Understanding World Population Dynamics

Assignment 1 - PSYC593

AUTHOR
Victoria Yao

PUBLISHED
September 8, 2023

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.*

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

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.1     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.2     ✔ tibble    3.2.1
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
```

```r
# Read data ----
world_data  <- readr::read_csv("../../data/raw_data/World.csv")
```

```
Rows: 30 Columns: 7
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): country, period, age
dbl (4): births, deaths, py.men, py.women

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
kenya_data  <- readr::read_csv("../../data/raw_data/Kenya.csv")
```

```
Rows: 30 Columns: 8
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): country, period, age
dbl (5): births, deaths, py.men, py.women, l_x

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
sweden_data <- readr::read_csv("../../data/raw_data/Sweden.csv")
```

```
Rows: 30 Columns: 8
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): country, period, age
dbl (5): births, deaths, py.men, py.women, l_x

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```r
# Create new variable for total person years
# Add addinitional line in 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
```

```r
# Create the CBR function
compute_cbr <- function (population_data) {
  population_data %>%
  group_by(period) %>%
  summarise(cbr = sum(births) / sum(py)) %>%
  pull()
}

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

```
[1] 0.03732863 0.02021593
```

```r
kenya_cbr <- compute_cbr(kenya_data)
kenya_cbr
```

```
[1] 0.05209490 0.03851507
```

```r
sweden_cbr <- compute_cbr(sweden_data)
sweden_cbr
```

```
[1] 0.01539614 0.01192554
```

The CBR for the World will be: 0.03732863 in 1950 - 1955 and 0.02021593 in 2005 - 2010
The CBR for the Kenya will be: 0.05209490 in 1950 - 1955 and 0.03851507 in 2005 - 2010
The CBR for the Sweden will be: 0.01539614 in 1950 - 1955 and 0.01192554 in 2005 - 2010
It looks like the CBRs in three conditions are all becoming smaller in 2005-2010 than 1950-1955. Sweden has the smallest CBR both before and after.

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

```r
# Create function to compute Age specific fertility rate (ASFR)
compute_asfr <- function (population_data) {
  population_data %>%
  mutate(start_age = as.numeric(str_extract(age, "^\\d+"))) %>%
  filter(start_age >= 15, start_age < 50) %>%
  mutate(asfr=births / py.women)
}

# Compute ASFR for each data set
world_data <- compute_asfr(world_data)
kenya_data <- compute_asfr(kenya_data)
sweden_data <- compute_asfr(sweden_data)

# Compare ASFRs for Kenya and Sweden
kenya_data$asfr
```

```
 [1] 0.16884585 0.35596942 0.34657814 0.28946367 0.20644016 0.11193267
 [7] 0.03905205 0.10057087 0.23583536 0.23294721 0.18087964 0.13126805
[13] 0.05626214 0.03815044
```

```r
sweden_data$asfr
```

```
 [1] 0.0389089519 0.1277108826 0.1252436647 0.0873641591 0.0486037714
 [6] 0.0162101857 0.0013418290 0.0059709097 0.0507320271 0.1162085625
[11] 0.1322744621 0.0625923991 0.0121600765 0.0006143942
```

It looks like both are having a smaller ASFR in 2005-2010, but Kenya has a generally larger ASFR than Sweden.

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 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

```r
# Function to compute the total fertility rate (TFR)

compute_tfr <- function (population_data) {
  population_data %>%
  group_by(period) %>%
  summarise(tfr=5 *sum(asfr)) %>%
  pull()
}

# Compute the TFR for each data set
world_tfr  <- compute_tfr(world_data)
world_tfr
```

```
[1] 5.007248 2.543623
```

```r
kenya_tfr  <- compute_tfr(kenya_data)
kenya_tfr
```

```
[1] 7.591410 4.879568
```

```r
sweden_tfr <- compute_tfr(sweden_data)
sweden_tfr
```

```
[1] 2.226917 1.902764
```

Below is the solution for computing the total change of women and birth:

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

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

```
  total_women total_births
1    2.694017     1.379818
```

```r
# Compare what percentage do totals change
changes_totals_percent <- ((totals_world[2,-1] - totals_world[1,-1])/totals_world[1,-1]
changes_totals_percent
```

```
  total_women total_births
1    169.4017     37.98179
```

In general, totals of women in 2005-2010 has increased to around 2.5 times of what it was in 1950-1955, which is about 152.5% increase in data.

Totals of birth in 2005-2010 has increased to around 1.38 times of what it was in 1950-1955, which is about 37.98% increase in data.

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

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

# Compute the CDR
world_cdr  <- compute_cdr(world_data)
world_cdr
```

```
[1] 0.007560667 0.002669479
```

```r
kenya_cdr  <- compute_cdr(kenya_data)
kenya_cdr
```

```
[1] 0.009272978 0.007324122
```

```r
sweden_cdr <- compute_cdr(sweden_data)
sweden_cdr
```

```
[1] 0.001812375 0.000751132
```

All three regions are having a 2005-2010 death rate smaller than the one in 1950-1955. However, Sweden seems to have a least decrease in the death rate with only 0.00012 difference between the data. Among three regions, Kenya seems to have the largest death rate no matter in 1950-1955 or in 2005-2010.

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

```r
# Function to compute Age specific death rate (ASDR)
compute_asdr <- function (population_data) {
  population_data %>%
  mutate(period_time = as.numeric(str_extract(period, "^\\d+"))) %>%
  filter(period_time >= 2005) %>%
  mutate(asdr=deaths/py)
}

# Compute ASDR for each data set
world_data  <- compute_asdr(world_data)
kenya_data  <- compute_asdr(kenya_data)
sweden_data <- compute_asdr(sweden_data)

#Show the ASDR data
world_data$asdr
```

```
[1] 0.001302818 0.001832602 0.002278500 0.002623982 0.003031563 0.003753402
[7] 0.005085583
```

```r
kenya_data$asdr
```

```
[1] 0.002942986 0.003885368 0.006558131 0.010603913 0.013881062 0.013474598
[7] 0.011288057
```

```r
sweden_data$asdr
```

```
[1] 0.0002687775 0.0004697344 0.0004941440 0.0005057066 0.0006689578
[6] 0.0010392562 0.0017696213
```

An interesting pattern is that in World and Kenya, the newborns (aged 0-4) seem to have higher death rates than the rest of at least 30 years; Swede newborn dearth rates is also much higher but then drops when it comes to 5-9 years old. Except newborn death rate, a gradual increasing pattern is observed in all three regions, and Kenya seems to have highest death rate in almost every period compared with World and Sweden.

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

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

# Compute population proportion for each data set
world_data <- compute_pop_prop(world_data)
kenya_data <- compute_pop_prop(kenya_data)
sweden_data<- compute_pop_prop(sweden_data)

# Compute Kenya CDR Kenya had Swede population distribution
kenya_cdrresweden <- mutate(kenya_data,
                     temp_cdr  = asdr * sweden_data$popP) %>%
                     group_by(period) %>%
                     summarise(cdrresweden = sum(temp_cdr))
kenya_cdrresweden
```

```
# A tibble: 1 × 2
  period    cdrresweden
  <chr>           <dbl>
1 2005-2010     0.00909
```

The counterfactual CDR is actually higher than the orignial CDR in Kenya in 2005-2010, meaning that given the same age distribution as Sweden, Kenya should have a higher CDR than the orignial one. Although the original CDR is lower than the conterfactual one, it is still higher than the Sweden original one.