Barcelona Summer School of Demography

*Module 2. Demography with R*

# 3. Standardization and decomposition

Standardization and decomposition

Tim Riffe
tim.riffe@gmail.com

7 July 2021

## Contents

# 1 Summary

Today we'll look at ways to make data more comparable (standardization) and ways to explain differences between summary measures (decomposition). As per the previous days, this material was originally prepared by the one-and-only Marie-Pier Bergeron-Boucher, credit is due to her for organizing the logic and rigor of this lesson. My contributions have been light edits to the text, occasional insertions where I though they would help, and doing an full overhaul of the code to a tidy approach.

# 2 Data

We will compare mortality in Taiwan and Japan. I downloaded their mortality rates from the HMD using the *demography* package (Hyndman et al. 2017) and tidified them using the approach from yesterday. I save you having to replicate that code and have posted the data as a `csv` on the github site. You can read it directly into `R`, below.

We *might* use the `DemoDecomp` package today if there's time and someone wants a demonstration of generalized decomposition. Just in case, feel free to install this, though it isn't strictly required for the prepared lesson.

```
install.packages("DemoDecomp")
```

I sometimes make updates to it without pushing to the main `R` repositories, so you could also get a more up to date version of the package here, if so inclined

```
install.packages("remotes")
library(remotes)
install_github("timriffe/DemoDecomp")
```

Get the data and load our beloved packages:

```
library(tidyverse)
library(readr)
library(DemoDecomp)
# will copy this link into the google doc too
DAT <- read_csv("https://raw.githubusercontent.com/timriffe/BSSD2021Module2/master/03_wednes
```

# 3 Standardization

Standardization is a commonly used procedure when comparing rates or probabilities for groups with differences in composition. This procedure is used to avoid the confounding effect of the population structure by simply equalizing structure for all groups.

## 3.1 Problems with crude measures

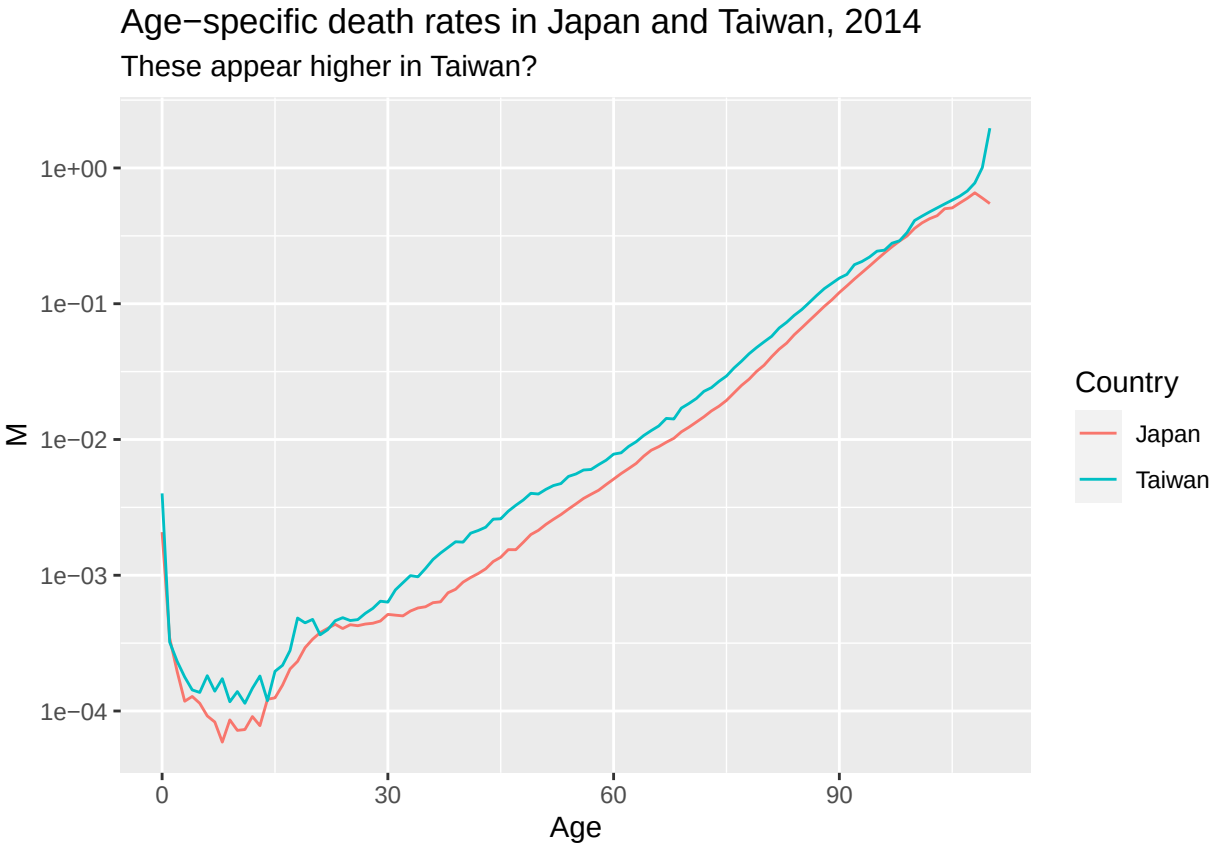
Let's start by comparing the crude mortality rates in Japan and Taiwan in 2014.

```
DAT %>%
  filter(Sex == "total",
         Year == 2014) %>%
  mutate(Deaths = M * Exposure) %>%
  group_by(Country) %>%
  summarize(CDR = sum(Deaths) / sum(Exposure))
```

```
## # A tibble: 2 x 2
##   Country     CDR
##   <chr>     <dbl>
## 1 Japan   0.0101
## 2 Taiwan  0.00698
```

Japan has a higher CDR than Taiwan, which *the many* would interpret as Japan having a higher mortality than Taiwan. However, if we look at the age-specific death rates, we have a different story.

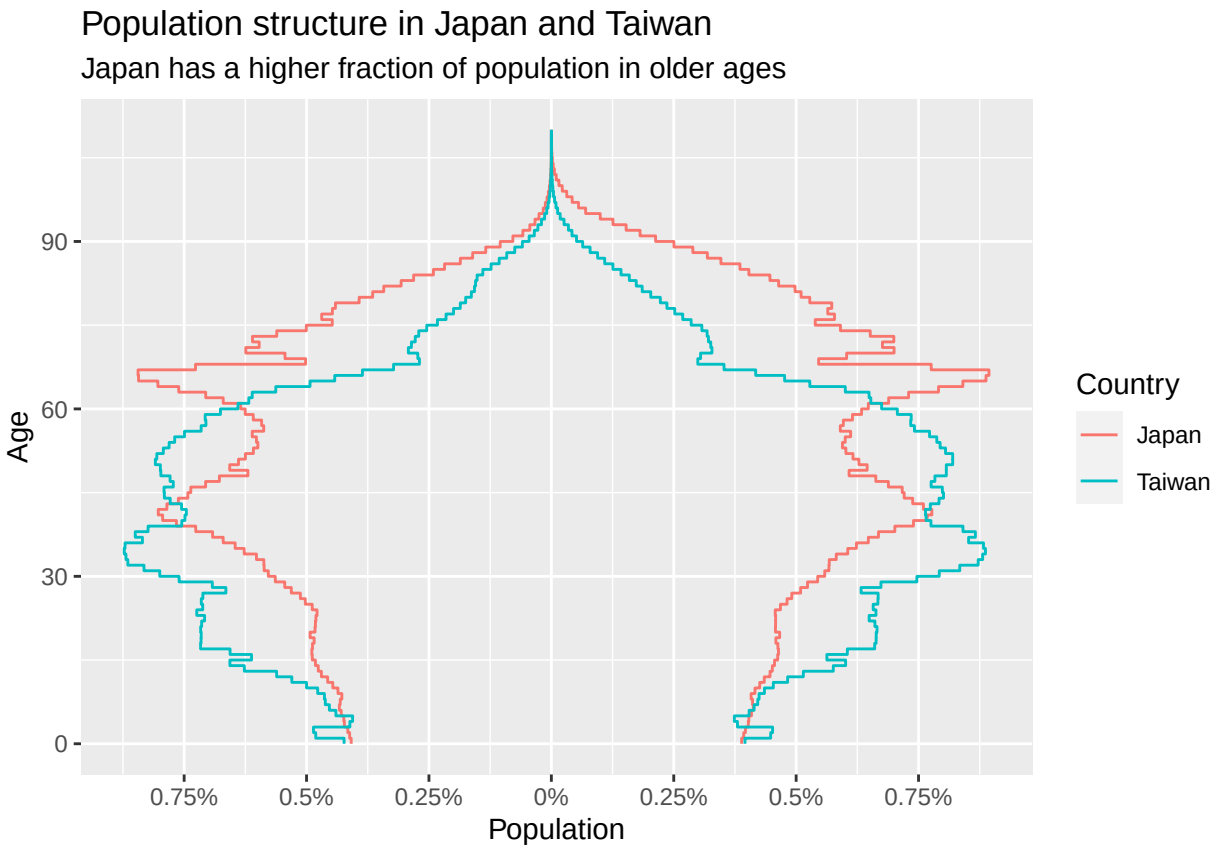
```
# Age-specific death rates
DAT %>%
  filter(Sex == "total",
         Year == 2014) %>%
  ggplot(aes(x = Age, y = M, color = Country)) +
  geom_line() +
  scale_y_log10() +
  labs(title = "Age-specific death rates in Japan and Taiwan, 2014",
       subtitle = "These appear higher in Taiwan?")
```



Here, we see that Japan has lower age-specific death rates than Taiwan at all ages, despite having a higher CDR. This occurs because 1) mortality has a strong age gradient: stronger than the international differences in this comparison, and 2) therefore the CDR is very sensitive to the population age structure, which is acting as an *implicit* weight.

```
breaks = seq(-0.01, 0.01, 0.0025)

DAT %>%
  filter(Year == 2014,
         Sex != "total") %>%
  group_by(Country) %>%
  mutate(Structure = Exposure / sum(Exposure),
         Population = ifelse(Sex == "male", -Structure, Structure)) %>%
  ungroup() %>%
  ggplot(aes(x = Age, y = Population, color = Country, group = interaction(Sex, Country))) +
  geom_step() +
  coord_flip() +
  scale_y_continuous(breaks = seq(-0.01, 0.01, 0.0025),
                     labels = paste0(as.character(
                         c(seq(.01, 0, -.0025), seq(0.0025, 0.01, 0.0025))*100), "%")) +
  labs(title = "Population structure in Japan and Taiwan",
       subtitle = "Japan has a higher fraction of population in older ages")
```




The age pyramids indicate that Japan has an older age structure than Taiwan. In 2014, 26% of Japanese population was aged 65 years old or higher, compared with 12% in Taiwan. As death rates are much higher at older ages than at younger age, older population will tend to have a

higher CDR than younger population.

### 3.2 Direct standardization

To avoid the confounding effect of population structure (e.g. age structure) when comparing rates, direct standardization can be used. This method allows us to estimate what the crude rate *would be* if both populations had the same structure.

An important relation between structure-specific rates ($r_c$) and crude rates (R) is:

$$R = \sum_{c}^{\infty} r_c s_c \tag{1}$$

where $s_c$ is the population structure by component $c$ (for example age, or age and sex). For the crude death rates,

$$CDR = \frac{\sum D_x}{\sum P_x} = \sum_{x}^{\infty} m_x s_x$$

.

where $s_x = \frac{P_x}{\sum(P_x)}$, i.e. the population structure net of its size.

The direct standardization method consists in:

- Finding a *standard* structure ($s_c^A$), e.g. an average structure between the population compared or the structure of one of these populations.
- Multiplying the component-specific rates ($r_c$) of the studied population by the standard structure.
- The standardized crude rates are found by summing $s_c^A r_c$

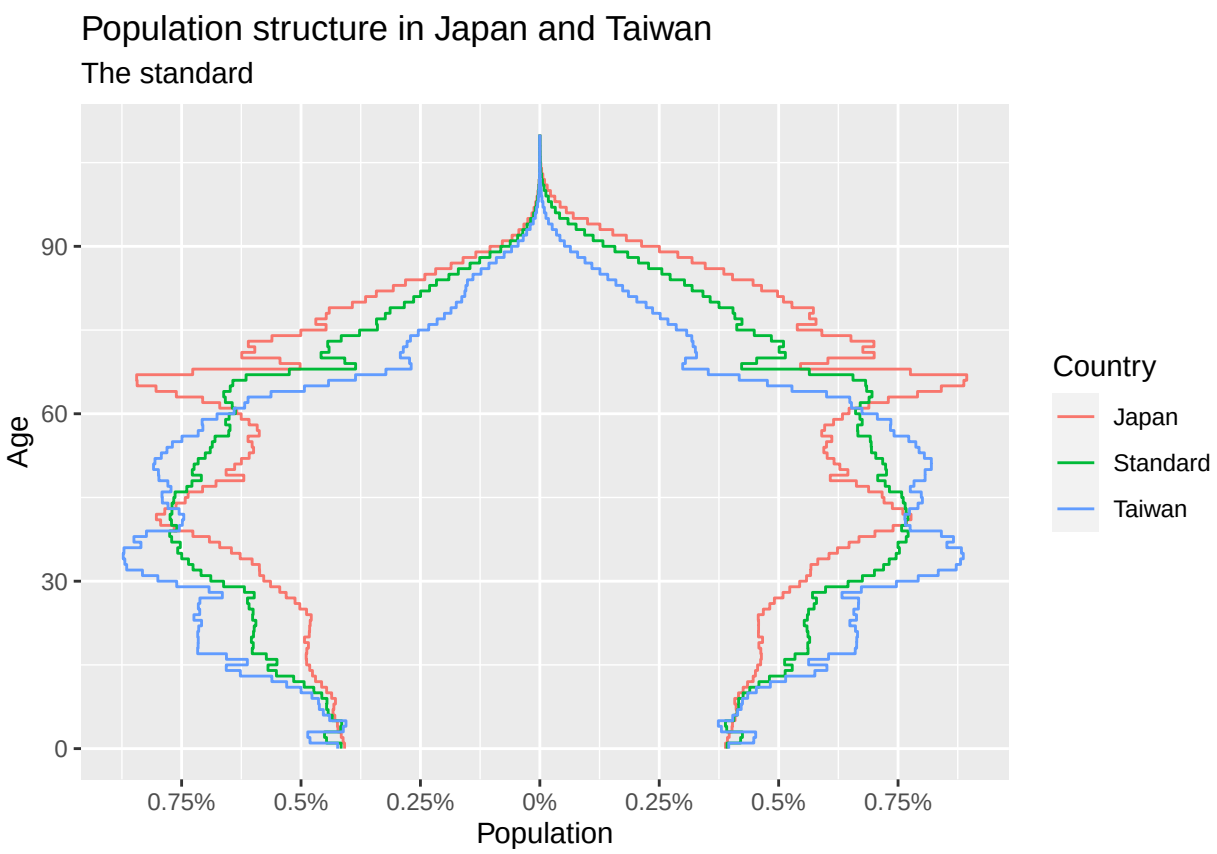
```
# Standardizing CDR of Taiwan and Japan

breaks = seq(-0.01, 0.01, 0.0025)

# calculate structure for each country
DAT2 <-
  DAT %>%
  filter(Year == 2014,
         Sex != "total") %>%
  group_by(Country) %>%
  mutate(Structure = Exposure / sum(Exposure))

# average structure (within age and sex) to get the standard
ST <-
  DAT2 %>%
  group_by(Age, Sex) %>%
  summarize(Structure = mean(Structure),
            .groups= "drop") %>%
  mutate(Country = "Standard")
```

```r
# stick together and plot to examine the standard against the original structure
DAT2 %>%
  bind_rows(ST) %>%
  mutate(Population = ifelse(Sex == "male", -Structure, Structure)) %>%
  ggplot(aes(x = Age, y = Population, color = Country, group = interaction(Sex, Country))) +
  geom_step() +
  coord_flip() +
  scale_y_continuous(breaks = seq(-0.01, 0.01, 0.0025),
                     labels = paste0(as.character(
                         c(seq(.01, 0, -.0025), seq(0.0025, 0.01, 0.0025))*100), "%")) +
  labs(title = "Population structure in Japan and Taiwan",
       subtitle = "The standard")
```




```r
# Step 2: Find the standardized CDR

# removing Country so we can join on age and sex
# Call structure Standard instead, because that's what
# it is now!
# Also, we're just standardizing on total, the sex-specific part was just for
# the pyramid, so we can aggregate it out. Alternatively we could recalculate it
# using total from DAT, but I found this more expedient.
ST2 <- ST %>%
  select(-Country, Standard = Structure, Age) %>%
  group_by(Age) %>%
  summarize(Standard = sum(Standard))
```

```r
# Filter down to our year, join the standard to it,
# then calculate within countries
DAT %>%
  filter(Year == 2014,
         Sex == "total") %>%
  mutate(Deaths = M * Exposure) %>%
  left_join(ST2, by= c("Age")) %>%
  group_by(Country) %>%
  summarize(CDR = 1000 * sum(Deaths) / sum(Exposure),
            ASDR = 1000 * sum(M * Standard))
```

```
## # A tibble: 2 x 3
##   Country   CDR  ASDR
##   <chr>   <dbl> <dbl>
## 1 Japan    10.1   7.45
## 2 Taiwan    6.98 10.6
```

After standardization, Japan has a lower CDR than Taiwan, the CDR being now consistent with what observed at the age-specific level.

### 3.3 Indirect standardization

The indirect standardization is used to estimate what would be the crude rates if both populations had the same component-specific rates. This method allows quantifying the effect of population structure on mortality.

The method consists in:

- Finding *standard* component-specific rates ($r_c^A$).
- Multiplying the population structures ($s_c$) of the studied population by the standard component-specific rates.
- The standardized crude rates are found by summing $s_c r_c^A$

```r
# Step 1: Find the standard age-specific rates
# We here use the average
STrates <-
  DAT %>%
  filter(Year == 2014,
         Sex == "total") %>%
  group_by(Age) %>%
  summarize(M_standard = mean(M))

DAT %>%
  filter(Year == 2014,
         Sex == "total") %>%
  left_join(STrates, by = "Age") %>%
  group_by(Country) %>%
  summarize(CDR = 1000 * sum(Exposure * M) / sum(Exposure),
            ASDRi = 1000 * sum(Exposure * M_standard) / sum(Exposure))
```

```
## # A tibble: 2 x 3
##   Country   CDR ASDRi
##   <chr>   <dbl> <dbl>
## 1 Japan   10.1  12.2
## 2 Taiwan   6.98  5.87
```

# 4 Decomposition methods

Decomposition methods are common tools in demography, used to understand differences in a demographic measure between two or more populations. These methods allow quantifying the exact contribution of specific components, such as ages and causes of death, to this difference between populations.

## 4.1 Kitagawa decomposition: Decomposing differences in crude rates

Kitagawa decomposition (Kitagawa 1955) aims at quantifying how much of the difference between two crude rates is due to composition effects (e.g. difference in the age-structures) and how much is due to differences in the component-specific rates.

The Kitagawa decomposition (Kitagawa 1955) was the first to decompose the difference between two rates by a composition effect and a rate effect, using multiple standardizations. It brings together both direct and indirect standardizations.

For example, when applied to the CDR, the decomposition is written as:

$$CDR^J - CDR^T = \underbrace{\sum_x^\infty (m_x^J - m_x^T)(\frac{s_x^J + s_x^T}{2})}_{RE\text{: } rate\ effect} + \underbrace{\sum_x^\infty (s_x^J - s_x^T)(\frac{m_x^J + m_x^T}{2})}_{CE\text{: } composition\ effect}$$

The left hand side of the equation (named RE) captures how much of the difference in the CDR between Japan and Taiwan is due to difference in age-specific death rates ($m_x$). This is the same process as finding the difference between the two crude rate after direct standardization, using the average population structure as standard.

The right hand side of the equation (named CE) captures how much of the difference in the CDR is due to age-structure ($s_x$) differences. This is the same process as finding the difference between the two crude rate after indirect standardization, using the average age-specific rate as standard.

```
# Get data in convenient format for side-by side calcs
DAT_Dec <-
  DAT %>%
  filter(Year == 2014,
         Sex == "total") %>%
  group_by(Country) %>%
  mutate(Sx = Exposure / sum(Exposure)) %>%
  ungroup() %>%
  select(-Exposure) %>%
  pivot_wider(names_from = Country, values_from = c(M, Sx))
```

```
DAT_Dec %>%
  mutate(
    # calculate standards
         M_st = (M_Taiwan + M_Japan) / 2,
         Sx_st = (Sx_Taiwan + Sx_Japan) / 2,
    # weight differences
         RE = (M_Japan - M_Taiwan) * Sx_st,
         CE = (Sx_Japan - Sx_Taiwan) * M_st) %>%
    # summarize decomp results, compare with original CDR
  summarize(RE = sum(RE),
            CE = sum(CE),
            CDR_Japan = sum(M_Japan * Sx_Japan),
            CDR_Taiwan = sum(M_Taiwan * Sx_Taiwan)) %>%
  mutate(CDR_diff = CDR_Japan - CDR_Taiwan)
```

```
## # A tibble: 1 x 5
##         RE      CE CDR_Japan CDR_Taiwan CDR_diff
##      <dbl>   <dbl>     <dbl>      <dbl>    <dbl>
## 1 -0.00313 0.00628    0.0101    0.00698  0.00316
```

The CDR is only one of few measures that can be decomposed with the Kitagawa method. The CBR, GFR, survival rates/probabilities, neonatal mortality rates, to names only a few, can also be decomposed using this method, as long as the relation between components-specific rates and the components structure, as expressed in equation (1), holds. The components can be age, socioeconomic status, race, etc.

More than one structure/composition effects can also be included. For more information see Kitagawa (1955) and Gupta (1978).

## 4.2 Arriaga decomposition: Decomposing differences in life expectancy

The Arriaga method (Arriaga 1984) allows to decompose the difference in life expectancy by age.

The method is based on survival probabilities ($l_x$) and person-years ($_nL_x$ and $T_x$) in the life table. We will calculate a lifetable as from Class 2.

```
radix = 1
LT <-
  DAT %>%
  filter(Year == 2014, Sex == "total") %>%
  group_by(Country) %>%
  mutate(M = ifelse(is.na(M), .5, M),          # hack
         n = 1,
         ax = case_when(
                Age == 0 & M < .02012 ~ .14916 - 2.02536 * M,
                Age == 0 & M < .07599 ~ 0.037495 + 3.57055 * M,
                Age == 0 & M >= .07599 ~ 0.30663,
                Age == 110 ~ 1 / M,
                TRUE ~ n / 2),
         ax = ifelse(is.infinite(ax),.5,ax),   # hack
```

```
        qx = (M * n) / (1 + (n - ax) * M),
        qx = ifelse(qx > 1, 1, qx),          # hack
        px = 1 - qx,
        lx = radix * c(1, cumprod(px[-n()])),
        dx = qx * lx,
        Lx = n * lx - (n - ax) * dx,
        Tx = Lx %>% rev() %>% cumsum() %>% rev(),
        ex = Tx / lx,
        ex = ifelse(is.nan(ex),ax,ex))  %>% # hack
  ungroup()
```

The difference in life expectancy between Japan and Taiwan is greater than 4 years. The Arriaga method can help figure out which ages (or age-groups) contribute to this difference.

```
#step 2: find the difference in life expectancy
LT %>%
  filter(Age == 0) %>%
  select(Country, ex)
```

```
## # A tibble: 2 x 2
##   Country    ex
##   <chr>   <dbl>
## 1 Taiwan   79.5
## 2 Japan    83.7
```

Let's select just the columns we'll need, and move them side by side, like before:

```
LT_arriaga <-
  LT %>%
  select(Country, Age, lx, Lx, Tx) %>%
  pivot_wider(names_from = Country, values_from = c(lx, Lx, Tx))
```

The method goes in two steps:

1) Find the direct effect.

The direct effect quantifies how much the difference in the number of years lived between age $x$ and $x+n$ contributes to the difference in life expectancy. It is the "*change in life years within a particular age group as a consequence of the mortality change in that age group*" (Arriaga 1984).

$$ {}_nD_x = \frac{l_x^T}{l_0^T}\left(\frac{{}_nL_x^J}{l_x^J} - \frac{{}_nL_x^T}{l_x^T}\right) $$

2) Finding the indirect effect

The indirect effect (and interaction effect) is the "*number of person-years added to a given life expectancy because the mortality change, within a specific age group, will produce a change in the number of survivors at the end of the age interval.*"(Arriaga 1984)

$$ {}_nI_x = \frac{T_{x+n}^J}{l_0^T}\left(\frac{l_x^T}{l_x^J} - \frac{l_{x+n}^T}{l_{x+n}^J}\right) $$

One way to do it with the tidy approach:

```
LT_arriaga <-
  LT_arriaga %>%
  mutate(direct = lx_Taiwan * (Lx_Japan / lx_Japan - Lx_Taiwan / lx_Taiwan),
         indirect = lead(Tx_Japan) *
           (lx_Taiwan / lx_Japan -
              lead(lx_Taiwan) / lead(lx_Japan)),
         # impute 0 in the final NA
         indirect = ifelse(is.na(indirect),0,indirect))
```

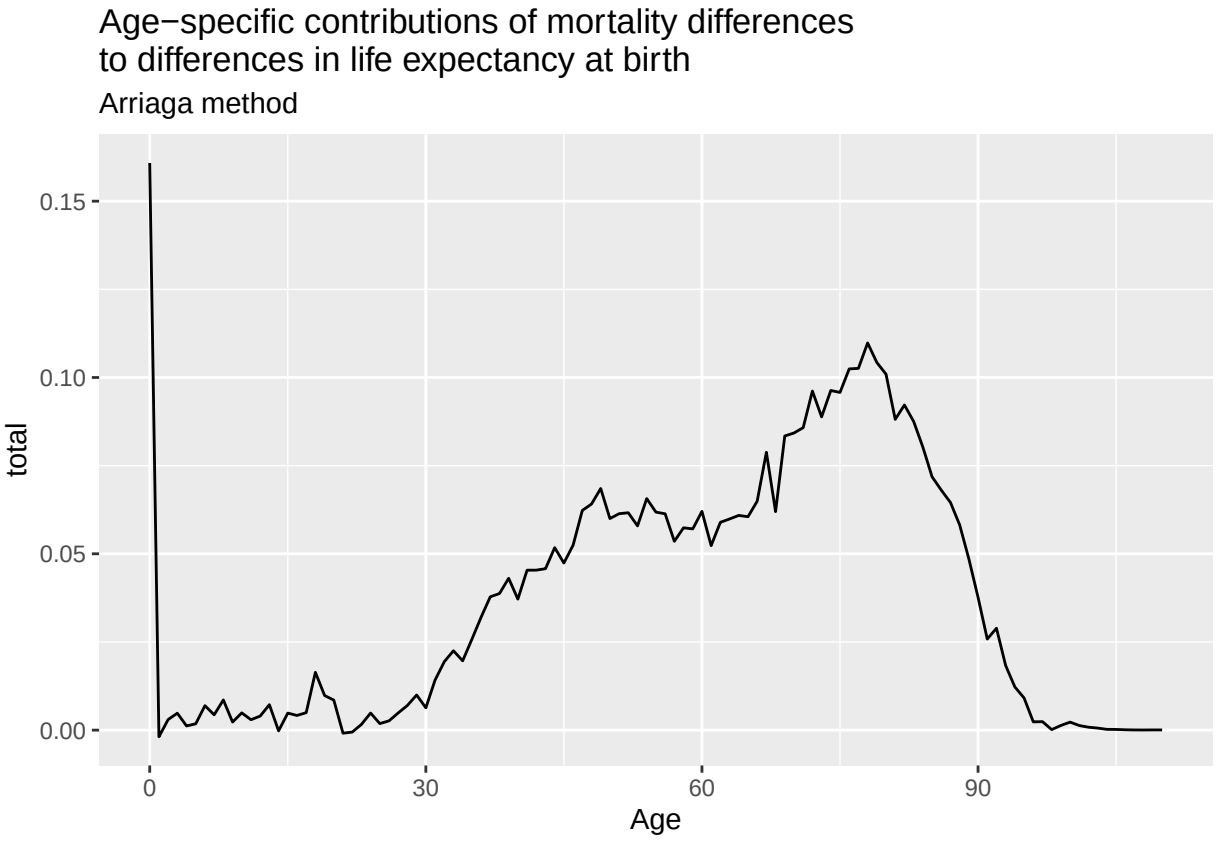
The direct and indirect contributions sum to the total differences. The Arriaga formula is then written as:

$$ {}_n\Delta_x = \frac{l_x^T}{l_0^T}\left(\frac{{}_nL_x^J}{l_x^J} - \frac{{}_nL_x^T}{l_x^T}\right) + \frac{T_{x+n}^J}{l_0^T}\left(\frac{l_x^T}{l_x^J} - \frac{l_{x+n}^T}{l_{x+n}^J}\right) $$

where ${}_n\Delta_x$ is the contribution to the difference in life expectancy at birth in age group x to x+n. The last (and open) age-interval consists only of the direct effect.

```
arriaga <-
LT_arriaga  %>%
  mutate(total = indirect + direct) %>%
  select(Age, total)

# age pattern
arriaga %>%
  ggplot(aes(x= Age, y= total)) +
  geom_line() +
  labs(title = "Age-specific contributions of mortality differences\nto differences in life
       subtitle = "Arriaga method")
```




```
# decomposition sum
arriaga$total %>% sum()
```

```
## [1] 4.182181
```

```
# it's exact!
LT %>%
  filter(Age == 0) %>%
  pull(ex) %>%
  diff()
```

```
## [1] 4.182181
```

An extension of the Arriaga method decomposing life expectancy by age AND cause of death is also available (see Preston, Heuveline, and Guillot (2001)). There are also generalized decomposition methods that you can use to decompose any function of parameters. Can be demonstrated on request.

## Exercises

Choose one country from the HMD and select 2 years (ideally over 15 years apart).

### Exercise 1

1) Create a function calculating the CDR, standardized (direct) CDR and the Kitagawa decomposition.

2) Calculate the age-specific rate effect and age-composition effect of the difference.
3) What factors allowed the CDR to decrease (or increase) over time?

### Exercise 2

1) Calculate the life table from these two years.
2) Create a function for the Arriaga decomposition.
3) Calculate the age-specific contributions for the change in life expectancy over time.
4) Plot and interpret the results.

Repeat exercises 1.2, 2.1 and 2.3 with a different time period. Try a for loop for many years using the functions you created.

## References


Arriaga, Eduardo E. 1984. "Measuring and Explaining the Change in Life Expectancies." *Demography* 21 (1): 83–96.

Gupta, Prithwis Das. 1978. "A General Method of Decomposing a Difference Between Two Rates into Several Components." *Demography* 15 (1): 99–112.

Hyndman, Rob J., Heather Booth, Leonie Tickle, and John Maindonald. 2017. *Package 'demography'*. https://cran.r-project.org/web/packages/demography/index.html.

Kitagawa, Evelyn M. 1955. "Components of a Difference Between Two Rates." *Journal of the American Statistical Association* 50 (272): 1168–94.

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