Understanding World Population Dynamics

Assignment 1 - PSYC593

Meichai Chen

2023-09-22

Understanding population dynamics is important for many areas of social science. We will calculate some basic demographic quantities of births and deaths for the world's population from two time periods: 1950 to 1955 and 2005 to 2010. We will analyze the following CSV data files - `Kenya.csv`, `Sweden.csv`, and `World.csv`. Each file contains population data for Kenya, Sweden, and the world, respectively. The table below presents the names and descriptions of the variables in each data set.

Name	Description
`country`	Abbreviated country name
`period`	Period during which data are collected
`age`	Age group
`births`	Number of births in thousands (i.e., number of children born to women of the age group)
`deaths`	Number of deaths in thousands
`py.men`	Person-years for men in thousands
`py.women`	Person-years for women in thousands

Source: United Nations, Department of Economic and Social Affairs, Population Division (2013). *World Population Prospects: The 2012 Revision, DVD Edition.*

We start by loading the necessary packages.

```r
# Load packages ----
library(tidyverse)
library(ggplot2)
```

We next set up paths to different directories so we can load data and save different outputs.

```r
# Path variables ----
here_path <- here::here()
code_path <- file.path(here_path, "src")
docs_path <- file.path(here_path, "doc")
data_path <- file.path(here_path, "data")
figs_path <- file.path(here_path, "results", "figures")
```

Now, we read in the three CSV data files - `Kenya.csv`, `Sweden.csv`, and `World.csv`.

```r
# Read data ----
world_data  <- readr::read_csv(paste0(data_path, "/raw_data/World.csv"))
kenya_data  <- readr::read_csv(paste0(data_path, "/raw_data/Kenya.csv"))
sweden_data <- readr::read_csv(paste0(data_path, "/raw_data/Sweden.csv"))
```

We want to next convert the variable `age` into a factor that follows the correct numerical order.

```r
# Convert age to factor ----
age_group <- unique(kenya_data$age)
kenya_data$age <- factor(kenya_data$age, levels = age_group, ordered = TRUE)
```

The data are collected for a period of 5 years where *person-year* is a measure of the time contribution of each person during the period. For example, a person that lives through the entire 5 year period contributes 5 person-years whereas someone who only lives through the first half of the period contributes 2.5 person-years. Before you begin this exercise, it would be a good idea to directly inspect each data set. In R, this can be done with the `View` function, which takes as its argument the name of a `data.frame` to be examined. Alternatively, in RStudio, double-clicking a `data.frame` in the `Environment` tab will enable you to view the data in a spreadsheet-like view.

Question 1

We begin by computing *crude birth rate* (CBR) for a given period. The CBR is defined as:

$$\text{CBR} = \frac{\text{number of births}}{\text{number of person-years lived}}$$

Compute the CBR for each period, separately for Kenya, Sweden, and the world. Start by computing the total person-years, recorded as a new variable within each existing `data.frame` via the `$` operator, by summing the person-years for men and women. Then, store the results as a vector of length 2 (CBRs for two periods) for each region with appropriate labels. You may wish to create your own function for the purpose of efficient programming. Briefly describe patterns you observe in the resulting CBRs.

Answer 1

We start by computing the total person-years for Kenya, Sweden, and the world.

```
# Create new variable py = total person years for each data set
world_data$py <- world_data$py.men + world_data$py.women
kenya_data$py <- kenya_data$py.men + kenya_data$py.women
sweden_data$py <- sweden_data$py.men + sweden_data$py.women
```

We then create a function that will compute the crude birth rate given a data set.

```
# Function to compute the Crude Birth Rate (CBR)
compute_cbr <- function(pop_data) {
  pop_data %>%
  group_by(period) %>%
  summarise(cbr = sum(births) / sum(py)) %>%
  pull()
}
```

Then, we use the function created in the previous code chunk to compute the CBR for Kenya, Sweden, and the world.

```
# Compute the CBR for each data set
world_cbr <- compute_cbr(world_data)
kenya_cbr <- compute_cbr(kenya_data)
sweden_cbr <- compute_cbr(sweden_data)

# Combine CBR into a table
cbr <- rbind(world_cbr, kenya_cbr, sweden_cbr)
row.names(cbr) <- c("World", "Kenya", "Sweden")
colnames(cbr) <- c("1950-1955", "2005-2010")
knitr::kable(cbr)
```

	1950-1955	2005-2010
World	0.0373286	0.0202159
Kenya	0.0520949	0.0385151
Sweden	0.0153961	0.0119255

We see that for Kenya, Sweden, and the whole world, crude birth rates decreased from the period 1950-1955 to the period 2005-2010. Within the period 1950-1955, we observe that

Kenya has a higher CBR than the world and Sweden has a lower CBR than the world. The same pattern persists for the 2005-2010 period.

Question 2

The CBR is easy to understand but contains both men and women of all ages in the denominator. We next calculate the *total fertility rate* (TFR). Unlike the CBR, the TFR adjusts for age compositions in the female population. To do this, we need to first calculate the *age specific fertility rate* (ASFR), which represents the fertility rate for women of the reproductive age range $[15, 50)$. The ASFR for age range $[x, x+\delta)$, where x is the starting age and δ is the width of the age range (measured in years), is defined as:

$$\text{ASFR}_{[x,\ x+\delta)} \ = \ \frac{\text{number of births to women of age } [x,\ x+\delta)}{\text{Number of person-years lived by women of age } [x,\ x+\delta)}$$

Note that square brackets, [and], include the limit whereas parentheses, (and), exclude it. For example, $[20, 25)$ represents the age range that is greater than or equal to 20 years old and less than 25 years old. In typical demographic data, the age range δ is set to 5 years. Compute the ASFR for Sweden and Kenya as well as the entire world for each of the two periods. Store the resulting ASFRs separately for each region. What does the pattern of these ASFRs say about reproduction among women in Sweden and Kenya?

Answer 2

We first create a function that will compute the age specific fertility rate given a data set.

```r
# Function to compute Age specific fertility rate (ASFR)
compute_asfr <- function(pop_data) {
  pop_data %>%
  mutate(asfr = births / py.women)
}
```

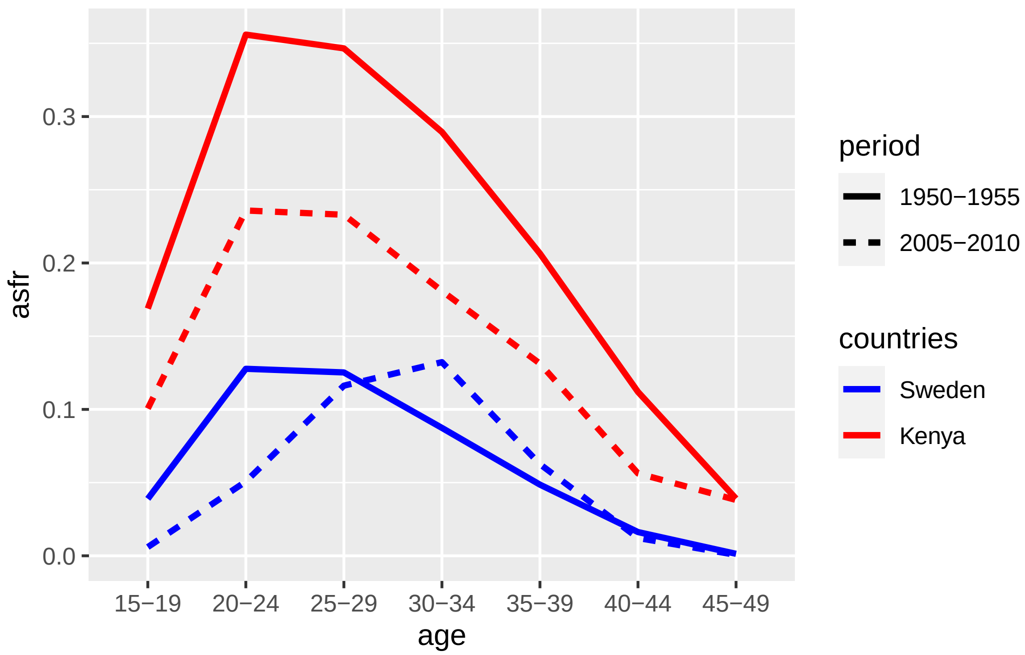
Since we are only interested in the reproductive age range of $[15, 50)$. We will compute the ASFR for Sweden, Kenya, and the world on data including only these age ranges.

```r
# Compute ASFR for each data set
world_asfr  <- compute_asfr(
  subset(
    world_data,
    age %in% unique(sweden_data$age)[4:10]
    )
  )
```

```r
kenya_asfr  <- compute_asfr(
  subset(
    kenya_data,
    age %in% unique(sweden_data$age)[4:10]
    )
  )
sweden_asfr <- compute_asfr(
  subset(
    sweden_data,
    age %in% unique(sweden_data$age)[4:10]
    )
  )
```

Because it is difficult to compare ASFRs across two time periods for Sweden and Kenya, we create the following line plot.

```r
# Create plot to compare asfr between Kenya and Sweden across two time periods
ggplot() +
  geom_line(data = kenya_asfr,
            aes(
              x = age,
              y = asfr,
              group = period,
              color = "red",
              linetype = period
              ),
            linewidth = 1.1) +
  geom_line(data = sweden_asfr,
            aes(
              x = age,
              y = asfr,
              group = period,
              color = "blue",
              linetype = period
              ),
            linewidth = 1.1) +
  scale_color_manual(labels = c("Sweden", "Kenya"),
                     values = c("blue", "red")) +
  guides(color = guide_legend("countries"))
```



Looking at the plot, we see that for both of the time periods, Kenya has a higher ASFR than Sweden. This difference is smaller in 2005-2010 than in 1950-1955. We also observe that for Kenya, the ASFR decreased from the period 1950-1955 to the period 2005-2010 for all age ranges. On the other hand, we see that for Sweden, the fertility rate is highest for the age range 20-24 in 1950-1955, while it is the highest for the age range 30-34 in 2005-2010.

Question 3

Using the ASFR, we can define the TFR as the average number of children women give birth to if they live through their entire reproductive age.

$$\text{TFR} = \text{ASFR}_{[15,\ 20)} \times 5 + \text{ASFR}_{[20,\ 25)} \times 5 + \cdots + \text{ASFR}_{[45,\ 50)} \times 5$$

We multiply each age-specific fertility rate by 5 because the age range is 5 years. Compute the TFR for Sweden and Kenya as well as the entire world for each of the two periods. As in the previous question, continue to assume that women's reproductive age range is $[15, 50)$. Store the resulting two TFRs for each country or the world as a vector of length two. In general, how has the number of women changed in the world from 1950 to 2000? What about the total number of births in the world?

Answer 3

We start by creating a function that will compute the TFR given a data set

```
# Function to compute the total fertility rate (TFR)
compute_tfr <- function (pop_data) {
  pop_data %>%
  group_by(period) %>%
  summarise(tfr = 5 * sum(asfr)) %>%
  pull()
}
```

We then use the above function to compute the TFR for Kenya, Sweden, and the world.

```
# Compute the TFR for each data set
world_tfr  <- compute_tfr(world_asfr)
kenya_tfr  <- compute_tfr(kenya_asfr)
sweden_tfr <- compute_tfr(sweden_asfr)

# Combine TFR into a table
tfr <- rbind(world_tfr, kenya_tfr, sweden_tfr)
row.names(tfr) <- c("World", "Kenya", "Sweden")
colnames(tfr) <- c("1950-1955", "2005-2010")
knitr::kable(tfr)
```

	1950-1955	2005-2010
World	5.007248	2.543623
Kenya	7.591410	4.879569
Sweden	2.226917	1.902764

We now take a look at the change in the number of women and the total number of births in the world from 1950 to 2000. We compute the change as the ratio of the total number in 1950 to the total number in 2000.

```
# Compute totals of women and births in the world by period
world_data %>%
  group_by(period) %>%
  summarise(
    total_women = sum(py.women),
    total_births = sum(births)
    ) ->
```

```r
  totals_world

# Compare how much these totals have changed
changes_totals <- totals_world[2, -1] / totals_world[1, -1]

knitr::kable(totals_world)
```

period	total_women	total_births
1950-1955	6555686	488891.5
2005-2010	16554781	674581.3

Based on the table, we see that TFR for Kenya, Sweden, and the world decreased from the period 1950-1955 to the period 2005-2010. To further investigate this change, we see the number of women in the world increased by 2.53 times from 1950 to 2000 and the number of births in the world increased by 1.38 times from 1950 to 2000. Even though both numbers are increasing from 1950 to 2000, the number of births is increasing at a slower rate than the number of women, resulting in a decrease in ASFRs from 1950 to 2000 in the world.

Question 4

Next, we will examine another important demographic process: death. Compute the *crude death rate* (CDR), which is a concept analogous to the CBR, for each period and separately for each region. Store the resulting CDRs for each country and the world as a vector of length two. The CDR is defined as:

$$\text{CDR} = \frac{\text{number of deaths}}{\text{number of person-years lived}}$$

Briefly describe patterns you observe in the resulting CDRs.

Answer 4

We first create a function that will compute the CDR given a data set.

```r
# Function to compute the Crude death rate (CDR)
compute_cdr <- function(pop_data) {
  pop_data %>%
  group_by(period) %>%
  summarise(cbr = sum(deaths) / sum(py)) %>%
  pull()
}
```

We then use this function to compute the CDR for Kenya, Sweden, and the world.

```
# Compute the CDR for each data set
world_cdr  <- compute_cdr(world_data)
kenya_cdr  <- compute_cdr(kenya_data)
sweden_cdr <- compute_cdr(sweden_data)

# Combine CDR into a table
cdr <- rbind(world_cdr, kenya_cdr, sweden_cdr)
row.names(cdr) <- c("World", "Kenya", "Sweden")
colnames(cdr) <- c("1950-1955", "2005-2010")
knitr::kable(cdr)
```

	1950-1955	2005-2010
World	0.0193189	0.0081661
Kenya	0.0239625	0.0103891
Sweden	0.0098448	0.0099685

We see that for Kenya and the whole world, crude death rate decreased from the period 1950-1955 to the period 2005-2010 while the CDR for Sweden from 1950-1955 to 2005-2010 were close to each other. Within the period 1950-1955, we observe that Kenya has a higher CDR than the world and Sweden had a lower CDR than the world. On the other hand, for the period 2005-2010, we observe that Kenya and Sweden both have a higher CDR than the world.

Question 5

One puzzling finding from the previous question is that the CDR for Kenya during the period of 2005-2010 is about the same level as that for Sweden. We would expect people in developed countries like Sweden to have a lower death rate than those in developing countries like Kenya. While it is simple and easy to understand, the CDR does not take into account the age composition of a population. We therefore compute the *age specific death rate* (ASDR). The ASDR for age range $[x, x + \delta)$ is defined as:

$$\text{ASDR}_{[x,\ x+\delta)} \ = \ \frac{\text{number of deaths for people of age } [x,\ x+\delta)}{\text{number of person-years of people of age } [x,\ x+\delta)}$$

Calculate the ASDR for each age group, separately for Kenya and Sweden, during the period of 2005-2010. Briefly describe the pattern you observe.

Answer 5

We start by creating a function that computes the ASDR given a data set.

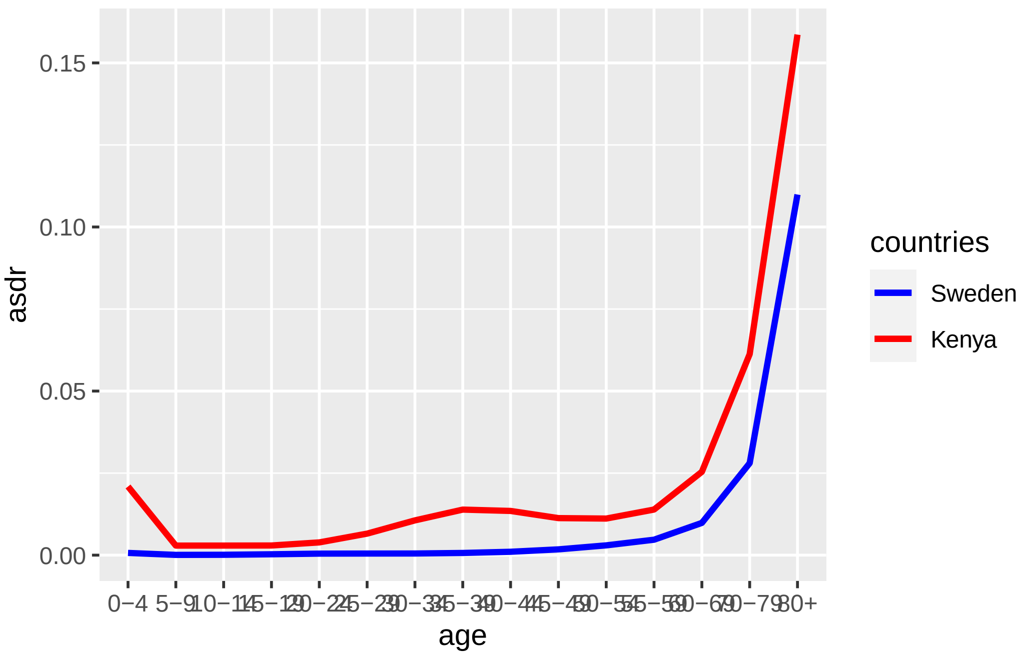
```r
# Function to compute Age specific death rate (ASDR)
compute_asdr <- function(pop_data) {
  pop_data %>%
  mutate(asdr = deaths / py)
}
```

We then use the above function to compute the ASDR for Kenya, and Sweden during the time period of 2005-2010.

```r
# Compute ASDR for each data set
kenya_asdr  <- compute_asdr(
  subset(
    kenya_data,
    period %in% c("2005-2010")
    )
  )
sweden_asdr <- compute_asdr(
  subset(
    sweden_data,
    period %in% c("2005-2010")
    )
  )
```

Because it is difficult to compare ASDRs for Sweden and Kenya across all of the age groups, we create the following line plot.

```r
# Create plot to compare the asdr between Kenya and Sweden for 2005-2010
ggplot() +
  geom_line(data = kenya_asdr,
            aes(x = age, y = asdr, group = period, color = "red"),
            linewidth = 1.1) +
  geom_line(data = sweden_asdr,
            aes(x = age, y = asdr, group = period, color = "blue"),
            linewidth = 1.1) +
  scale_color_manual(labels = c("Sweden", "Kenya"),
                     values = c("blue", "red")) +
  guides(color = guide_legend("countries"))
```



From the plot, we see that across all of the age groups, the ASDR for Kenya is higher than that of Sweden. This difference is larger for age groups 0-4, 70-79, and 80+.

Question 6

One way to understand the difference in the CDR between Kenya and Sweden is to compute the counterfactual CDR for Kenya using Sweden's population distribution (or vice versa). This can be done by applying the following alternative formula for the CDR.

$$\text{CDR} \;=\; \text{ASDR}_{[0,5)} \times P_{[0,5)} + \text{ASDR}_{[5,10)} \times P_{[5,10)} + \cdots$$

where $P_{[x,x+\delta)}$ is the proportion of the population in the age range $[x, x+\delta)$. We compute this as the ratio of person-years in that age range relative to the total person-years across all age ranges. To conduct this counterfactual analysis, we use $\text{ASDR}_{[x,x+\delta)}$ from Kenya and $P_{[x,x+\delta)}$ from Sweden during the period of 2005-2010. That is, first calculate the age-specific population proportions for Sweden and then use them to compute the counterfactual CDR for Kenya. How does this counterfactual CDR compare with the original CDR of Kenya? Briefly interpret the result.

Answer 6

We start by creating a function that will compute the age-specific population proportion given a data set

```
# Function to compute population proportion by period
compute_pop_prop <- function(pop_data) {
  pop_data %>%
  group_by(period) %>%
  mutate(pop_prop = py / sum(py)) %>%
  ungroup()
}
```

We then use the above function to compute the age-specific population proportions for Sweden.

```
# Compute population proportion for each data set
kenya_asdr <- compute_pop_prop(kenya_asdr)
sweden_asdr <- compute_pop_prop(sweden_asdr)
```

Using this information, we compute the counterfactual CDR for Kenya.

```
# Compute Kenyas CDR when Kenya had Sweden's population distribution
mutate(kenya_asdr, temp_cdr = asdr * sweden_asdr$pop_prop) %>%
 group_by(period) %>%
 summarise(cdr_re_sweden = sum(temp_cdr))
```

The counterfactual CDR for Kenya is 0.023, which is higher than the original CDR of 0.010. We see that using Sweden's population distribution, which consisted of a higher proportion of elderly, Kenya's CDR increased. This is consistent with the pattern shown previously, where the age specific death rate for Kenya is much higher than Sweden when looking at infants and elderly.