Exercise 1-3

your_name
European Doctoral School of Demography 2019-20

Barcelona, March 30 – April 3 2020

Assignments are due Friday April 3 at midnight (CET). Send your assignment via email to alburezgutierrez[at]demogr.mpg.de with the subject line "EDSD assignment".

Read Familinx data

```
# Read the Familinx data to your Global Environment

# For large datasets, we'll use data.table.
# Data.table is a state-of-the-art package for working with
# large datasets in R.
# In this exercise we'll only use it to read the data  and then we transform to
# data.frame since data.table objects obey different rules in R.
# If you have the time and interest, I strongly recommend that you check out the data.table
# documentation and do the exercises using data.table instead of the tidyverse:
# https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html

library(data.table)
library(tidyverse)

# This is a reduced version of the original Familinx data, filtered to include
# only cases from Sweden and keeping only the essential columns for the exercise.
# The data is described in details in:
# Kaplanis, J., et al. (2018). Quantitative analysis of population-scale family
# trees with millions of relatives. Science 360(6385):171-175.

# Load using data.table and then converto to data.frame
# Make sure that getwd() is the directory where this script is stored
prof <- data.table::fread("../Data/sweden_genealogy.csv", stringsAsFactors = F) %>%
  data.frame
```

Exercise 1

Focusing on people born between 1750 and 1850, consider the following: how has lifespan developed historically in Sweden, according to the online genealogies?

1. Compute the lifespan average by birth cohort and sex. For this exercise, I recommend you group birth cohorts by 15 years (e.g. 1750-1774; 1775-1799; etc.).

```
head(prof)
```

```
##   profileid   father   mother gender is_alive birth_year death_year burial_year
## 1       136       NA       NA   male        0         NA         NA          NA
## 2       264 57536639 69836161 female        0       1875         NA          NA
## 3       708 83768131 48140261   male        0         NA         NA          NA
## 4       722       NA       NA   male        0       1694       1767          NA
## 5       812       NA 66543089   male        0       1871       1933          NA
## 6       860       NA       NA   male        0         NA         NA          NA
##   baptism_year birth_month death_month burial_month baptism_month
## 1           NA          NA          NA           NA            NA
## 2           NA          12          NA           NA            NA
## 3           NA          NA          NA           NA            NA
## 4           NA          NA           2           NA            NA
## 5           NA          11          10           NA            NA
## 6           NA          NA          NA           NA            NA
```

2. Include a short description of your findings (max 200 words) and one figure that summarises them.

Exercise 2

What is the difference between lifespan and life expectancy? The two readings from Thursday seem to conflate both terms at points.

1. Write a short paragraph (max 150 words) describing the connection between lifespan and life expectancy for a given birth cohort.
2. Is it possible to evaluate this empirically using data from online genealogies? Write a short paragraph (max 250 words) indicating how you would do it
3. For extra points, compute the cohort life expectancy for any given birth cohort using the genealogies (optional)

Exercise 3

What are potential sources of bias in the online genealogies? How can we evaluate these biases?

1. Write a short paragraph (max 250 words) describing three potential sources in bias in online genealogies.
2. Focus on one of the three biases identified above and provide evidence of its existence using empirical data
3. Write a short paragraph with a potential solution to overcome this bias (max 200 words).