Methodological documentation of the educational microdata set from the 2022 Census[1]

HCSO aims to continuously follow new user needs and meet them as quickly and as fully as possible. Today, there is a growing demand for microdata sets from our records. The 'raw' disclosure of such datasets is not possible for data protection reasons, so either the scope of users and the place of use must be very accurately defined and then the research results must be subject to a data protection controls, or the microdata sets must be anonymised to such an extent that they do not allow the identification of our data providers or the disclosure of new information about them, despite their wide availability.

Since 2007, HCSO has been providing research opportunities for scientific purposes on detailed microdata sets that are not suitable for direct identification in the central research room of HCSO. A research plan and scientific background are required to start the research. HCSO may charge a fee for sets prepared for individual needs[2]. A 10% sample of the 2011 and 2022 census population is available for researchers in the research room environment.

As the first step towards opening up to a wider range of users, we published the *test sets*[3] from the 2011 Census in March 2014 and the *test sets* from the 2022 Census in 2024. Although these sets can freely be downloaded by anyone, the target users are researchers who wish to analyse the census sets in a secure environment (research room access, remote execution). The test sets are intended to facilitate the preliminary writing and testing of program sets and are not suitable for analysis or research purposes.

1. The census microdata set for educational purposes

The **educational** microdata set can primarily be utilised by educational institutions during statistical courses, but it is also recommended for those who wish to learn about statistical methods in practice.

The education data set from the 2022 Census contains

- 10,000 records - a selection rate of about 1‰ (one thousandth) - and

- 13 variables[3].

The definitions of the variables and the nomenclatures can be found in the file ' *Rekordleírás_Népszámlálás 2022_Oktatási mikroadat-állomány.xlsxmikroadat-állomány '*.

The level of detail of the census microdata file for educational purposes is significantly lower than the 10% sample of persons and dwellings that can be searched in a secure environment in the research room of HCSO. However, a major advantage of the microdata file for educational purposes is that it is easier to access: after accepting the terms of use, anyone can freely download it from HCSO website.

2. Sampling

[1] A file consisting of a series of records, which contains data on observation units (here: persons).
[2] For details, see https://www.ksh.hu/kutatoknak_kutatoszoba
[3] 10 variables are personal characteristics and 3 variables are housing characteristics.

The personal sample of 10,000 elements is a probability-based sample selected from the total census population (9,603,634 persons). The persons were selected without stratification, in a single-stage, simple randomisation without replacement, using the SAS *survey select* procedure. Thanks to the simple selection method, the accuracy of the estimates calculated from the sample can easily be characterised using the well-known formula standard deviation.

3. Data protection

HCSO is deeply committed to the protection of individual data collected from data providers. Our data disclosures are made in accordance with Act CLV of 2016 on Statistics, Government Decree 184/2017 (5 July 2017) on its implementation, and Act CXII of 2011 on the Right of Informational Self-Determination and the Freedom of Information. In order to ensure that the persons in the microdata set cannot be identified, we have taken the following measures:

- We applied a *low selection rate* (around 1‰). If a user finds a record in the set that matches himself or a person known to him, based on certain characteristics, then the probability that he has actually identified himself or the person known to him is negligible. The reason for this is that there are at least three occurrences of the keys that allow identification in the population. Therefore, it is not at all certain that the person being identified is the one the user thinks and not another person with similar characteristics.
- *The set contains no direct identifier.*
- Variables that qualify as *indirect identifiers*[4] (e.g. exact date of birth, municipality of residence, age) or that fall under the definition of *specific data*[5] (e.g. religion, disability, nationality) *have been removed* or are provided in *a less detailed form* compared to the specifications in the questionnaire (e.g. the place of residence is only shown at NUTS2 region level).

4. Limitations of the use of the census microdata set

Since the census microdata set for educational purposes is a *simple random* sample of about 1‰, the estimates made from the set (e.g. average educational attainment) are subject to *sampling error*. When interpreting the results of analyses or drawing conclusions on the population characteristics, one should always pay attention to the error resulting from the sampling.

Although univariate distributions reproduce the values observed in the population with a fairly high degree of accuracy[6], if one is researching the responses obtained from the Census for scientific purposes, we suggest that access in a secure environment is considered. In this case, as we mentioned earlier, we can provide much more detailed data (10% sample of the personal, housing, family and household population, instead of non-aggregated variables - e.g. age group -, full set of variables corresponding to the questionnaires).

[4] A variable of a microdata set that can help identify statistical units as a key part.
[5] For the definition of special data, see Article 3, Section 3 in Act CXII of 2011.
[6] See Appendix 1.

Appendix: Representativeness of the census microdata set for educational purposes

For illustrative purposes, we present the univariate distributions derived from the microdata set and compare them with the distribution characteristics calculated from the full census population[7].

- Region (REGION)

REGION	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
11	17,4	17,5	0,1
12	13,6	13,9	0,3
20	11,0	11,0	0,0
30	10,4	10,2	0,2
40	8,8	8,9	0,1
50	11,6	11,4	0,2
60	14,9	14,6	0,3
70	12,3	12,5	0,2

- Sex (SEX)

SEX	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	48,3	48,1	0,2
2	51,7	51,9	0,2

- Age group (AGE GROUP)

AGE GROUP	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	9,4	9,7	0,3
2	10,1	9,9	0,2
3	12,0	11,2	0,8
4	12,2	12,9	0,7
5	16,8	16,4	0,4
6	13,4	13,5	0,1
7	12,5	12,7	0,2
8	9,0	9,3	0,3
9	4,6	4,5	0,1

- Family status (FAMILY STATUS)

FAMILY STATUS	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	44,3	44,1	0,2
2	36,9	37,1	0,2
3	8,4	8,4	0,0
4	10,4	10,5	0,1

[7] The sum of the distributions is not always exactly 100% due to rounding.

- Number of children born alive (CHILDREN BORN ALIVE)

CHILDREN BORN ALIVE	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
0	43,0	43,0	0,0
1	17,9	18,1	0,2
2	27,1	27,0	0,1
3	8,6	8,8	0,2
4	2,1	2,1	0,0
5	0,8	0,6	0,2
6	0,4	0,4	0,0

- Do you go to school? (STUDIES)

STUDIES	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
0	83,6	84,0	0,4
1	16,4	16,0	0,4

- Educational attainment (EDUCATION)

EDUCATION	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	15,9	16,0	0,1
2	18,8	18,6	0,2
3	17,3	18,1	0,8
4	28,6	28,3	0,3
5	19,4	19,1	0,3

- Economic activity (ECONOMIC ACTIVITY)

ECONOMIC ACTIVITY	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	49,3	49,1	0,2
2	2,5	2,5	0,0
3	23,5	23,6	0,1
4	24,8	24,8	0,0

- Current occupation, job title (OCCUPATION)

OCCUPATION	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	12,1	12,0	0,1
2	12,0	12,0	0,0
3	7,1	6,9	0,2
4	1,1	1,1	0,0
5	5,8	5,9	0,1
6	11,2	11,2	0,0
	50,7	50,9	0,2

- Family composition (FAMILYCOMPOSITION)

FAMILY	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	24,8	25,0	0,2
2	48,0	48,2	0,2
3	12,9	12,7	0,2
4	14,4	14,0	0,4

- Construction year of the dwelling (CONSTRUCTION YEAR)

CONSTRUCTION YEAR	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	13,5	14,2	0,7
2	11,3	11,0	0,3
3	33,0	33,3	0,3
4	24,1	24,1	0,0
5	10,3	9,9	0,4
6	6,0	5,7	0,3
	1,7	1,7	0,0

- Floor area of the dwelling (FLOOR AREA OF DWELLING)

FLOOR AREA	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	4,2	4,6	0,4
2	6,0	6,0	0,0
3	13,1	13,0	0,1
4	20,4	19,9	0,5
5	20,5	20,5	0,0
6	34,2	34,2	0,0
	1,7	1,7	0,0

- Number of persons living in the dwelling (NUMBER OF PERSONS IN DWELLING)

NR OF PERSONS	Distribution of educational file, %	Distribution of population file, %	Difference, percentage point
1	14,8	14,8	0,0
2	22,7	23,4	0,7
3	21,3	20,8	0,5
4	19,5	19,5	0,0
5	10,5	10,5	0,0
6	9,4	9,3	0,1
	1,7	1,7	0

If you have any questions or comments about the use of the microdata file and the information provided about it, please contact us.

Methodology Department
6 June 2024