Merging, exploring, and batch processing data from the Human Fertility Database and Human Mortality Database

Dr Jon Minton

Sunday, February 22, 2015

# Introduction

This briefing paper will demonstrate how to use the R package plyr to merge all available data from the Human Mortality Database (HMD) and Human Fertility Database (HFD) into single datasets that follow a 'tidy data' format. With the data arranged in this format it then becomes easier to make comparisons between genders and groupings, and to automate the production of figures and other analyses for a large number of permutations of groups.

## Accessing all available data from the HFD

* Go to humanfertility.org and select login from the Registration section of the column on the left of the page.
* Once logged in, select 'Zipped Data Files' under the 'DATA' section of the column on the left of the page.
* Within the table 'Data by type', look for the link to the data type 'All types of HFD' data. Click on this link to begin the download.

The size of the HFD file is around 25Mb. Once unzipped, this increases to around 144Mb.

## Accessing all available data from the HMD

* Go to mortality.org and log in.
* Select 'Zipped Data Files'; scroll to the bottom of the page and click on the link in the table 'All countries for the HMD'.
* Once logged in, select 'Zipped Data Files' under the 'DATA' section of the column on the left of the page.
* Within the table 'Data by type', look for the link to the data type 'All types of HFD' data. Click on this link to begin the download.

The size of the HMD file is around 311 Mb. Once unzipped, this increases to around 1.24 Gb.

## Data preparation

This tutorial assumes that: - the two zipped files above have been downloaded and unzipped - the HFD file has been unzipped into a directory called hfd - the HMD file has been unzipped into a directory called hmd - that both the hfd and hmd directories are within a directory called raw\_data

The directory structure assumed is:

base\_dir - data -- raw\_data --- hfd --- hmd

## Structure of hfd directory

The HFD has a relatively straightforward directory structure: within 'hfd' is a directory called 'Files', and within this directory is a directory called 'zip\_w'. A total of 56 files are in this directory, and there are no additional directories.

## Structure of the hmd directory

The hmd directory has a more complex, branched structure. Once unzipped, the directory opens to 46 separate folders, each labelled with the country code of the country whose data they contain. Each of these country folders then has the same internal directory structure. As an example here is the directory structure of the first country by code, AUS:

* AUS -- CHECKS -- DOCS -- InputDB -- LexisDB -- STATS

The STATS directory then contains 55 data files in comma-separated value (CSV) format. Within this technical document, the aim will be to extract data from the files Deaths\_1x1.txt and Populations.txt from each of these subdirectories, but the methods described can be generalised to other operations.

# Data Management Concepts

## Tidy Data

Hadley Wickham defines a dataset as 'tidy' if: 1. Each variable forms a column. 2. Each observation forms a row. 3. Each type of observational unit forms a table.

[REF: p4 of <http://vita.had.co.nz/papers/tidy-data.pdf>]

Each row within a 'tidy data' format table contains two types of variable:

1. Variables defining the location of the observation
2. Variables defining the value of the observation.

In the case of the demographic data of interest in this exercise, 'location' variables include:

1. Age;
2. Year;
3. Country;
4. Sex (for the HMD).

Each row therefore begins with the three (HFD) or four (HMD) location variables, followed by a number of observation variables for that particular location. These could include:

1. Death counts.
2. Population counts.
3. Births

With data arranged this format, it becomes easier to derive additional variables, to compare between groups, and to automate the production of outputs for each country separately.

# Getting data from the HFD into 'tidy data' format

With this information, the following tables were identified as all including the same three location variables Code, Year, and Age: -asfrRR -birthsRR -cpfrRR - exposRR