Cohort Fertility and Education Database
Methods Protocol

Kryštof Zeman[1], Zuzanna Brzozowska, Tomáš Sobotka, Éva Beaujouan, Anna Matysiak

Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), Vienna Institute of Demography/Austrian Academy of Sciences

last update 20-09-2017

1. Introduction

The Cohort Fertility and Education (CFE) database provides a free access to aggregate data on completed cohort fertility and parity distribution by level of education. The data come from censuses and large-scale surveys conducted in countries with generally high levels of education and relatively low fertility. As the database covers women and men who have completed or almost completed their family building, the data focus on people aged between 40 and 80. Most datasets come from the 2010 census round, corresponding to birth cohorts 1930 to 1970. For many countries we have collected data from the 2000 and 1990 census rounds; for some countries the data are available from even earlier censuses, encompassing the late 19th century birth cohorts.

[1] krystof.zeman@oeaw.ac.at

2. Database structure & use in a nutshell

The database is available online at www.cfe-database.org and is freely accessible without registration. The data are structured by country and survey (one has to choose first the country and then the survey). Having selected the survey of interest, the users can access a range of indicators, which are generated automatically using the formulas specified in section 5 of this protocol. Specifically, the database features:

- the completed fertility rate (CFR), specified by birth order (ranging from birth order 1 to 8+),
- the absolute and relative distribution of respondents by the number of children ever born (including childless respondents),
- parity progression ratios (PPR).

All indicators are specified by the highest level of education achieved by the time of the given census or survey and, when available, by sex and country of birth or citizenship. The source data are available by clicking the ***Get input data as CSV*** button. They contain the absolute number of women (and, if available, men) by level of education, birth cohort, the number of children ever born and, if available, country of birth or citizenship (distinguishing between Native and Foreign). These source data are not standardised and include the unknown categories (unknown number of children, birth cohort, level of education and country of birth). For each country a brief description of each survey or census as well as the country's education system is provided in the Country Documentation file.

The CFE database supports enhanced possibilities to explore and visualise data. Users can download a CSV file with the source data, but they can also compile and download tables defined by themselves (with a self-specified aggregation level, self-selected indicators and certain cohorts/education categories filtered out), by clicking the ***Get data & indicators as CSV*** button. In addition, the database allows users to generate interactive graphs with selected indicators for each country and survey, which provide first-hand visualisation of inter-cohort fertility trends, patterns and differences.

3. File system and labelling of input (source) data

Each dataset in the CFE database is provided as a source data file containing absolute numbers of respondents by parity, tabulated further by other characteristics. The filename is constructed as `COUNTRY_SURVEY YEAR.csv` where:

- `COUNTRY` is the name of the country
- `SURVEY` is the type of survey (usually Census, Survey or Register)
- `YEAR` is year of census or survey

For example, the file containing the 2001 census data for Austria is labelled `Austria_Census 2001.csv`.

The data are stored as comma-separated values (CSV) files in a long format. In the statistical package R the following command can be used to import the data:

```
read.csv('Austria_Census 2001.csv', header = TRUE)
```

The first line contains a header, with the following categories listed:

```
country, data_source, cohort_from, cohort_to, edu_eurrep, edu_from, edu_to, sex,origin, stat, value
```

They include the following information:

- `country` – country name;
- `data_source` – simple label of data source including the year of survey/census (e.g. Census 2001);
- `cohort_from, cohort_to` – respondents' birth cohort range (e.g. 1946-1950); these two columns are equal when one-year birth cohorts are displayed; unknown cohorts are labelled as -1;
- `edu_eurrep, edu_from, edu_to` – education coding, described in section 4;
- `sex` – `F` for women, `M` for men (when included);
- `origin` – `Native` for respondents born in the country or with the country's citizenship, `Foreign` for respondents born in another country or with a foreign citizenship, Total when the origin is not distinguished;
- `stat` – name of the indicator listed in the next column, labelled `value`:
    - `women_total / men_total` – total number of respondents;
    - `children_total` – the total number of children ever born to all respondents;

- parity_0 – the number of childless respondents;
- parity_i (i=1,2…,i_max) – number of respondents with i children; the maximum-parity category i_max differs across surveys and includes all respondents with a number of children higher than i (i+); i_max should not be higher than 20;
- parity_unknown – number of respondents for whom the number of children is not known;

- value – number of cases.

The database lists all data by cohort, education and sex; and sometimes (when available) also by place of origin. Totals can always be computed as a sum of all specified cases.

The source data are extracted directly from the survey or census records, with very few, if any, computations.

4. Coding of education

The highest education attained (the edu_eurrep or edu column in the CSV files) is coded according to the International Standard Classification of Education 1997 (ISCED97). There are 7 broad levels of education, ranging from 0 to 6, and further categorised into C/B/A sublevels (see Table 1). In the documentation files for each country there is a table translating the original educational categories (as given in the questionnaire) to the ISCED97 levels.

Table 1 The 1997 ISCED levels and their short description

ISCED level	Description
0	Early childhood education; no education
1	Primary education
2C-2A	Lower-secondary education (Second stage of basic education)
3C	Upper-secondary education: programmes designed to lead directly to labour market, not to ISCED 5A or 5B (e.g. Apprenticeship, Secondary technical school)
3B	Upper-secondary education: programmes designed to provide direct access to ISCED 5B (e.g. Vocational training)
3A	Upper-secondary education: programmes designed to provide direct access to ISCED 5A (e.g. Higher general secondary/Higher technical secondary school)
4B	Post-secondary non-tertiary education not giving access to level 5 (primarily designed for direct labour market entry, e.g. Schools for medical services, Schools for nursing)
4A	Post-secondary non-tertiary education programmes that prepare for entry to ISCED 5 (e.g. follow-up courses, language schools)
5B	First stage of tertiary education (not leading directly to an advanced research qualification; e.g. Higher technical school, polytechnic)
5A	Bachelor's, Master's or equivalent level
6	Doctoral or equivalent level

For detailed information on ISCED 1997 visit http://uis.unesco.org/en/topic/international-standard-classification-education-isced

Users can aggregate the educational levels manually or they can use the three- or four-level EURREP classification. The four-level classification is specified as follows:

Level 1: lowest compulsory education; encompasses ISCED levels 0, 1 and 2; also labelled basic or primary education;

Level 2: usually ISCED3C[2];

Level 3: upper-secondary and post-secondary education; includes ISCED3A, ISCED3B and ISCED4 levels;

Level 4: ISCED5 and ISCED6 levels; tertiary or university education.

The three-level EURREP classification merges all ISCED3 (3A, 3B and 3C) and ISCED4 levels into one category; other levels remain the same as in the four-level EURREP classification. In some cases (e.g. Switzerland), the Medium-low and Medium-high cannot be distinguished, and only the three-level EURREP classification is specified (Level 1, 2, 3).

5. Computing indicators

This section specifies all the indicators provided in the CFE database and the equations used to compute them.

- CFR is the completed fertility rate;
- $SHARE$ is the share of women/men with i children, including the childless and ranging up to the maximum open-ended group i_max, which also includes all women/men with a higher number of children;
- PPR is the parity progression ratio.

Generally, the highest parity/birth order group considered in the CFR and $SHARE$ indicators is an open-ended group 8+, if not explicitly stated otherwise (for instance, when the original data contain a narrower set of available categories). For the parity progression ratios, the highest progression rate computed is from 6th to 7th child ($PPR67$), unless specified otherwise.

The CFE database uses the following notations:

c...birth cohort

[2] In some countries a different categorisation is used to reflect the underlying differences in educational systems and to achieve maximum comparability. For example for Austria ISCED3B (apprenticeship and vocational training) is included into the Level 2, while for the Czech Republic ISCED3B is included into the Level 3.

e...educational level

o...country of origin/citizenship

s…sex

i...birth order of child or parity of the mother

i_{max}...highest birth order considered (used as an open interval i_{max+})

$parity\_i$...number of respondents with i children

$women\_total\ (men\_total)$...total number of respondents

$children\_total$...total number of children

$SHARE_i$...proportion of respondents with i children

CFR_i...completed fertility rate for birth order i

$PPR_{i-1,i}$...parity progression ratio; conditional probability of having another (i^{th}) child among respondents achieving parity $i-1$

For each cohort, educational level, origin and sex (not shown in the formulas below), the following indicators are computed:

Completed fertility rate by birth order:

$$CFR_i = \frac{\sum_i parity\_i}{women\_total} \quad (1)$$

Completed fertility rate for all women/men:

$$CFR = \frac{children\_total}{women\_total} \quad (2)$$

The share of women/men by the number of children ever born (summing up to 1):

$$SHARE_i = \frac{parity\_i}{women\_total} \quad (3)$$

The parity progression ratio to first birth and to higher-order births:

$$PPR_{0,1} = CFR_1 \quad (4)$$

$$PPR_{i-1,i} = \frac{CFR_i}{CFR_{i-1}}, \text{ for } i>1 \quad (5)$$

Indicators for the highest birth order/parity i_{max} are computed as open intervals in the case of CFR_{8+} and $SHARE_{8+}$ and as a single-parity progression rate in the case of $PPR_{7,8}$ (if CFR_8 is available):

$$CFR_{8+} = CFR - \sum_{i=1}^{7} CFR_i \tag{6}$$

$$SHARE_{8+} = \frac{parity\_8p}{women\_total} \tag{7}$$

$$PPR_{7,8} = \frac{CFR_8}{CFR_7} \tag{8}$$

In most cases respondents with an unknown number of children are disregarded in computations of all fertility indicators:

$$women\_total \sim women\_total - parity\_unknown \tag{9}$$

In practice this recalculation gives the same result as if these respondents were redistributed proportionally to the parity distribution of respondents for whom the number of children ever born has been recorded. In special cases, where there is a good reason to presume that respondents with unknown number of children are in fact childless (e.g. in the censuses for the Czech Republic and Slovakia), the number of respondents with an unknown number of children has been added to the number of childless respondents:

$$parity\_0 \sim parity\_0 + parity\_unknown \tag{10}$$

These country or survey-specific computations are specified in country documentation files.

6. User defined output data

The user defined output data can be downloaded as comma-separated values (CSV) files, which you can easily read using the following command in the statistical package R:

```
read.csv('file_name.csv', header = TRUE)
```

If you open the file with Calc within the LibreOffice office suite, you should use the option 'separated with comma'; if you open it with Microsoft Office Excel and your operating system uses a delimiter other than comma (e.g. tabulation or semi-colon), mark the column with entries, go to *Data* on the ribbon, click *Text to Columns*, choose *Delimited*, press *Next*, tick off *Comma* as a delimiter, press *Next* and then *Finish*. Here you can find a PowerPoint presentation showing the procedure.

The first line contains a header, with the following categories listed:

- `cohort` – respondent's birth cohort either as single-year or five-year cohorts;
- `edu` – education coded as described in section 4;
- `sex` – `F` for women, `M` for men (when included);
- `origin` – Native for respondents born in the country or with the country's citizenship, Foreign for respondents born in another country or with a foreign citizenship, Total when Native/Foreign not distinguished (most countries);
- `women_total` – the total number of women (sex `F`) /men (sex `M`);
- `children_total` – the total number of children ever born to all respondents in a given cohort and educational group (and of a given origin status);
- `parity_0` – the number of childless women/men;
- `parity_i (i=1,2…,i_max)` – the number of respondents with `i` children; the maximum-parity group `i_max` differs across surveys and includes all respondents with a higher number of children `i+`;
- `parity_unknown` – number of respondents for whom the number of children is not known (not for all countries);
- `CFR, CFRi (i=1,2…,i_max)` - the completed fertility rate, total and by birth order;
- `SHAREi (i=0, 1,2…,i_max)` – proportion of women/men by parity *i*;
- `PPRi-1i (i=1,2…,i_max)` – parity progression ratio (see section 5 for more details).

7. Terms of use

The data in the Cohort Fertility and Education database are provided free of charge to all interested users under the CC BY+SA 4.0 license. We kindly ask you to comply with the following requests and restrictions:

1. CFE data are not to be used for commercial purposes. The user may copy and retain data solely for scholarly, educational or research purposes or for personal use. The original tabulations of absolute number of respondents by cohort, education and number of children ever born should not be used for commercial gain or re-published in any form without the explicit permission of the data owners.
2. In all published work and presentations, please acknowledge the Cohort Fertility and Education database as either the source or the intermediary of the data. When downloading the data, you should note the date for future reference.

When using the database, please cite the full name of the database (Cohort Fertility and Education database, CFE database) and refer to:

Zeman, K., Z. Brzozowska, T. Sobotka, E. Beaujouan, A. Matysiak. 2017. *Cohort Fertility and Education database. Methods Protocol.* Available at www.cfe-database.org (accessed on [date]).

3. Please do not pass your copy of these data to other users. Rather refer them to the www.cfe-database.org website, where they can download the same data for themselves. Since these data are updated on a regular basis (including corrections if needed), this practice helps to prevent the circulation of multiple outdated or incorrect versions. It also ensures that each user has full access to complete information about the data, methodology, citation procedures, etc.
4. All data contained in the Cohort Fertility and Education database have been carefully checked to avoid errors. However, the database accepts no responsibility for any harm suffered by the user as a result of using these data, even if such harm results from errors on our part. We ask you kindly to inform us of any errors that you may identify, so that they can be corrected in a timely fashion at krystof.zeman@oeaw.ac.at

8. Acknowledgments

The Cohort Fertility and Education database has been developed as a part of the EURREP project, funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n° 284238.

We are grateful for the support we have received from individuals and institutions providing the data or agreeing with their use. You find the list of our collaborators in the ACKNOWLEDGMENTS section of the database at
www.cfe-database.org/about/acknowledgments