Barcelona Summer School of Demography

Module 2. Demography with R

1. Basic Demographic Measures

Basic Demographic measures

Tim Riffe
tim.riffe@gmail.com

7 July 2025

Contents

1 Preliminary remarks

This aim of this *Demography with R* module is to provide you the basics of demographic thinking, and to help you start producing basic demographic measures in R. The class will focus on three major demographic products: the life table, decomposition, and projections.

As for the previous module, lectures are scheduled every morning from 10:00 to 14:00. We will take a 10 minute break every hour. Each lecture will be a mixture of theory, illustration, and practical exercises.

This material was initially hand crafted by the great Marie-Pier Bergeron-Boucher, and is typically masterfully delivered by her. I'm covering for her this season of BSSD, but I can't help myself but to deviate the content just a bit. Specifically, it will be nice to do as much as possible of this work in the so-called-tidy framework. But for the sake of time, this won't be as explicit an introduction as I'd like. If you're so inclined, you can supplement this superficial tidyverse intro check out my tidy-intro materials from previous such workshops.

- BSSD 2023 Module 1: https://github.com/timriffe/BSSD2023Module1
- EDSD 2023 Data Wrangling module: https://github.com/timriffe/EDSD2023data

- BSSD 2024 Module 2: https://github.com/timriffe/BSSD2024Module2

Materials for this module will be hosted here: https://github.com/timriffe/BSSD2025Module2 This repository will be built over the course of this week and updated after each live session.

2 Review of R concepts and functions

2.1 Vector, data.frame, matrix and list

- **data.frame** A `data.frame` is a table where the columns are variables and the rows are observation. Each column/variable can be of different types (e.g. factors, numeric, etc.). Can be created with `data.frame()`. `tibble` is another kind of `data.frame` we'll introduce.
- **vector** A combination of elements. Can be created with `c()`, `seq()`, `rep()`.
- **matrix** A matrix, is a two-dimensional object where a variable, generally numeric (e.g. birth counts), is arranged into rows and columns. Can be created with `matrix()`.
- **list** A list can contain elements of different types and dimensions. Can be created with `list()`.

2.2 R commands

`as.character()`: Transforms an object into a character type object.

```
#Example

A<-seq(1,10,2)
A
```

```
## [1] 1 3 5 7 9
```

```
class(A)
```

```
## [1] "numeric"
```

```
B<-as.character(A)
B
```

```
## [1] "1" "3" "5" "7" "9"
```

```
class(B)
```

```
## [1] "character"
```

`as.numeric()`: Transform an object into a numeric type object (e.g. 1,2,10,etc.).

```r
#Example

C<-as.numeric(B)
C
```

```
## [1] 1 3 5 7 9
```

```r
class(C)
```

```
## [1] "numeric"
```

Other useful goodies include `as.integer()`, `as.factor()`, and so forth. Likewise, a `data.frame` can be coerced to a `matrix` with `as.matrix()` (or in the other direction, with `as.data.frame()`), `as.list()` separates the columns of a `matrix` or `data.frame` into `vector` elements of a `list`. When in doubt, try `as.*` whenever you're feeling coercive. My feeble attempts to *tidy* this module will certainly introduce the `tibble()` object, which you can think of as another name for `data.frame`.

`expand.grid()` (equivalent `expand_grid()`: Create a data frame from all combinations of supplied vectors or factors.

`read_csv()` from `readr` package: Reads a comma-separated file in table format and creates a data frame from it with pretty good default settings. `read_delim()` from the same package is more general.

2.3 tidyverse concepts

Tidy refers to a particular data format. You may have come across long vs wide format data in other contexts. `tidy` data is closer to long format, but in a particular way. It is a rectangular data layout, where *observations* are in rows and *variables* are in columns. That's literally all there is to it, and it really is somewhat more flexible than it looks at first, since it's not always clear what constitutes an observation. For panel survey data, an observation might be a respondent, a respondent-wave, or a respondent-wave interval! For aggregate demography, an observation might be as as simple as a combination of year, age, and sex. Variables are in this case, structuring variables, such as year, age, and sex, as well as value variables on hand, such as population counts, deaths counts, and the like. A tidy data structure fits many demographic data applications very well. I'll do my best to merge the material of this block with a tidy approach, as this will give you a strong tool kit for data analysis.

`dplyr` verbs 1. `pivot_longer()` stack a range of columns 2. `pivot_wider()` take a single column and spread it over columns 3. `group_by()` declare subgroups for independent operations 4. `mutate()` make new columns, no loss of rows 5. `summarize()` aggregates over rows. Usually reduces nr of rows 6. `select()` selects columns 7. `filter()` selects rows (subsets) 8. `ungroup()` removes group declaration 9. `group_modify()` can be used to apply a `data.frame`-in `data.frame`-out function 10. `reframe()` is like `mutate()` or `summarize()` but more flexible with respect to output length. 9. `left_join()` merges two data sources, preserving the rows of the left-side object. Also interesting joins: `right_join()`, `inner_join()`, `full_join()`, `cross_join()`

These and other such statements can be strung together into data *pipelines* using the pipe operator `|>` or `|>` (shortcut: `Ctrl + Shift m` [`Cmd + Option + m` on mac])

I have a habit of rewarding the completion of a pipeline with a plot of some kind using `ggplot2`. I will do this on the fly for the sake of the output. Ilya will give an explicit and more complete intro to `ggplot2` concepts in Module 3. My peppering it into this module will at least prime you for his master class! Now let's talk about demography!

Of course there's a cheat sheet that can help you figure out which concept you need: https://posit.co/resources/cheatsheets/. Of these, the `tdyr`, `dplyr`, and `ggplot2` cheat sheets might be the most useful overviews.

3 General demographic concepts

3.1 What is demography?

According to the Max Planck Institute for Demographic Research: "*Demography is the science of populations. Demographers seek to understand population dynamics by investigating three main demographic processes: birth, migration, and aging (including death).*" [https://www.demogr.mpg.de/en/about_us_6113/what_is_demography_6674/]

According to the Oxford Dictionary demography is "*The study of statistics such as births, deaths, income, or the incidence of disease, which illustrate the changing structure of human populations.*"

One more... "*Demography is the study of the size, territorial distribution, and composition of population, changes therein, and the components of such changes.*" (Hauser and Duncan 1959)

In summary, demographers use quantitative approaches to study population dynamics and changes by investigating their

- Size
- Growth
- Composition/structure (age, socioeconomic status, etc.)
- Processes: Fertility, Mortality and Migration

And maybe a thousand more things because really the field is a cognate of many other sciences.

3.2 What is a population?

Broadly defined, demographers refer to a population as "*the collection of persons alive at a specific point in time who meet certain criteria*" (Preston, Heuveline, and Guillot 2001).

Often we just deal with populations defined by administrative boundaries, but this is not necessarily the case.

3.3 Three dimensions of changes

Generally, demographers study demographic events (birth, death and migration) on three interrelated aspects of temporal structure: age, period, and cohort.

3.3.1 Age and period

In censuses, civil registries, and most surveys, data and events will generally be recorded by year and age of occurrence. For example, in 2014, 1202 infants died before age 1 in Spain (Human Mortality Database 2018).

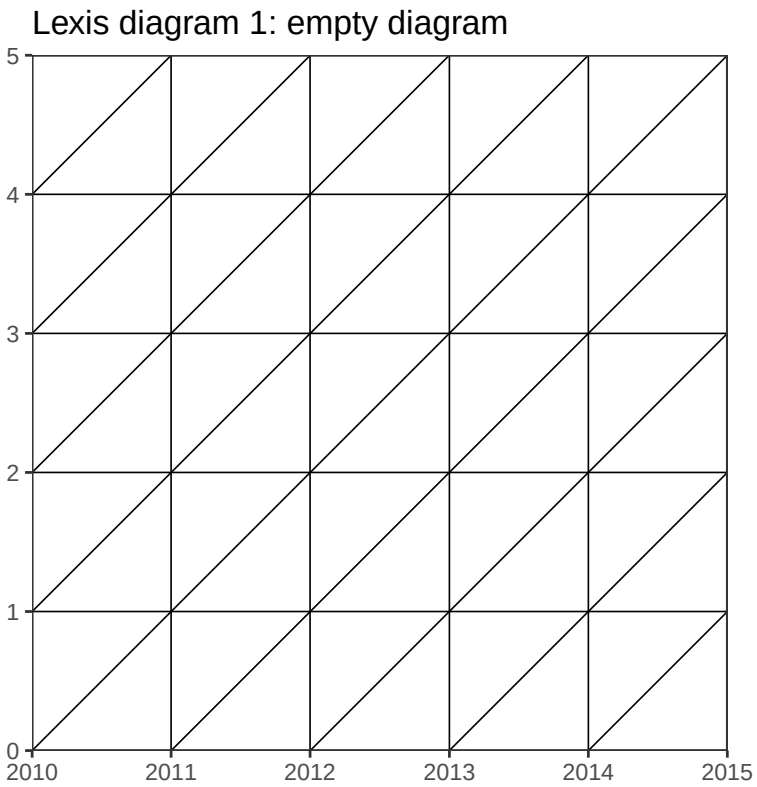
3.3.2 Cohort

A cohort is defined as "*the aggregate of all units that experience a particular demographic event during a specific time interval*" (Preston, Heuveline, and Guillot 2001). For example, the 1950 Spanish birth cohort consist of all individuals born in 1950 in Spain.

3.3.3 Lexis diagram

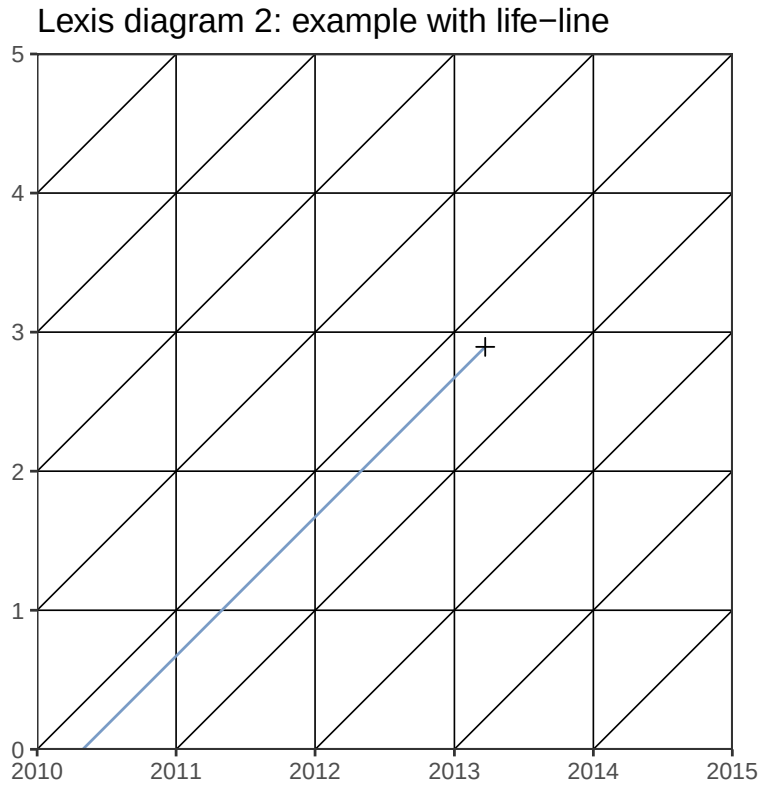
The Lexis diagram is a two-dimensional figure used to represent demographic events across year, age, and cohort. On the diagram, age is one dimension and calendar year is the other. Cohorts advance through life along a 45-degree line.

```
library("LexisPlotR")
library(ggplot2)
mylexis <- lexis_grid(year_start = 2010, year_end = 2015,
                      age_start = 0, age_end = 5)+
  labs(title = "Lexis diagram 1: empty diagram")
mylexis
```



On a Lexis diagram, each individual is represented as a *life-line*, advancing through time and age as a parallel line to the cohort line.

```
mylexis<-lexis_grid(year_start = 2010, year_end = 2015,
                    age_start = 0, age_end = 5)+
  labs(title = "Lexis diagram 2: example with life-line")
lexis_lifeline(lg = mylexis,
               birth = "2010-05-01",
               exit="2013-03-23",
               lineends = TRUE)
```



4 Data

We will use the Spanish females data from the HMD (Human Mortality Database 2018) and HFD (Human Fertility Database 2018) for the year 2014 (and 2015 population counts). Please read in the raw data using these lines:

```
library(tidyverse)
library(readr)
#Birth counts in 2014
B2014 <- read_csv("data/ES_B2014.csv") |>
  mutate(sex = "total") |>
  select(year, age, sex, births = total)
```

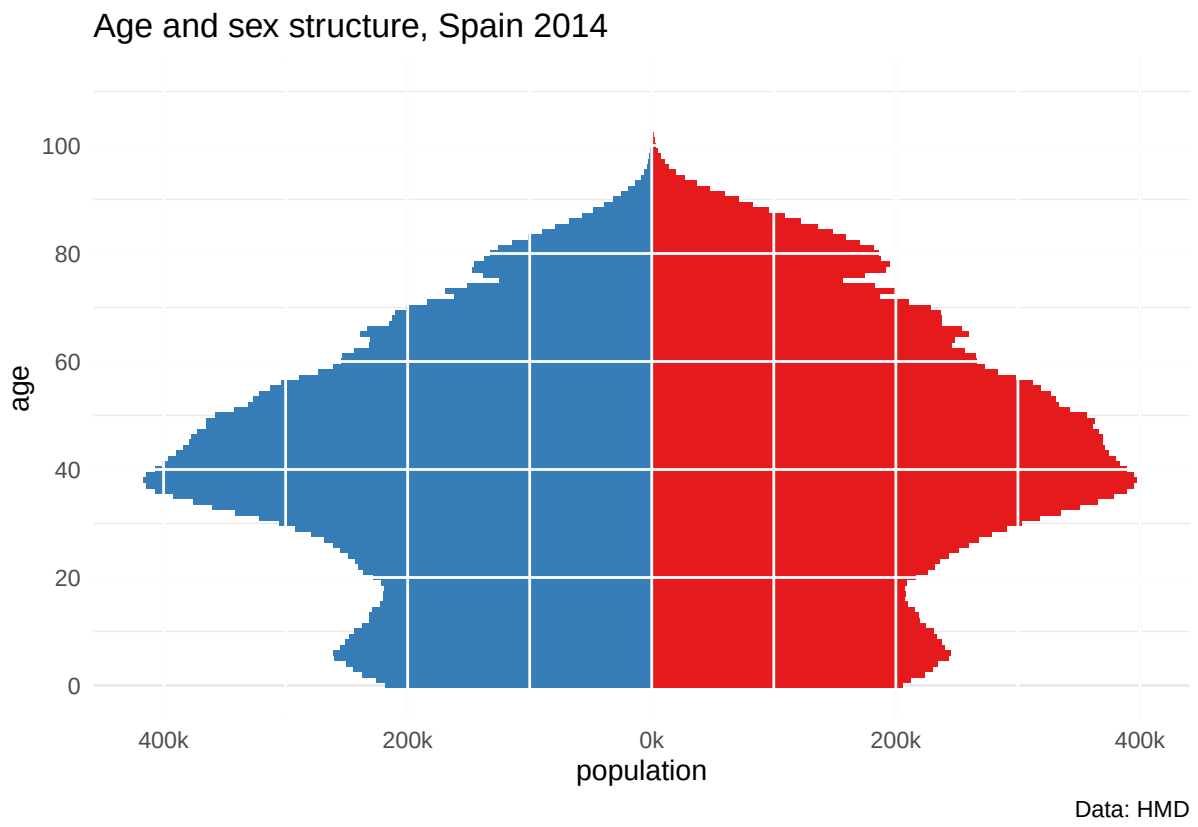
```r
# Death counts in 2014
D2014 <- read_csv("data/ES_D2014.csv")  |>
         pivot_longer(female:total,
                      names_to = "sex",
                      values_to = "deaths")

# Population counts on January 1st, 2014 and 2015,
# get exposure as the mean
E2014 <- read_csv("data/ES_P2014.csv") |>
         pivot_longer(female:total,
                      names_to = "sex",
                      values_to = "pop") |>
  arrange(sex, age) |>
  group_by(sex, age) |>
  summarize(exposure = mean(pop))

# fertility rates by age
Fx2014 <-
  B2014 |>
  mutate(sex = "female") |>
  left_join(E2014, by = join_by(sex, age)) |>
  mutate(asfr = births / exposure)
```

5 Population pyramid

The population pyramid or age pyramid is the most common tool to visualize the age structure of a population by sex (for better or worse). Each age is represented by a horizontal bar and male age distribution is plotted on the left and female age distribution on the right.

```r
library(ggplot2)
E2014 |>
  select(sex, age, pop = exposure) |>
  filter(sex != "total") |>
  mutate(population = ifelse(sex == "male", -pop, pop)) |>

  ggplot( aes(x = age, y = population, fill = sex)) +
  geom_col(width = 1) +
  # axis labels and grids

    scale_y_continuous(breaks = seq(-400000, 400000, 200000),
                       labels = paste0(as.character(c(seq(400, 0, -200), seq(200, 400, 200))),
  scale_x_continuous(breaks = seq(0, 100, 20))+

  # extra nice grids (2m per box)- needs to be worked out by hand
  geom_vline(xintercept = seq(20, 100, by = 20),
             color = "white",
             linewidth = .5) +
  geom_hline(yintercept = seq(-400000, 400000, by = 100000),
```

```
            color = "white",
            linewidth = .5) +
guides(fill = "none") +

# final flip x and y
coord_flip() +
scale_fill_brewer(palette = "Set1") +
theme_minimal() +
labs(caption = "Data: HMD",
     title = "Age and sex structure, Spain 2014")
```



6 Demographic rate vs probability

6.1 Probability

A probability is a concept for cohorts only. It refers to the chance that an event will occur (Preston, Heuveline, and Guillot 2001) and is calculated as:

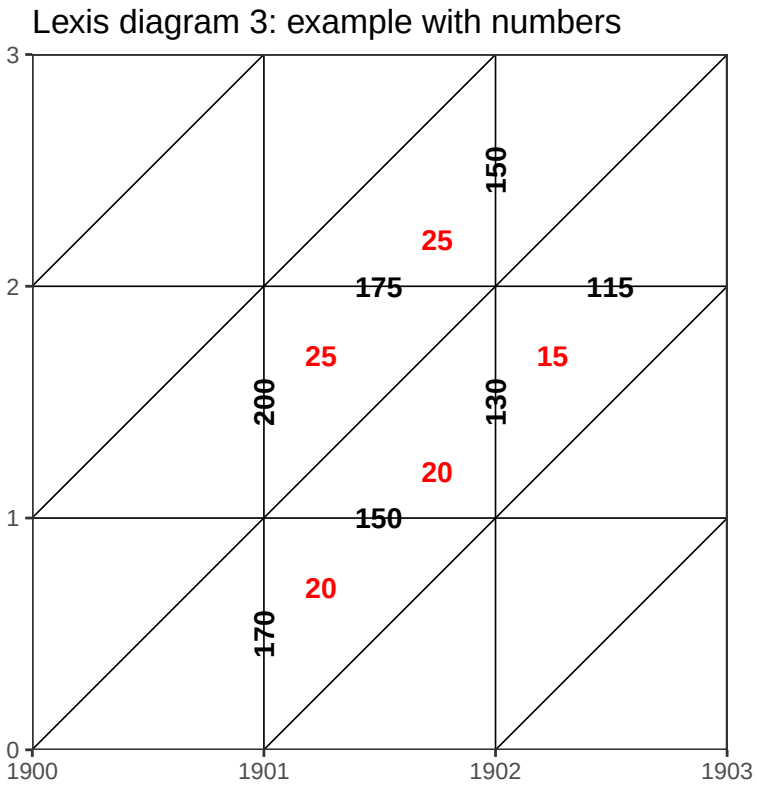
$$Probability = \frac{Number\ of\ Occurences}{Number\ of\ preceeding\ Events\ or\ trials}$$

For example, the probability of dying at a specific age - e.g. age 5 - is the number of deaths occurring at age 5 (between the exact age 5 and exact age 6) in a cohort, divided by the number of people who reached the exact age of 5. In the Lexis diagram below, the numbers in red are

the events (deaths) and the numbers in black are the number of people alive at an exact age or date. Using this example, the death probability at age 1 from the birth cohort 1900 is:

```
deaths<-20+15
preAlive<-150
probability<- deaths/preAlive
probability
```

```
## [1] 0.2333333
```



6.2 Rate

Rather than the chance that an event occurs, as the probability, we here look at the rate at which an events occur. Rates can be both applied to period and cohort analysis. However, as most demographic information is periodic, period rates are often used. Other period-perspective demographic quantities, such as period life expectancy, or conditional probabilities, are then derived as needed from rates.

In demography, rates are normally known as occurrence-exposure rates (events relative to exposure to risk). Often, exposure is referred to person-years lived (Preston, Heuveline, and Guillot 2001). Rates differ from probabilities in the denominator:

$$Rate = \frac{Number\ of\ Occurences}{Number\ of\ person-years\ lived}$$

With rates, the number of occurrences is scaled to the population size AND per unit time.

6.2.1 Person-years

The person-years concept refers to the number year a person lived - i.e. is *exposed* to the event - in a specific interval. For example, if a person lives 1 year in an interval 0 to 1 year, then this person contributed one person-year. If another person lived 5 days in the same interval, this person contributed (about) 5/365th of a person-year in the same time interval.

In the Lexis diagram 2, the life-line illustrates the case of a person who contributed 0.89 (326/365) person-year of exposure to the risk of dying at age 2 for the birth cohort 2010. The same person contributed 2.89 person-years of exposure in the full 2010 birth cohort.

When person-years are used, the rate is an *annualized* rate.

Person-years are rarely observed or counted directly (Preston, Heuveline, and Guillot 2001). Thus, person-years (PY) are often calculated by assuming that the persons (P) who experienced an event (E) in a given interval ($t : t + n$), experienced it at the mid-point of the interval:

$$ {}_nPY_t = \frac{P_t + P_{t+n}}{2} $$

For example, the death rate at age 1 for the birth cohort 1900 on the Lexis diagram 3 is:

```
deaths <- 20+15
Pyear  <- (150+115)/2
rate   <- deaths/Pyear
rate
```

```
## [1] 0.2641509
```

The period death rate at age 1 in 1901 is:

```
deaths <- 25+20
Pyear  <- (200+130)/2
rate   <- deaths/Pyear
rate
```

```
## [1] 0.2727273
```

Sometimes statistical offices release midyear population estimates, derived either directly from population registers or on the basis of demographic accounting. There are also plenty of other approximations out there.

7 Death rates (all-cause)

- Events (occurrences): deaths
- Exposure: All person-years lived in the time interval

7.1 Crude death rate

The crude death rate (CDR) is a measure of the risk of death for a whole population:

$$CDR[t,n] = \frac{Number\ of\ deaths\ in\ the\ population\ between\ times\ t\ and\ t+n}{Number\ of\ person-years\ lived\ in\ the\ population\ between\ times\ t\ and\ t+n}$$

```
# Person-years
E2014 |>
  left_join(D2014,
            by = join_by(sex, age)) |>
  group_by(sex) |>
  summarize(exposure = sum(exposure),
            deaths = sum(deaths)) |>
  mutate(CDR = 1000 * deaths / exposure)
```

```
## # A tibble: 3 x 4
##   sex     exposure deaths   CDR
##   <chr>      <dbl>  <dbl> <dbl>
## 1 female 23605597. 193620  8.20
## 2 male   22838292. 200114  8.76
## 3 total  46443889. 393734  8.48
```

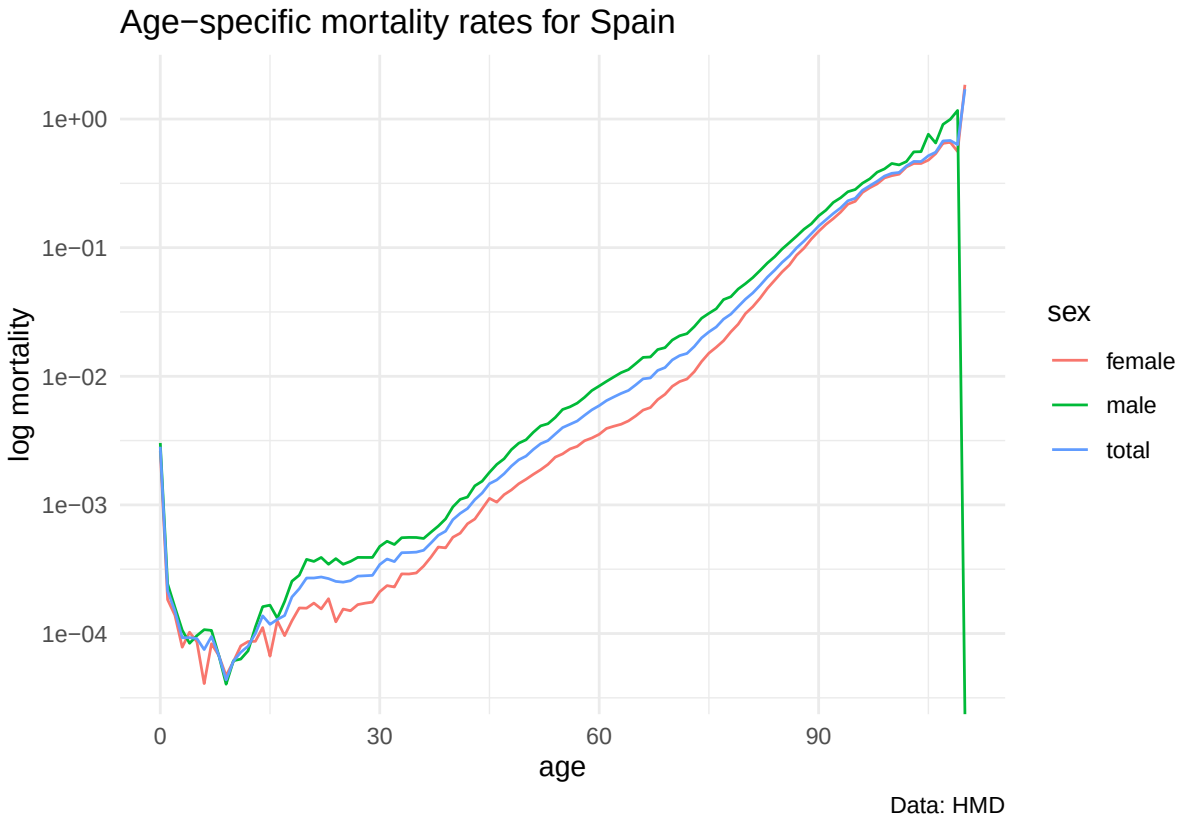
7.2 Age-specific death rates

Age-specific death rates (M) measure the risk of death by age x (or between age x and $x+n$) in the population:

$${}_nM_x[0,T] = \frac{Number\ of\ deaths\ in\ the\ age\ range\ x\ to\ x+n\ between\ times\ T\ and\ T+t}{Number\ of\ person-years\ lived\ in\ the\ age\ range\ x\ to\ x+n\ between\ times\ T\ and\ T+t}$$

This measure controls for the effect of population age-structure, i.e. some populations have older age-structure than others resulting in higher CDR even if the ${}_nM_x$ are smaller at most ages. The age interval width n and calendar interval t are up to you and the data.

```
E2014 |>
  left_join(D2014, by = join_by(sex, age)) |>
  mutate(mx = deaths / exposure) |>
  ggplot(aes(x = age, y = mx, color = sex)) +
  geom_line() +
  scale_y_log10() +
  labs(title = "Age-specific mortality rates for Spain",
       caption = "Data: HMD",
       y = "log mortality") +
  theme_minimal()
```



8 Fertility rates

- Events: births
- Exposure: every women alive in their reproductive age ($\approx$ 15 to 50 years old)

8.1 Crude birth rate

The crude birth rate (CBR) is a loose measure of occurrence/exposure of fertility:

$$CBR[0,T] = \frac{Number\ of\ births\ in\ the\ population\ between\ times\ T\ and\ T+t}{Number\ of\ person-years\ lived\ in\ the\ population\ between\ times\ T\ and\ T+t}$$

```
# Person-years
B2014 |>
  right_join(E2014,by=join_by(sex, age)) |>
  filter(sex == "total") |>
  summarize(exposure = sum(exposure),
            births = sum(births, na.rm = TRUE)) |>
  mutate(CBR = 1000 * births / exposure)
```

```
## # A tibble: 1 x 3
##    exposure births   CBR
##       <dbl>  <dbl> <dbl>
## 1 46443889. 426076  9.17
```

8.2 General fertility rate

The general fertility rate (GFR) is generally considered a better measure of fertility, as only women in their reproductive ages can give birth, and are thus at risk of experiencing the event.

$$GFR[0,T] = \frac{Number\ of\ births\ in\ the\ population\ between\ times\ T\ and\ T+t}{Number\ of\ person-years\ lived\ by\ women\ aged\ 15\ to\ 50\ between\ times\ T\ and\ T+t}$$

```
# Female population in 2014 and 2015 at reproductive ages
Fx2014 |>
  filter(between(age, 15, 50)) |>
  summarize(exposure = sum(exposure),
            births = sum(births)) |>
  mutate(GFR = 1000 * births / exposure)
```

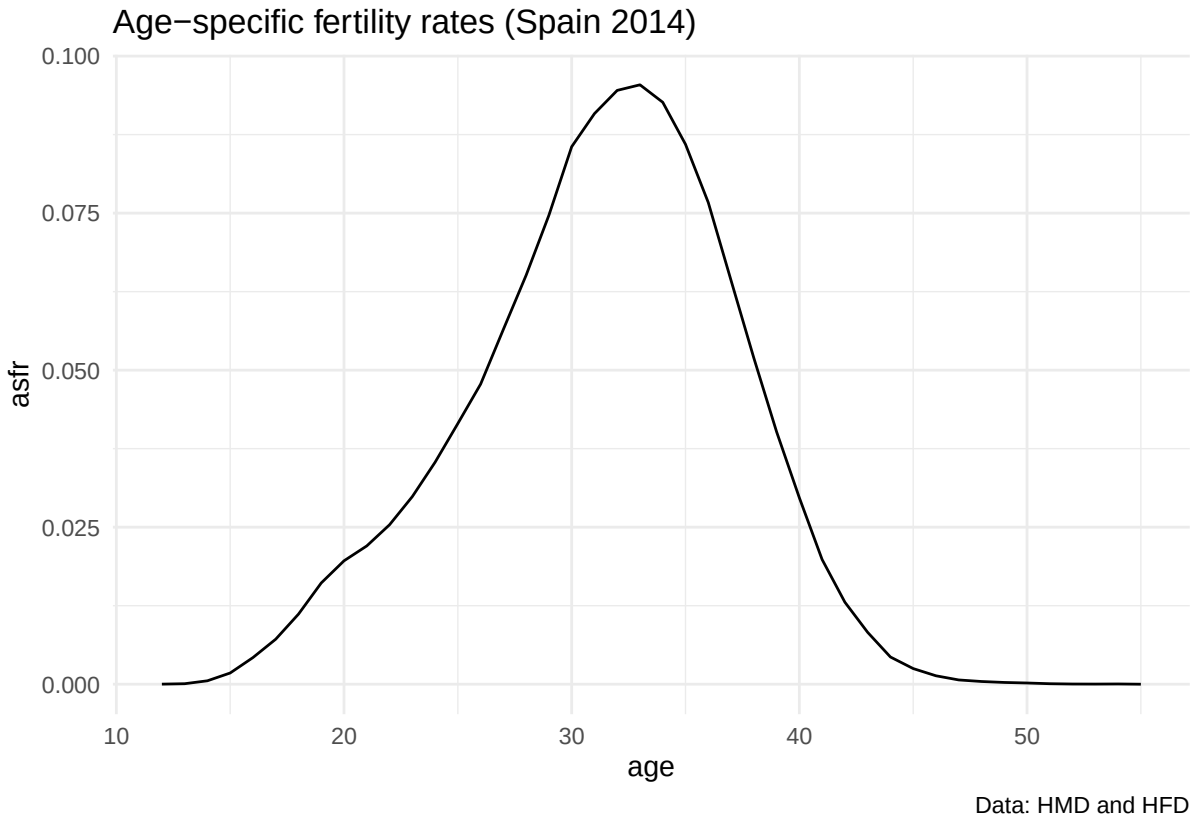
```
## # A tibble: 1 x 3
##    exposure births   GFR
##       <dbl>  <dbl> <dbl>
## 1 11253028. 425888  37.8
```

8.3 Age-specific fertility rates

As with age-specific death rates, age-specific fertility rates (F) are less sensitive to the age structure of the population. This measure provides the risk of given birth from women age x to $x+n$:

$${}_nF_x[0,T] = \frac{Number\ of\ births\ between\ times\ T\ and\ T+t\ to\ women\ aged\ x\ to\ x+n}{Number\ of\ person-years\ lived\ by\ women\ aged\ x\ to\ x+n\ between\ times\ T\ and\ T+t}$$

```
Fx2014 |>
  ggplot(aes(x = age, y = asfr)) +
  geom_line() +
  labs(title = "Age-specific fertility rates (Spain 2014)",
       caption = "Data: HMD and HFD") +
  theme_minimal()
```



8.4 Total fertility rate

The total fertility rate (TFR) is the average number of children a woman would have if she experienced the a particular set of age-specific fertility rates and survived until the end of her reproductive age. "*The TFR is the single most important indicator of fertility*" (Preston, Heuveline, and Guillot 2001). It is also the area under the ASFR curve.

$$TFR[T, T+t] = n \sum_{x=a}^{B-n} {}_nF_x[T, T+t]$$

where a and B are the minimum and maximum age at childbearing.

```
# TFR
Fx2014 |>
  mutate(n = 1) |> # make this 5 if you have 5-year age groups!!
  summarize(TFR = sum(asfr * n))
```

```
## # A tibble: 1 x 1
##     TFR
##   <dbl>
## 1  1.32
```

8.5 Mean age at childbearing

The mean age at childbearing is not a rate, but is based on the age-specific fertility rates. The mean age at childbearing (MA) is the average age of mothers at childbearing, standardized for the age-structure of the female population at reproductive age (Human Fertility Database 2018).

$$MA[T, T+t] = \frac{\sum_{x=a}^{B-n} \bar{x} * {}_nF_x[T, T+t]}{\sum_{x=a}^{B-n} {}_nF_x[T, T+t]}$$

where $\bar{x}$ is the mid-age of interval $x : x+n$, i.e. $\bar{x} = x + n/2$.

```
Fx2014 |>
  mutate(n = 1,
         age_mid = age + n / 2) |>
  summarize(MAB1 = sum(age_mid * asfr) / sum(asfr),
            MAB2 = sum(age_mid * births) / sum(births))
```

```
## # A tibble: 1 x 2
##    MAB1  MAB2
##   <dbl> <dbl>
## 1  31.8  32.8
```

Exercises

Load the data `ES_P.csv.gz` and `ES_B.csv.gz` (no need to unzip)

1. Calculate age-specific fertility rates from 1950 until 2014 (annually) from age 12 to 55. Skip the year 1975!
2. Plot the age-specific rates in 1950 and 2014.
3. Calculate the TFR at each year.
4. Plot TFR over time.
5. Calculate the mean age at birth over time.
6. Plot the MAB over time.

References

Hauser, Philip M., and Otis Dudley Duncan. 1959. *The Study of Population, an Inventory and Appraisal*. Chicago: University Press (London, Cambridge University Press.

Human Fertility Database. 2018. "Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria)."

Human Mortality Database. 2018. "University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany)."

Preston, Samuel H., Patrick Heuveline, and Michel Guillot. 2001. *Demography: Measuring and Modeling Population Processes*. Oxford: Blackwell.