01,\ \forall i, j$	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			

43 Some expected trends can be observed in all four cases. For example, $p_{21} \sim 2/3$, that acts against male underrepresentation that was found when using the Kish grid (since p_{21} is

the selection probability of the first adult in a two-person household, and the first one tends to be male because of the ordering procedure).

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

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

46 Table 11 presents the modified selection table obtained by function f with the integer constraint.

Table 11: 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

Other Countries

47 Since the performance of the grid depends on the household structure of the target population, its modification varies country to country. In the following, those countries are considered in which the Kish grid is used in health surveys or in other surveys. The source for the national datasets was the database of the Luxembourg Income Study (LIS)[4]. The optimal solutions were obtained by optimizing function f with the integer constraint. Table 12 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 worst in Italy, and 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.

Table 12: 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
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

48 The fourth column presents the distance between the pseudopopulation and the expected sample obtained by using the optimal solution (D2). The lists of the countries sorted by D1 and sorted by D2 can be compared. One can see that Russia moved from a middle position into the last one; the improvement in its case was more significant than in the case of other countries.

49 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 porrer 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; the relative improvement obtained through optimization. The distance from the pseudopopulation 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.

Summary

50 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 has been proven that the phenomenon is caused by the sampling method and not by practical problems.

The grid can be successfully modifiable if our aim is to adjust the sample to the population by sex and age.

The problem treated is of international significance. The trends observed in the

Hungarian household structure are global trends. Size of households is currently decreasing, and the proportion of single persons is on the rise.

51 The general lesson of our work is that in cases when the implementation of both address-based samples and population-register samples are feasible, it may be worth considering the above mentioned problems and deciding to choose a population-register sample. Especially in case of determination of sampling guidelines for international surveys, it may be advisable to take into account the considerable variation among countries in availability of sampling frames.

BIBLIOGRAPHY

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

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.

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.

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

Népszámlálás 2001. 2 Részletes adatok. Központi Statisztikai Hivatal, Budapest, 2001.

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.

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

NOTES

1. Some examples: (a) international interviewer surveys: World Health Survey conducted by the WHO in 2002-2003 in more than 70 participating countries. Address-based sampling design and the Kish grid were recommended in the Sampling Guidelines for all the participating countries, see http://www3.who.int/whs/P/SamplingGuidelines.pdf. Further international surveys, in which the Kish grid was used by many countries: International Social Survey Programme, European Social Survey, National Fertility and Family Survey, International Social Justice Project, Los Medios y Mercados de Latinoamérica; (b) national interviewer surveys: Australian national social science survey, Health Survey for England, Scottish Health Survey, Northern Ireland social attitudes survey, ICPE study on mental health (Canada), Oregon Health Behavior Surveys (USA),

Risk Factors for Sleep Bruxism in the General Population (Italy); (c) telephone surveys: the grid is often used in surveys on politics, health, etc. in many countries, such as the USA and Hong Kong.

2. A preliminary, partial version of this paper was published (in Hungarian) as Németh, R, Rudas, T.: Sample selection with the Kish grid, *Statisztikai Szemle*, 80, 309-327.

3. 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.

4. Computations are based on Luxembourg Income Study samples.

ABSTRACTS

The problem of drawing a person from a household often occurs at the final stage of an address-based sample survey design; e.g., in telephone surveys after the households are contacted. The Kish grid gives an algorithm for this random selection. When evaluating the representativeness of samples obtained by this design, researchers often refer to the undersampling of males and overrepresentation of elderly people, the phenomenon originating from the practical realisation of the interview; e.g., males are more difficult to find at home, and less willing to participate. In the paper, some theoretical evidence will be given that explains the representation problems in the case of address-based samples without considering these assumptions. We found that, contrary to the opinion held by some researchers, the grid is not capable of providing representativeness by gender and age. The misconception stems from the fact that when the Kish grid was developed in the 1950's, both randomness and representativeness could be achieved using the method, due to the household structure of the USA. We show that today this does not hold for most countries. Finally, a modification of the Kish grid is suggested that 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 to country. In the paper, those countries are considered where the Kish grid is in use. The main lesson is that in cases when the implementation of both address-based samples and population register samples is feasible, it may be worth considering the above mentioned aspects and deciding to choose a population register sample.

Problèmes de représentativité inhérentes des échantillons basés sur adresses et une modification de la grille de Leslie Kish: Le problème de tirage d'une personne dans un foyer apparait souvent dans la stage finale de l'échantillonnage basé sur adresses, par exemple dans les enquêtes par téléphone apres avoir contacté le foyer. La grille de Kish fournit un algorithème pour cette sélection au hasard. En évaluant la représentativité de tels échantillons, on mentionne souvent la sous-représentation des hommes et la sur-représentation des personnes âgées, attribuant ce phénomène à des problèmes de passation du questionnaire ; les hommes sont plus difficiles à trouver à la maison et sont moins inclins à répondre. Cet article fournit un cadre théorique qui explique ces problèmes de représentation. Contraire aux opinions de beaucoup de chercheurs, nous avons trouvé que la grille ne peut pas fournier une représentation fidèle par âge et sex. Ceci est dû au fait que la grille de Kish a été développé dans les années 1950 aux Etats-Unis quand le hasard et la représentativité étaient garantis par la structure des foyers de l'époque. Nous montrons que ce n'est pas le cas aujourd'hui dans plusieurs pays. En fin, une

modification de la grille de Kish est proposée qui améliore sa performance mais les modifications varient de pays en pays. Des pays où la grille est utilisés sont considérés. Le principal résultat est que dans le cas où des échantillons basés sur adresses et ceux sur des registres de la population sont faisables, il peut valoir la penne de tenir compte de ces questions de représentativité et de décider en faveur de l'utilisation du registre de la population.

INDEX

Mots-clés: Echantillon basé sur adresses, Grille de Kish, Tables de Kish, Echantillon basé sur des registres de la population
Keywords: Address-based Sample, Kish Grid, Kish Tables, Population Register Sample

AUTHOR

RENÁTA NÉMETH

(Hungarian National Center for Epidemiology & Eötvös Loránd University; Hungarian National Center for Epidemiology, H-1966 Budapest Pf. (PO.Box) 64, Hungary; nemethr@oek.antsz.hu)