Formats of HFD Data Files

Contents

Introduction

The raw data on births and the female population that have been collected from data providers are transformed into files of a standard HFD format. These data files, together with documents that describe these data and their sources, constitute the Input Database. The input data files are posted in the tab entitled "Input Data" on the HFD country page. The structure of these files is explained in section 2 below.

The data stored in the input data files are further processed using the standard HFD methodology (for more details about data processing in the HFD, see the Methods Protocol). The processing of the data produces a wide range of fertility and female population estimates. All these data are stored in separate files, which together make up the Output Database. The output data files are displayed on the HFD website in three tabs on the country page: "Summary Indicators", "Age-Specific Data", and "Fertility Tables". The following section provides more information on the format of these files.

1. Output Data

Output data files are space-delimited text (ASCII) files. The data dimensions are age, birth order, calendar year (period), and/or year of mother's birth (birth cohort).

Live births, female population exposure, fertility rates, and fertility tables are provided for the age range ≤12 (denoted as 12- in the data files) to 55+ by single years of age. Live births, fertility rates, and fertility tables are provided for birth orders one to 5+. Parity-specific female exposure to risk is provided for parities zero to 4+ by single years of age. Total fertility rates (including the completed cohort fertility), mean ages at birth, and standard deviation in mean age at birth are computed for the whole range of reproductive years, and by age 40.

Birth counts are provided with two decimal places, and should be interpreted as estimates of real integer values. Non-integer values are produced by transforming the raw birth data—which originally varied in terms of Lexis shapes, age ranges, birth orders, specification of unknown ages and birth orders, etc.—into Lexis triangles (i.e., in an age-year-year dimension). From these data, births and rates by Lexis squares, and vertical and horizontal parallelograms are computed (for further details, see the Methods Protocol).

1

Missing values in the output files are coded as dots ('.'). The first two lines of each output file contain information about data included in the file and the date of the last modification. The header is on the third row.

The names of the output data files are provided in Appendix 1, in the order the files appear in the three output data tabs on the country page. 'XXX' in the file name stands for a standard (uppercase) ISO code, representing a country or area name (AUT for Austria, CAN for Canada, etc.). For countries, the ISO 3166-1 alpha-3 codes are used (a freely accessible list is available at, for example, Wikipedia: http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3). For subpopulations or areas not corresponding to an ISO country, the ISO 3166-2 region codes are used (where available) as an extension of the country code. In cases in which a code of this kind is not available, such as for Eastern and Western Germany, the code used and its description must be defined in the Background and Documentation file. The country, area, and subpopulation codes used in the HFD are usually identical to the codes used in the Human Mortality Database (HMD) (www.mortality.org).

To facilitate rapid downloads of large amounts of HFD output data, two series of zipped data files are offered to HFD users (see "Zipped Data Files" on the Main Menu). For users who only want information of a given data type (e.g., birth counts) for all countries, the zipped files "By data type" are recommended. In addition to the data that can be found in a regular output file of the same type, such as a file of a single country, there is an extra column showing a country ISO code. This makes it possible to identify the country to which the data refer. For users who want to obtain all of the available data for an individual country or for all countries, the zipped data files labeled "By country" are recommended.

2. Input Data

The data block "Input Data" stores the raw data originally collected for each country converted into a standard input format, and the accompanying documentation. For each country included in the HFD, there are <u>four types of input data files</u> containing the following:

1) births by calendar year, and/or age of mother, and/or mother's year of birth, and (if available) birth order;
2) the age-parity distribution of the female population (from population registers, population censuses, or large-scale surveys);
3) births by calendar year and month; and
4) the female population by age (for countries and periods not available in the HMD).

In addition, there are <u>three files</u> with documentation for each country:

1) *Background and documentation*, which provides information about data quality, coverage, territorial changes, historical peculiarities, and other important comments;
2) *Notes*, which provides details about specific data points in the data files; and
3) *References,* which provides information about the data sources.

All of the input files are available on the HFD website and are freely accessible to HFD users. For the convenience of HMD[1] and HFD users, the HFD files are named in the same manner as the HMD files. The files available for each country are as follows:

- XXXbirths.txt (birth data);
- XXXparity.txt (age-parity distribution of women from population registers, censuses, or surveys);

---

[1] Human Mortality Database available at http://www.mortality.org

- XXXmonthly.txt (births by month of birth);
- XXXfpop.txt (female population by age), for countries and periods not available in the HMD;
- XXXcom.pdf (background and documentation);
- XXXnote.pdf (notes for individual data points); and
- XXXref.pdf (references).

The data are provided as comma-separated ASCII files, and the documentation files (XXXcom.pdf, XXXnote.pdf, and XXXref.pdf) are available in Portable Document Format. The data files have headings (first line) that are the same for all HFD countries. Comma (',') is used as a field delimiter. Missing values are coded as a single dot ('.'). A detailed description of each file is given below.

2.1 File for births

File name: XXXbirths.txt

Heading: *PopName, Area, Year, YearReg, Age, AgeInt, Lexis, BirthOrder, OrderInt, PrevBirth, DurInt, Vital, Births, Access, Note1, Note2, Note3, RefCode, LDB*

1. *PopName* (the same code 'XXX' as in the file name): country or area code (see full description above).
2. *Area* (numeric): territorial/country coverage. The description of these codes must be provided in the Background and Documentation file.
3. *Year* (four-digit numeric or '.'): calendar year in which the birth occurred. This may be '.' in cases in which the published births data are by year of registration only. In such cases, the country expert decides whether the year of registration can be treated as the year of occurrence (i.e., that treating *Year* as equivalent to *YearReg* does not significantly bias the calculated rates). If not, some adjusted birth counts corresponding to the year of occurrence should be deduced.
4. *YearReg* (four-digit numeric or '.'): registration year of the birth. Some countries report births both by the year of occurrence and by the year of registration, or by the year of registration only. In such cases, *Year*, *Age,* etc., refer to the date of occurrence; and *YearReg* refers to the registration year if this year is known. If the birth was registered in the year of its occurrence, *YearReg* is equal to *Year*.
5. *Age* (numeric; 'UNK' for unknown, 'TOT' for total): age of the mother. For age groups other than the lower open age interval (*AgeInt* value '-'), the value of *Age* is the lower age limit. It should be noted that the age of the mother in the input file always refers to the beginning of the year[2].
6. *AgeInt* (one- or two-digit numeric; '-' or '+' for the lower or the upper open age interval, respectively, always '.' for *Age* 'UNK' or 'TOT'): the length of the age interval.
7. *Lexis* (two characters, '.' allowed for *Age* 'TOT' or 'UNK'): shape of the Lexis element. 'TL' stands for lower triangle, 'TU' for upper triangle, 'RR' for rectangle/square, 'VV' for vertical parallelogram, and 'VH' for horizontal parallelogram.
8. *BirthOrder* (one- or two-digit numeric; 'UNK' for unknown, 'TOT' for total): 'true' (biological) birth order of the child.
9. *OrderInt* (one- or two-digit; '+' for open interval, always '.' for *BirthOrder* 'UNK' or 'TOT'): length of the birth order interval.
10. *PrevBirth* (four-digit numeric; '.' indicates that duration data are not available): calendar year in which the previous birth occurred.

---

[2] If in the original source these data were provided by calendar year and mother's age reached during the year (ARDY), the age must be (re)calculated and the data are reorganized to meet the standard HFD input format.

11. *DurInt* (one- or two-digit numeric; '−' or '+' for lower or upper open intervals, respectively; '.' indicates that duration data are not available): length (in years) of the interval between the births.
12. *Vital* (one-digit numeric; '1' for live births, '2' for total births): vitality of births. If the total birth counts are given in the source data, these may be included in the input file with *LDB* '0'; if these counts are not known, estimates of the number of live births should be made. If *LDB* is '1', *Vital* should also be '1'.
13. *Births* (numeric; '.' allowed where LDB is '0'): birth counts. In cases in which births are not published for a certain age/order category (e.g., out of concerns about confidentiality when the number of births is low), this may be '.', but *LDB* should be '0', and a note (see fields Note1-3), as well an additional row containing an estimation of the true value of this field for use in HFD calculations, should be included. The country-specific method applied in the estimation of the true values should be documented in the country Background and Documentation file.
14. *Access* (one-character: 'C' for confidential, 'O' for publicly accessible, 'U' for unknown): confidentiality/accessibility of the data. Birth counts with access fields 'C' (confidential) are replaced by dots '.' in the public version of the data files (published on the website).
15. *Note1* (numeric or '.'): each of the three fields ("*Note1*", "*Note2*", "*Note3*") contains a code that links to specific notes in the file "XXXnote.txt". These fields may be empty (denoted by a single dot '.'). All codes must be defined in the note file.
16. *Note2*: see *Note1* for details.
17. *Note3*: see *Note 1* for details.
18. *RefCode* (numeric): reference to the source of data. Complete references to the sources are stored in the file XXXref.txt. All of the codes used in this field must be defined in the reference file.
19. *LDB* (one-digit numeric): indicator of whether the data cell is used for the HFD Lexis database production ('1' for data used for producing the HFD Lexis database, '0' for other data). For example, LDB '0' may be used for raw data in which country-specific adjustments have been made, or for duplicate data from additional sources.

2.2 File for births by month of birth

File name: XXXmonthly.txt

<u>Headings:</u> *PopName, Area, Year, YearReg, Month, Vital, Births, Access, Note1, Note2, Note3, RefCode, LDB*

1. *PopName*: as for *Births* file.
2. *Area*: as for *Births* file.
3. *Year*: as for *Births* file.
4. *YearReg*: as for *Births* file.
5. *Month* (one- or two-digit numeric; 1-12, 'UNK' for unknown, 'TOT' for total): the month of birth.
6. *Vital*: as for *Births* file.
7. *Births*: as for *Births* file.
8. *Access*: as for *Births* file.
9. *Note1*: as for *Births* file.
10. *Note2*: as for Births file.
11. *Note3*: as for Births file.
12. *RefCode*: as for *Births* file.
13. *LDB*: as for *Births* file.

2.3 File for the female population by age and parity

File name: XXXparity.txt

Headings: *PopName, Area, Day, Month, Year, Age, AgeInt, Cohort, CohInt, Parity, ParityInt, Population, Access, Note1, Note2, Note3, RefCode, LDB*

1. *PopName*: as for *Births* file.
2. *Area*: as for *Births* file.
3. *Day* (one- or two-digit numeric; 1-31): day corresponding to the date of the population estimate or of the census. This field can be denoted by a single dot '.' if the data are the exposure-to-risk estimates, and thus are not female population counts pertaining to a particular date in a given year.
4. *Month* (numeric; 1-12): month corresponding to the date of the population estimate or of the census. This field can be denoted by a single dot '.' if the data are the exposure-to-risk estimates, and thus are not female population counts pertaining to a particular date in a given year.
5. *Year* (four-digit numeric): calendar year corresponding to the date of the population estimate or of the census,
6. *Age* (numeric; 'UNK' for unknown, 'TOT' for total, '.' where the population counts are by cohort and not by age): age of the woman. For age groups other than the lower open age interval (*AgeInt* value '-'), the value of *Age* is always equal to the lower age limit.
7. *AgeInt* (one- or two-digit numeric; '-' or '+' for the lower or upper open age intervals, respectively; always '.' for *Age* '.', 'UNK' or 'TOT'): length of the age interval.
8. *Cohort* (four-digit numeric; 'UNK' for unknown, 'TOT' for total, '.' where population counts are by age and not by cohort): woman's year of birth.
9. *CohInt* (one- or two-digit numeric; '−' or '+' for open interval, always '.' for *Cohort* '.', 'UNK' or 'TOT'): length of the cohort interval.
10. *Parity* (one- or two-digit numeric; 'UNK' for unknown, 'TOT' for total, 'CHL' for number of children ever born): parity of women (0, 1, 2, …). If, in addition to the number of women by age and parity, the total number of children ever born to each woman is provided, the latter data are denoted by 'CHL'.
11. *ParityInt* (one- or two-digit numeric; '+' for open interval, always '.' for *Parity* 'UNK', 'TOT' or 'CHL'), the length of the parity interval.
12. *Population* (numeric; '.' allowed where *LDB* is '0'): population size (number of women in a given category). In cases in which the population sizes are not published for a certain category (e.g., due to concerns about confidentiality when the numbers are low), this may be '.'. In such cases, an estimation of the true value of this field for use in the HFD calculations should be made using a country-specific method documented in the country Background and Documentation file.
13. *Access*: as for *Births* file.
14. *Note1*: as for *Births* file.
15. Note2: as for *Births* file.
16. Note3: as for *Births* file.
17. RefCode: as for *Births* file.
18. *LDB*: as for *Births* file.

2.4 File for the female population by age

File name: XXXfpop.txt

Headings: *PopName, Area, Age, AgeInt, Day, Month, Year, Population, Access, Note1, Note2, Note3, RefCode, LDB*

1. *PopName*: as for *Births* file
2. *Area*: as for *Births* file.
3. *Age*: as for *Parity* file.
4. *AgeInt*: as for *Parity* file.
5. *Day*: as for *Parity* file.
6. *Month*: as for *Parity* file.
7. *Year*: as for *Parity* file.
8. *Population*: as for *Parity* file.
9. *Access*: as for *Births* file.
14. *Note1*: as for *Births* file.
15. *Note2*: as for *Births* file.
16. *Note3*: as for *Births* file.
10. *RefCode*: as for *Births* file.
11. *LDB*: as for *Births* file.

2.5 File for notes

File name: XXXnote.pdf

This file contains specific notes that are identified by a note code, and are relevant to specific data points in the data files. The first line of each record contains a numeric code with the corresponding text provided below. The text notes may be written in a free form, but they should not include blank lines, because blank lines are used for separating the note records.

An example:

1
In the 1991 census, the number of women by parity was not published if the number in a given birth cohort and parity category was lower than four. Therefore, some figures for the birth cohorts 1975-1966 and for women at parities 2-6+ are substituted by the code for missing value ".".

2
The distribution of women by the number of live-born children was recalculated from the census date to 1 January of the given year (1991 or 2001) using the exact numbers of live-born children by birth order, age of mother, and date of birth between 1 January and the census date.

2.6 File for references

File name: <u>XXXref.pdf</u>

References to the data sources are stored in this file. Each source is displayed in a separate record that contains the following fields:

- *RefCode*: numeric reference code (the same as in the data files).
- *Source*: complete (or as complete as possible) reference to the data source.
- *Comments*: list of data types (e.g., births, population estimates, etc.) and comments from the country specialist.
- *Date*: date when the data were added to the database (format: dd.mm.yyyy).
- *Reference person*: person who placed this information in the *References* file.

An example:

RefCode
1
Source
Statistics Sweden (1976). Swedish Population History. Urval No 8.
Comments
Live births by calendar year and mother's age, 1891-1950
Date
24.07.2009
Reference person
Aiva Jasilioniene (Jasilioniene@demogr.mpg.de)

2.7 File for background description and documentation

File name: <u>XXXcom.pdf</u>

This file is posted on the website as part of each country page. It contains general information about the data, including about data quality, coverage, territorial changes, historical peculiarities, and country-specific methods (usually as an appendix); as well as other important comments.

APPENDIX 1: Names of input and output data files

1st data block: Summary Indicators

a) Period summary indicators

	All birth orders combined	By birth order
Total number of live births	XXXtotbirthsRR.txt	XXXtotbirthsRRbo.txt
Crude birth rate	XXXcbrRR.txt	XXXcbrRRbo.txt
Total fertility rate	XXXtfrRR.txt	XXXtfrRRbo.txt
Tempo-adjusted TFR	XXXadjtfrRR.txt	XXXadjtfrRRbo.txt
Mean age at birth	XXXmabRR.txt	XXXmabRRbo.txt
Standard deviation in mean age at birth	XXXsdmabRR.txt	XXXsdmabRRbo.txt

b) Cohort summary indicators

	All birth orders combined	By birth order
Completed cohort fertility	XXXtfrVH.txt	XXXtfrVHbo.txt
Parity progression ratios	---	XXXpprVHbo.txt
Mean age at birth	XXXmabVH.txt	XXXmabVHbo.txt
Standard deviation in mean age at birth	XXXsdmabVH.txt	XXXsdmabVHbo.txt

2nd data block: Age-Specific Data

a) Birth counts, population exposures, and rates: period

	All birth orders combined			By birth order		
	year, age, cohort	year, age	year, cohort	year, age, cohort	year, age	year, cohort
Birth counts	XXXbirthsTR.txt	XXXbirthsRR.txt	XXXbirthsVV.txt	XXXbirthsTRbo.txt	XXXbirthsRRbo.txt	XXXbirthsVVbo.txt
Female population exposure	XXXexposTR.txt	XXXexposRR.txt	XXXexposVV.txt	---	---	---
Age-specific fertility rates	XXXasfrTR.txt	XXXasfrRR.txt	XXXasfrVV.txt	XXXasfrTRbo.txt	XXXasfrRRbo.txt	XXXasfrVVbo.txt
Cumulative fertility rates	---	XXXcpfrRR.txt	XXXcpfrVV.txt	---	XXXcpfrRRbo.txt	XXXcpfrVVbo.txt

b) Birth counts, population exposures, and rates: cohort

	All birth orders combined	By birth order
Birth counts	XXXbirthsVH.txt	XXXbirthsVHbo.txt
Female population exposure	XXXexposVH.txt	---
Age-specific fertility rates	XXXasfrVH.txt	XXXasfrVHbo.txt
Cumulative fertility rates	XXXccfrVH.txt	XXXccfrVHbo.txt

3rd data block: Fertility Tables

a) Period fertility tables

	HFD parity estimates	Census- or register-based parity estimates
Fertility tables	XXXpft.txt	XXXpftc.txt
Female population exposure by parity	XXXexposRRpa.txt	XXXexposRRpac.txt
Conditional age-specific fertility rates	XXXmi.txt	XXXmic.txt
Parity- and age-adjusted TFR (PATFR)	XXXpatfr.txt	XXXpatfrc.txt
Table mean age at birth	XXXpmab.txt	XXXpmabc.txt

b) Cohort fertility tables

	HFD parity estimates
Fertility tables	XXXcft.txt

4th data block: Input Data

a) Input data

	Years
Births	XXXbirths.txt
Births by month	XXXmonthly.txt
Women by age and parity	XXXparity.txt
Population size and deaths	Either a link to the HMD or XXXfpop.txt
Notes	XXXnote.pdf
References	XXXref.pdf

b) Lexis maps

	Map
Births	XXXbirths.pdf
Population size	XXXpop.png
Death counts	XXXdeath.png