SAMPLING,
RELIABILITY OF RAW DATA

Balázs Kapitány

SAMPLING

The present study outlines the process of sampling design and describes the major characteristics of the sampling technique applied in the research entitled "Turning points of the Life-course" carried out by the Demographic Research Institute.(HCSO-DRI, Hungary)

As a rule, sampling design is defined by the intentions and possibilities of the research. As the first text of our volume presents the research objective provided exceptional difficulties in our case.

- Upon defining the required number of sampling elements, on the one hand we had to make sure that the incidence of the significant social groups (e.g. unmarried couples, young singles living on their own) regarding the demographic behaviour intended to be analysed, is suitably large enough. On the other hand, we also had to ensure that the number of demographic events (e.g. childbirth , marriage, moving away from home) between subsequent panels in every third year occur at a frequency that is appropriate for analysis.

In the light of these implications, we concluded that the number of necessary successful interviews has to add up to at least 15 000 pieces. In other words, we determined the size of the minimal sample to be 15 000 individuals.

- Due to the nature of panel surveying, it is vital to decrease as much as possible the bias in data arising from either failed or refused interviews. For even if the database is seemingly correctable by weighting at present, the sample would become rather biased by the time of the subsequent survey due three years later. Moreover, following from the nature of longitudinal studies, we had to be prepared for a higher refusal rate.

In order to minimise the number of failed and refused interviews, we had to make alterations in both questioning (for details see the study on data collection in the present volume) and sampling techniques. The most essential question of sampling techniques is whether one should substitute the failed and refused questionnaires, and if yes, how this should be carried out. Certain amount of substitution is inevitable due to the aforementioned characteristics of panel surveying. The traditional method employed by the (Hungarian) Central Statistical Office (HCSO) in these cases is the so called "address (person) substitution" technique. According to this method, one collects "multiple" samples, that is, behind each individual of the sample, there stands a person (or possibly more people) of the same gender, similar age and area of residence (and possibly marital status, qualifications), who may replace the original subject if necessary. This method, however, has a great disadvantage: past experience has shown that if aware of an alternative possibility, interviewers do not do all they can in order to access and/or persuade subjects to participate in the interviews. For this reason, the above method may result in a rather large-scale implicit bias in the sample that is impossible to disclose by primary demographic variables.

Based on the above, we chose rather to distort the address list in advance. In other words, based on earlier sampling experiences, we over-represented the groups where a larger amount of failed and refused interviews was to be expected. By this process one

can expect the composition of the actual database to come closest to the ratios of the population. This procedure is called "dropout" (or decrase) sampling. Nevertheless, successful employment of this method is constrained by the fact that one is unaware of the number of successful interviews beforehand. Its primary advantage over address substitution is that each respondent is regarded as being irreplaceable, so interviewers are unable to filter out subjects providing difficulties.

Even so, it is important to note that no sampling method or subsequent weighting process is able to correct the potential biases arising from the fact that respondents of successful and for whatever reason unsuccessful interviews can differ in numerous features. Consequently, the primary aim on all occasions is to decrease as much as possible the proportion of unsuccessful interviews.

Summary of the main characteristics of this sampling technique: taking the refusal and failure of interviews into account, we used a previously distorted multi-stage sample stratified by region, size of settlement and age. The sample unit was a person and Hungarian citizens born between the 1st of January, 1926 and the 31st of December, 1983 formed the population.

Data was gathered from two distinct sources: data concerning the population was provided by the HCSO and the authentic data of respondents was made available by the National Election Office (NEO). The population was based on the estimated number of inhabitants on the 1st of January, 2000, as this was the most current data at the time. The 2001 improved data by settlements was not yet available and the preliminary results of the 2001 Census came out only in December, 2001.

Unfortunately no data is available on the exact age composition of settlements (this data can be found in the census). For this reason, we had to estimate the number of 18-75 (further broken down into 18-29 and 30-75 year olds) year olds within the total population of the settlement. These calculations were conducted by the Demography and Methodology Section of the HCSO based on the HCSO's 2000, fourth quarterly and the 2001 first quarterly results of the so-called Uniform Population Surveying System (UPSS) database (N=84 000). Larger settlements such as Budapest and the following seventy settlements with the highest population in Hungary all participated in the UPSS database, so one was able to make direct estimations of their distribution by age. In the cases of other, smaller settlements, indirect procedures were implemented: the estimated proportion of age groups was defined on all settlement-size categories by regions and this was uniformly projected onto all settlements belonging to that specific settlement-size category.

According to our experiences, personal data (e.g. citizen has moved away, non-existent address) provided by the NEO were incorrect, so the proportion of unavoidably failed interviews forms 10-20% of the database at present. This is a rather large fraction, yet since the updating of the address list of the 1990 year Census of Population had not yet been carried out and the address list of the 2001 census was not ready yet, these were the most reliable data in Hungary at the time of designing the sample[1]. What is more, the interviews failed for this reason are not randomly distributed but are notably more frequent in certain regions and social groups (e.g. amongst geographically exceptionally

[1] Due to the parliamentary and municipal elections, the database of the National Election Office has also been updated since then.

mobile youth; or amongst the poorest inhabitants). There is also a large number of data that is not entirely faulty, yet provides difficulties in carrying out the interviews. The most characteristic example worth mentioning is the group of those inhabitants who do not live at their declared address of permanent residence, yet whose temporary address can be found out at these addresses. In order to avoid losing these people in our research, we also separately regulated the means by which our interviewers are able access these subjects.

Two stratification variables were taken into account when selecting the settlement sample.

- The first stratifying variable was the regional division of Hungary (West-Transdanubia, Mid-Transdanubia, South-Transdanubia, Central region, North Hungary, North-Lowlands, South-Lowlands). Stratification taking these into account increases the reliability of regional data and thus provides the possibility of high quality analysis if a sufficiently large number of sampling elements are available.
- The second stratification variable was the category of settlement-size. The sample may be broken down into ten settlement-size categories in which the total population approaches a tenth of the overall population of the country. In other words, it means 1 004 322 people according to the data of the time. It is important to note that the outstandingly highly populated city of Budapest appears in the sample broken down into its 23 districts. Consequently, the total number of settlements adds up to 3156.

Belonging to categories 6-10. are 70 independent settlements and the 23 districts of Budapest. These are the so called "self-representing" settlements, for each was included in the sample individually and sampling was carried out within every one of them. However, 3036 settlements belong to categories 1-5. Owing to difficulties in organisation and other problems, this high number makes individual sampling of settlements practically impossible in all cases. The interviewers of the HCSO usually work on single settlements. This is due to the fact that for a number of reasons such as organisation, costs and quality of data, interviewing more than 14 people from a settlement is not considered very successful.

In the light of these considerations, the designation of the settlements themselves was carried out in a way as to fit the ratio of selection probability by region and settlement-size category. In other words, as if all settlements of a settlement-size category of a region were self-representing settlements. Altogether the number of settlements involved in the interviewing was 441 (counting Budapest as a single settlement), of which 370 belong to the five non-self-representing settlement categories.

Table 1
Settlement-size categories by number of inhabitants
(based on the state of affairs on the 1st of January, 2000)[2]

size category	size of settlement (persons)	number of settlements in category (pc)	number of people belonging to category (persons)
1.	-1214	1919	1 003 218
2.	1215-2376	585	1 004 887
3.	2377-4353	320	1 003 983
4.	4354-9182	161	1 002 245
5.	9183-17816	78	993 484
6.	17 817-32 169	42	1 007 243
7.	32 170-63 337	21	956 499
8.	63 338-85 877	14	1 023 488
9.	85 878-12 5941	9	950 793
10.	12 5942-20 3648	7	1 097 384

By selecting the discussed sampling process, in addition to the cases of failures we naturally also had to estimate the distorting effect of the refusals. Previous research suggests that the refusal of interviews is related to age (it is higher amongst the young[3]), type of settlement (it is higher in larger settlements), gender of the respondent (higher amongst males) and many other factors as well.

We had the chance of defining a different proportion of address selection for age and settlement groups. We thus had the opportunity to take into consideration not only the frequency of failed interviews, but also that of refused ones. Table 2 presents the number of addresses necessary for carrying out an indisputably successful interview according to our estimations. The calculations include both groups of possible failures.

Table 2
Safety coefficient of the sampling procedure

Settlement-size category	Respondents between 19-29 years	Respondents between 30-75 years
1.	1,44	1,2
2-3.	1,56	1,3
4-5.	1,68	1,4
6.	2,16	1,8
7-8.	2,4	1,8
9-10[4]	2,4	2

The estimated population is 7 454 196 people[5]. The selection ratio is 15 000 / 7454 196, that is roughly 0,2%. If the theoretically obtained precise selection ratio is multiplied by the values present in Table 2, the real selection ratio necessary by settlements-size

[2] Source: Mihályffy 2001
[3] These relations are of course not as simple and neither are they linear.
[4] We included a few of the under-populated inner districts of Budapest in this group, where the number of successful interviews has traditionally been low (districts I., VI., VIII., XII.). The other districts that usually caused problems belonged to this group originally due to their large number of inhabitants.
[5] According to the results of the 2001 census, the population is 7 618 280 people. The difference comprises the improved data as well as the projection of the divergence of the census results by some 200 000 inhabitants onto the population.

categories and age groups is calculated. Accordingly, the number of utilised addresses was 25 510, of which 7247 belonged to the category of 18-29 year olds.

The organisation of the steps of sampling and questioning processes were all aimed at making the results of the research as reliable as possible, even in the social subgroups that are usually difficult to approach for carrying out interviews. Controlling the number of non-self-representational settlements with a considerable Roma population to be sufficiently represented in the sample settlements according to their proportions, also serves this goal[6].

We achieved finding at least a part of the respondents not living at their declared address of permanent residence by the so called address-tracking technique. Consequently, we believe that we managed to include the young, mobile stratum into the research better than is generally done. (Details of the questioning techniques are elaborated on in another study of the present volume.)

In the light of these considerations, we requested 25 510 addresses from the NEO, hoping that according to our estimations, we shall subsequently acquire 15 200 successful interviews.

RELIABILITY OF RAW DATA

The actual data was the following: 95,7% (24 417 addresses) of the requested 25 510 addresses were distributed among the interviewers, who filled out "address cards" for 24 138 people (94,7%). The proportion of failed and refused interviews compared to all the addresses is 30,4%, while the proportion of completed interviews is 64,3%, that is, 16 394 pieces. (In other words, the proportion of successful interviews is 67,9% of the addresses visited by interviewers.) Upon interpreting the results, one has to keep in mind the prior distortion in favour of the more problematically accessible groups that in itself had already rendered interviewing more difficult.

Although the useful amount of data decreased to a certain extent after inspection, the 16 394 successful interviews did not only meet our expectations but even slightly exceeded them. The level based on previous estimations was not met in four counties. The county of Borsod-Abaúj-Zemplén presented the greatest difference compared to our expectations, as the expected 1338 completed interviews came short by 92 pieces. The county of Fejér highly surpassed our expectations for the expected 589 successful interviews were exceeded by 213 pieces. It is important to note, however, that these dissimilarities may not only be consequences of the individual performances of the counties, but can also be explained by divergence in the levels of probability calculated depending on the sizes of the settlements and the proportion of young people

Nevertheless, the truly vital question remains how successful the estimation of the previously distorted addresses had been. In other words, to what extent do the results approximate actual distribution. The best way to answer this question is by comparing the whole database to the final results of the 2001 Census. At first glance, comparing the distribution of the sample population (1st of January, 2001) with the results may

[6] We made our estimations on the proportion of the Roma population based on Kertesi-Kézdi (1998).

seem more logical. By doing so, however, one would compare the results to a database accepted because of technical constraints, one that has meanwhile turned out to be imprecise. The fact that by employing the sample population one would have to renounce comparison of certain characteristics (e.g. qualification) for which there is no improved data available, provides a further disadvantage. For these reasons, we have chosen to make our calculations with the more accurate census results that are also closer to the ideal date of our research, when interpreting the raw, unweighted distribution of the results. Naturally, slight imprecision is unavoidable even when employing the census data for almost a year has passed between the census and our data collection, during which the population distribution had undergone some alterations.

The distribution by gender in the database is 45,4% male, 54,6% female. Although there is a higher number of women (52,4%) in the population too, due to the higher mortality of men, this ratio has been further biased in our sample as it is more difficult to get hold of and making male respondents speak. Comparison by age groups provides the following results. (Table 3)

Table 3
Composition of database by generations

	1926-1941	**1942-1951**	**1952-1961**	**1962-1971**	**1972-1983**	**Total**
census 2001	1 574 583	1 346 282	1 510 999	1 323 724	1 862 692	7 618 280
census %	20,7 %	17,7 %	19,8 %	17,4 %	24,5 %	100 %
Turning points of life-course	3495	2882	3091	2605	4321	16 394
Turning points of life-course %	21,3 %	17,6 %	18,9 %	15,9 %	26,4 %	100 %

It is striking that the intentionally over-represented proportion of young people during the process of sampling approximated, even exceeded by 2% their proportion in census. This excess, however, does not evenly set the shortage present in the other age groups off, but is focused around the young middle-aged people. It seems that the slight over-representation of this age group (approximately between 30 and 40 year olds) during the sampling process would have been justified.

Considering the composition by both gender and age simultaneously, one may declare that the young middle-aged are most under-represented in the database, since the proportion of this group would ideally be 8,7%, whereas it is only 7,4% in our sample.

We have the chance of comparing the population and the raw database with regards to marital status, as well as age and marital status. Distribution of respondents by marital status was as follows: unmarried, single 25,2%, married 56,7%, divorced 9%, widowed 9,1%. Census distribution was as follows: unmarried, single 25,7%, married 55,7%, divorced 9,9%, widowed 8,7%. Considering the ratios, it is apparent that the group least accessible and most difficult to interview was that of the divorced, yet the divergence projected onto the category is merely 10 percent in this case too. There is fortunately no such concentration in the bias of the pattern of marital status by age group as in the case of young middle aged men.

Considering qualification, 13,5% had obtained a degree, while this proportion was 12,6% according to the census results. We calculated high school graduates to be 31,9% compared to the 28,6% of the population. One possible reason responsible for the perceived divergence may be that a school year had come to an end between the ideal date of the census and the collection of data, so many members of the sample could have finished high school or obtained a degree during that this time period. (The study on weighting elaborates on this problem and its solution in detail.) Nevertheless, it is also possible that people with higher qualifications were more easily made to talk. Nonetheless, this approximation may be considered rather good compared to results of other researches.

As for qualification and age group the picture may be further specified. It can be observed that the raw database maps the proportion of those with a higher degree into every age group rather precisely. Divergence is present between high school graduates and non-high school graduates, to the benefit of the former. This bias is further amplified by the fact that the number of middle aged people is generally lower in the sample. Consequently, the non-high school graduates between 30 and 50 years forms 21% of the population, while only 17,9% of the raw database.

It is of course impossible to discuss all potential combinations of fitting from every possible viewpoint. Even if one tries, it is still impossible to state with utmost certainty that the goodness of fit is just as fine in the cases of the incomparable aspects. However, the study of a sub-sample is necessary, namely, that of the population of the capital.

It is common knowledge that refusal of interviews and people not living under their declared address of permanent residence is exceedingly high in the capital (especially in the inner districts and the rich mountainous regions of Budapest) is a widely known (e.g. Waffenschmidt 2001) problem that has been worsening since the change of regime in 1989. This circumstance severely endangers the reliability of the research data acquired in Budapest. We had therefore made our calculations with the lowest ratio of successful interviews in Budapest when designing the “dropout” sampling procedure. Our estimations have turned out to be more-or-less precise, since the proportion of the population of the capital met our expectations (17,5% inhabitants of Budapest were represented in the raw sample, while we had previously calculated 18%). The precision of the internal ratios, however, remains questionable.

Table 4 presents the distribution of the population of Budapest by gender, age and qualification based on the results of the census on the one hand, and on the raw database on the other hand.

The comparison reveals that the raw database fits more precisely in the case of women than in that of men. Only two groups of the former show noteworthy divergence: more of the older women of Budapest with a higher degree were represented in the sample compared to their number in the population, whereas less of the young non-high school graduate women were represented. Similar facts may be observed in the case of male respondents. Namely, the group of young uneducated men shows shortage, while the non-middle aged uneducated group shows excess compared to their number in the population . Under-representation may again be a result of the fact that the qualifications of the population may have grown in number (e.g. a cohort may have

graduated from high school) during the time between the ideal dates of the census and the collection of data.

All aspects considered, it can be maintained that the raw data aptly approximates the distribution of the population, and that precise fitting of the marginal distributions may be attained without any difficulties with the help of weighting in the major dimensions. The "dropout" sampling procedure distorted beforehand has lived up to our expectations. The sample has sufficiently mapped the distribution of the population with special regard to the dimensions under study.

Table 4
Distribution of the population of Budapest according to generations and qualification by gender

Gender	Qualification	Database	**1926-1941**	**1942-1951**	**1952-1961**	**1962-1971**	**1972-1983**	Total
Male	**higher education**	**Census %**	2,6	2,5	2,2	2,1	1,5	11,0
		Turning points of life-course %	*3,5*	*3,3*	*2,7*	*2,2*	*3,9*	*15,6*
	high school graduation	**Census %**	2,1	2,6	2,5	2,8	5,8	15,7
		Turning points of life-course %	*2,9*	*2,3*	*2,2*	*2,5*	*6,9*	*16,9*
	lower than high school	**Census %**	3,8	3,3	3,4	3,4	5,6	19,5
		Turning points of life-course %	*3,5*	*3,3*	*2,7*	*2,2*	*3,9*	*15,6*
female	**higher education**	**Census %**	1,7	2,4	2,6	2,5	2,0	11,1
		Turning points of life-course %	*2,6*	*3,0*	*3,4*	*2,8*	*2,4*	*14,3*
	high school graduation	**Census %**	3,8	4,6	4,1	3,7	6,8	23,0
		Turning points of life-course %	*4,2*	*4,7*	*3,9*	*3,8*	*7,4*	*23,9*
	lower than high school	**Census %**	7,2	3,4	2,8	2,3	4,0	19,7
		Turning points of life-course %	*7,4*	*3,6*	*2,7*	*1,4*	*2,9*	*18,0*
Total		**Census %**	21,1	18,8	17,7	16,7	25,7	100,0
		Turning points of life-course %	*23,6*	*20,0*	*16,8*	*14,4*	*25,1*	*100,0*

References

Kertesi, Gábor-Gábor Kézdi (1998) *A cigány népesség Magyarországon. Dokumentáció és adattár.* Budapest: Socio-typo.

Mihályffy, László (unknown date of publication) A *Demográfiai Panel mintájának terve.* Manuscript.

Waffenschmidt, Jánosné (2001) *Adatgyűjtés és az adatok minősége.* Statisztikai Szemle, 2001/9:741-751.