KOSTAT-UNFPA Summer Seminar on Population

Workshop 1. Demography in R

Day 7: Demographic standardization and decomposition

Instructor: Tim Riffe
tim.riffe@ehu.eus

Assistants:
Jinyeon Jo: jyjo43043@gmail.com
Rustam Tursun-Zade: rustam.tursunzade@gmail.com

4 August 2022

Contents

1 Summary

Today we will use tidy processing skills to age-standardize different demographic rates in order to make them more comparable. We will also explain some demographic decomposition approaches, and see how these work well using the function-writing and tidy processing skills that we've learned so far. Credit goes to Prof. Marie-Pier Bergeron-Boucher, for originally organizing the logic of this lesson as of 2019. I've redone the entire `R` code here, and this is now the second iteration of this material moving through me.

2 Data

We will compare mortality in USA and Japan. I downloaded their mortality rates from the HMD and combined them into a single tidy dataset. 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'll be doing some decomposition exercises today, and for some methods we'll want the `DemoDecomp` package, which implements some generalized decomposition methods in a standard way.

```
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
# I cut the url so it would fit on a pdf line...
raw_repo      <- "https://raw.githubusercontent.com/timriffe/KOSTAT_Workshop1"
specific_file <- "master/Data/Decomp_inputs.csv"
full_url      <- paste(raw_repo, specific_file,sep = "/")
M <- read_csv(full_url)
```

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 USA in 2014.

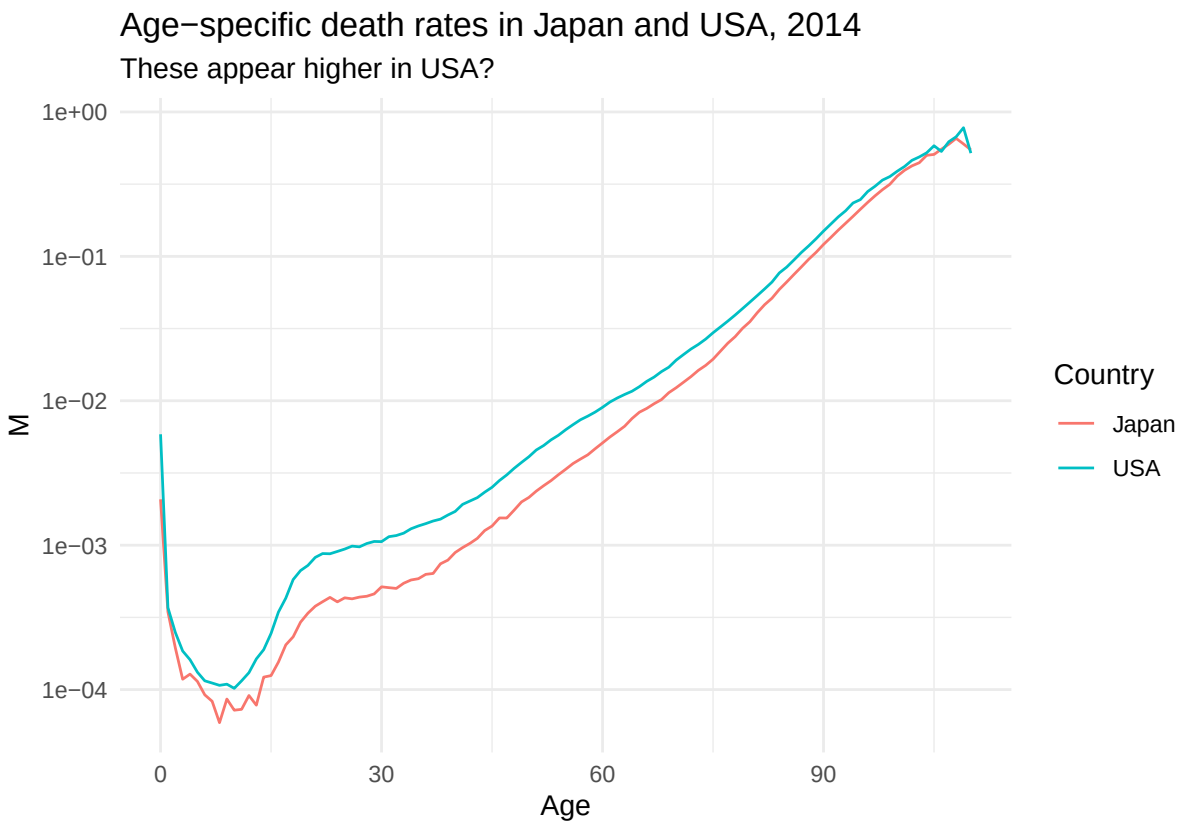
```
M %>%
  filter(Sex == "total",
```

```
         Year == 2014) %>%
  mutate(Deaths = M * Exposure) %>%
  group_by(Country) %>%
  summarize(CDR = sum(Deaths) / sum(Exposure) * 1000)
```

```
## # A tibble: 2 x 2
##   Country   CDR
##   <chr>   <dbl>
## 1 Japan    10.1
## 2 USA       8.26
```

Japan has a higher CDR than USA. On the surface it may seem as if Japan had *higher mortality* than USA. However, if we look at the age-specific death rates, we have a different story.

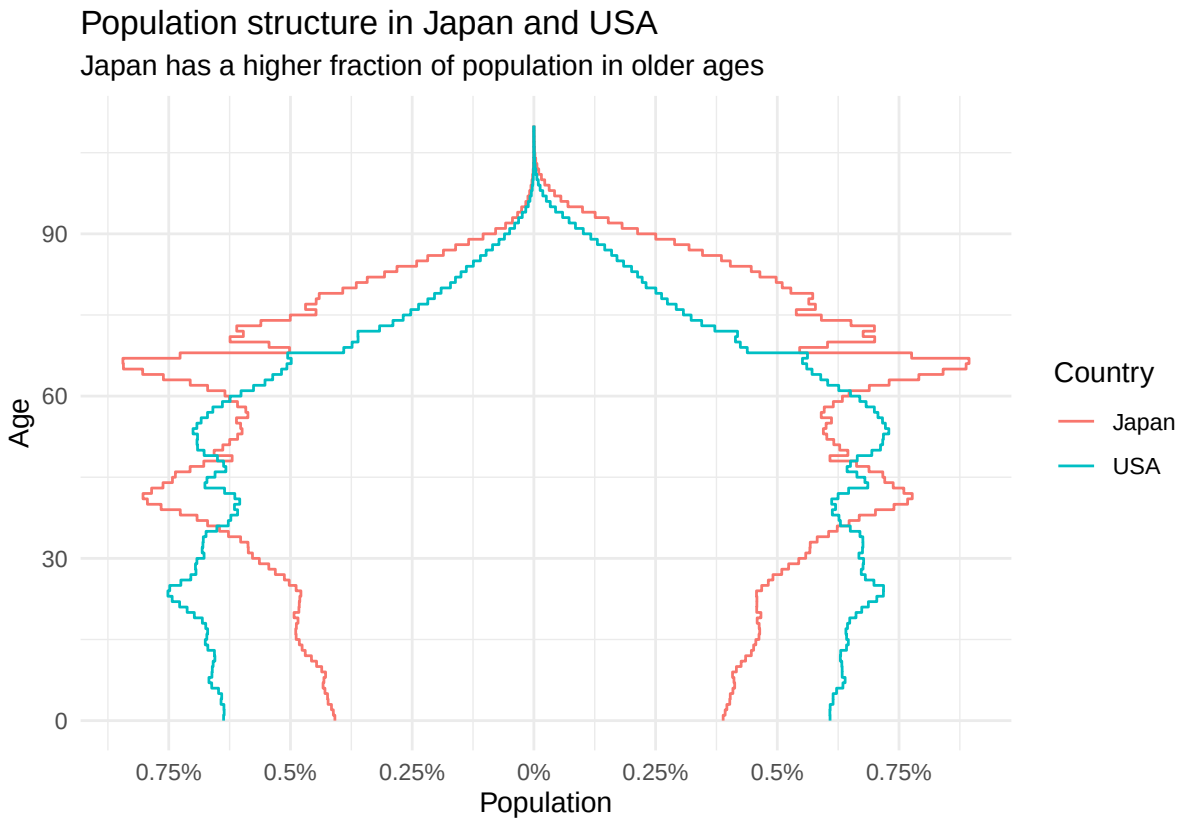
```
# Age-specific death rates
M %>%
  filter(Sex == "total",
         Year == 2014) %>%
  ggplot(aes(x = Age, y = M, color = Country)) +
  geom_line() +
  scale_y_log10() +
  labs(title = "Age-specific death rates in Japan and USA, 2014",
       subtitle = "These appear higher in USA?") +
  theme_minimal()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



Here, we see that Japan has lower age-specific death rates than USA 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 acts as weight for the CDR.

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

M %>%
  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 USA",
       subtitle = "Japan has a higher fraction of population in older ages") +
  theme_minimal()
```



The age pyramids indicate that Japan has an older age structure than USA. In 2014, 26% of Japanese population was aged 65 years old or higher, compared with 12% in USA. 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 age 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 $s_c^A r_c$

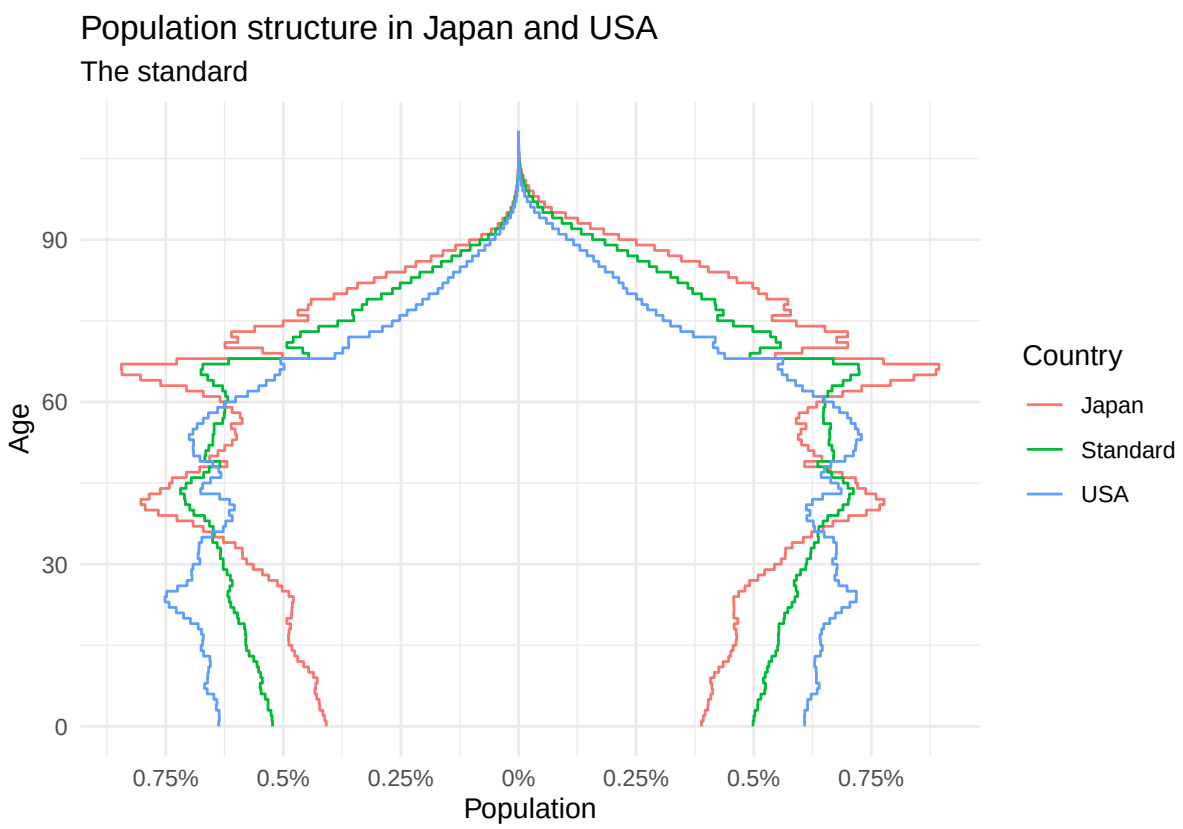
```
# calculate structure for each country
Str <-
  M %>%
  filter(Year == 2014,
         Sex != "total") %>%
  group_by(Country) %>%
  mutate(Structure = Exposure / sum(Exposure))

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

Now let's combine the country-specific and mean structures into a single data object and plot them again as a pyramid so that we see what the shared structure looks like:

```
# object to help with plotting breaks and labels
my_breaks <- seq(-0.01, 0.01, 0.0025)
Str %>%
  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 = my_breaks,
                     labels = paste0(as.character(
                         c(seq(.01, 0, -.0025), seq(0.0025, 0.01, 0.0025))*100), "%")) +
  labs(title = "Population structure in Japan and USA",
       subtitle = "The standard") +
  theme_minimal()
```




```
# Step 2: Find the standardized CDR

M %>%
  filter(Year == 2014)
```

```
## # A tibble: 666 x 6
##    Country  Year Sex      Age Exposure        M
##    <chr>   <dbl> <chr>  <dbl>    <dbl>    <dbl>
##  1 Japan    2014 female     0  488556. 0.00199
##  2 Japan    2014 female     1  493710. 0.000332
##  3 Japan    2014 female     2  498220. 0.000199
##  4 Japan    2014 female     3  505671. 0.000105
##  5 Japan    2014 female     4  506952. 0.000109
##  6 Japan    2014 female     5  513712. 0.000107
##  7 Japan    2014 female     6  518432. 0.000077
##  8 Japan    2014 female     7  516488. 0.000058
##  9 Japan    2014 female     8  512793. 0.000045
## 10 Japan    2014 female     9  522742. 0.000069
## # ... with 656 more rows
## # i Use `print(n = ...)` to see more rows
```

```
ST2 <- ST %>%
  select(-Country, Structure, Age) %>%
  group_by(Age) %>%
  summarize(Standard = mean(Structure))

# Filter down to our year, join the standard to it,
# then calculate within countries
M %>%
  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   3.96
## 2 USA      8.26  5.58
```

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

3.2.1 World standards

If you need to compare multiple populations, it may be convenient to apply an international standard population structure, of which there are many. In this case, it's *usually* preferable to apply a recognized standard rather than your own (unless you have reasons to make your own), and the steps would be really similar to what we just did. Here is a decent list and source for international standards https://seer.cancer.gov/stdpopulations/stdpopdic.html, there may of course be others. This list has standards in age groups as well as two single-age standards, as of this writing. Data are delivered in fixed width files, which we no know how to read in :-).

```
specs <-
  fwf_widths(widths = c(3,3,8),
             col_names = c("standard_code",
                           "age",
                           "pop"))
WHO_standard <-
  read_fwf("https://seer.cancer.gov/stdpopulations/stdpop.singleagesthru99.txt",
           col_positions = specs,
           show_col_types = FALSE) %>%
  filter(standard_code == "012") %>%
  mutate(age = as.integer(age),
         pop = as.integer(pop),
         pop = pop / sum(pop))
```

If you have single-age data, don't limit yourself to single-age standards if you prefer one of the others: instead you can *expand* and age-grouped standard to single ages assuming uniformity within age groups. Here's a trick that may help. (`WHO_5` is a convenient aggregation of the above in order to demonstrate the trick)

The steps imply repeating each 5-year value 5 times, then scaling down by 5 (uniform distribution within each age class). The later lines compress 0-104 into 100+.

```
head(WHO_5)
```

```
## # A tibble: 6 x 2
##     age    pop
##   <dbl>  <dbl>
## 1     0 0.0886
## 2     5 0.0869
## 3    10 0.0860
## 4    15 0.0847
## 5    20 0.0822
## 6    25 0.0793
```

```
WHO_5_expanded <- tibble(age = 0:104,
                         pop = rep(WHO_5$pop, each = 5) / 5) %>%
  mutate(age = if_else(age > 100, 100L, age)) %>%
  group_by(age) %>%
  summarize(pop = sum(pop),
            .groups = "drop")
```

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 $s_c r_c^A$

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

M %>%
  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  12.1
## 2 USA      8.26  6.98
```

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. There are two broad families of decompositions:

1. Decompositions that explain differences in terms of individual parameters of functions. For example, a difference in life expectancy can be broken down into contributions from differences in each age and cause of death.
2. Decompositions that re-express and explain differences phenomena in terms of abstractions of the raw parameters. For example, differences in life expectancy could be broken down into three additive components for *ontogenscence*, *young adult*, and *senescent* mortality. The catch is that these three components need to be estimated: they are not observed parameters. Many decompositions fall into this category. Some examples will be given in the session.

In this session, we only have time to cover the first sort of decompositon: We will see how to explain differences in functions of empirically observed demographic parameters, such as rates.

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 standardization. It brings together both *direct* and *indirect* standardization.

For example, when applied to the CDR (using `J` and `T` for Japan and USA), the decomposition is written as:

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

The left hand side of the equation (named RE) captures how much of the difference in the CDR between Japan and USA 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
M_Dec <-
  M %>%
  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))


M_Dec %>%
  mutate(
    # calculate standards
         M_st = (M_USA + M_Japan) / 2,
         Sx_st = (Sx_USA + Sx_Japan) / 2,
    # weight differences
         RE = (M_Japan - M_USA) * Sx_st,
         CE = (Sx_Japan - Sx_USA) * M_st) %>%
    # summarize decomp results, compare with original CDR
  summarize(RE = sum(RE) * 1000,
            CE = sum(CE) * 1000,
            CDR_Japan = sum(M_Japan * Sx_Japan) * 1000,
            CDR_USA = sum(M_USA * Sx_USA) * 1000) %>%
  mutate(CDR_diff = CDR_Japan - CDR_USA)
```

```
## # A tibble: 1 x 5
##      RE    CE CDR_Japan CDR_USA CDR_diff
##   <dbl> <dbl>     <dbl>   <dbl>    <dbl>
## 1 -3.24  5.12      10.1    8.26     1.88
```

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, case fatality 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 using a generalization of this approach. 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. Except we need to hand-calculate `nAx`, being reasonably sensitive for age 0. There are standard rules for $a(0)$ for infant ages. What you see below in the `case_when()` is a simplification of the HMD protocol at http://www.mortality.org/Public/Docs/MethodsProtocol.pdf (Human Mortality Database 2018, 37) for more details. I have simplified the HMD piecewise approach by averaging male and female model results.

```
# loads lifetable from session 4!
specifc_file2  <- "master/04_lifetable_functions.R"
full_url2 <- paste(raw_repo, specifc_file2, sep = "/")
source(full_url2)


LT <-
```

```
  M %>%
  filter(Year == 2014, Sex == "total") %>%
  # need to change names to those anticipated by our function!
  rename(nMx = M) %>%
  mutate(AgeInt = 1,

         nAx = case_when(
                Age == 0 & nMx < .02012 ~ .14916 - 2.02536 * nMx,
                Age == 0 & nMx < .07599 ~ 0.037495 + 3.57055 * nMx,
                Age == 0 & nMx >= .07599 ~ 0.30663,
                Age == 110 ~ 1 / nMx,
                TRUE ~ AgeInt / 2)) %>%
  group_by(Country) %>%
  group_modify(~my_lifetable(Data = .x, radix = 1)) %>%
  ungroup()
```

The difference in life expectancy between Japan and USA 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 Japan    83.7
## 2 USA      78.9
```

Let's select just the columns we'll need, and move them side by side, like before. Except, rather than typing `Japan` and `USA` so many times.... let's just type `j` and `u`?

```
LT_arriaga <-
  LT %>%
  mutate(Country = if_else(Country == "Japan","j","u")) %>%
  select(Country, Age, lx, nLx, Tx) %>%
  pivot_wider(names_from = Country, values_from = c(lx, nLx, 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).

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

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)

$$ {}_nI_x = \frac{T^J_{x+n}}{l^U_0}(\frac{l^U_x}{l^J_x} - \frac{l^U_{x+n}}{l^J_{x+n}}) $$

One way to do it with the tidy approach:

```
LT_arriaga <-
  LT_arriaga %>%
  mutate(direct = lx_u * (nLx_j / lx_j - nLx_u / lx_u),
         indirect = lead(Tx_j) *
           (lx_u / lx_j -
              lead(lx_u) / lead(lx_j)),
         # impute 0 in the final NA
         indirect = ifelse(is.na(indirect),0,indirect))
```

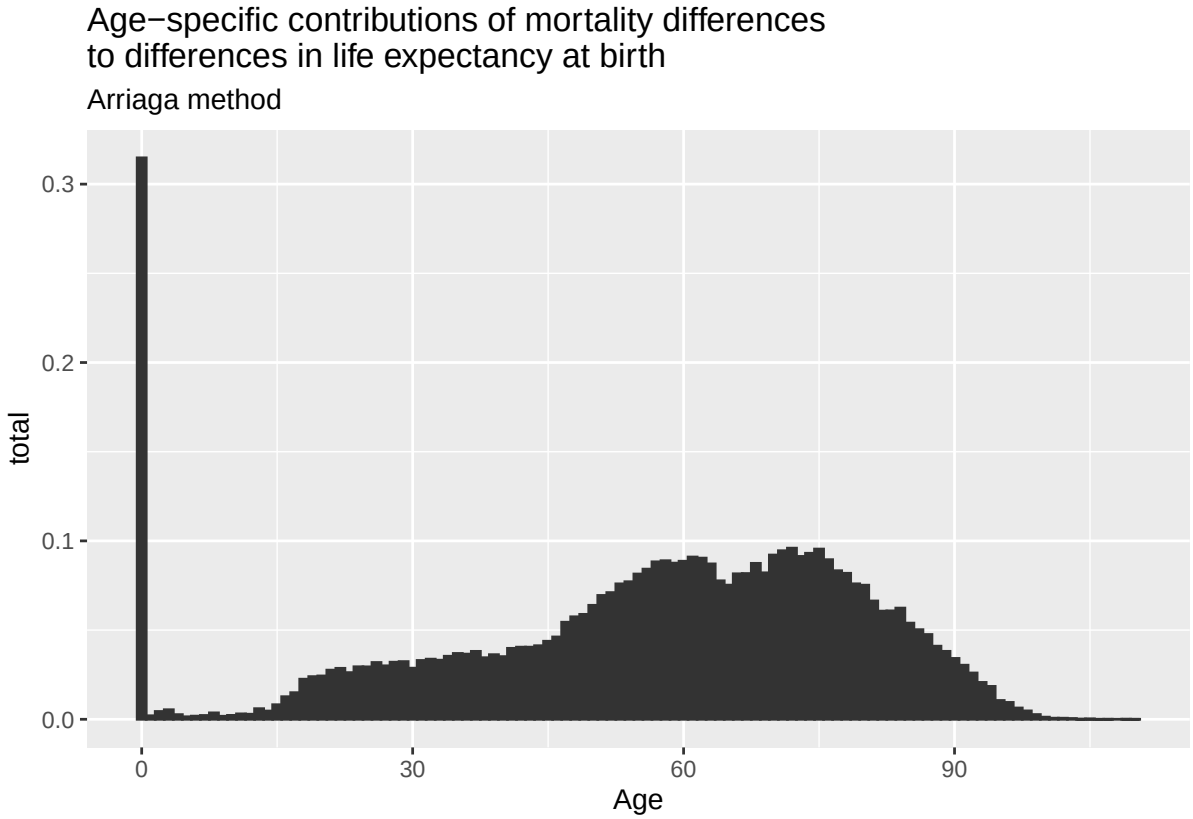
The direct and indirect contributions sum to the total differences. The Arriaga formula is then written as:

$$ {}_n\Delta_x = \frac{l^U_x}{l^U_0}\left(\frac{{}_nL^J_x}{l^J_x} - \frac{{}_nL^U_x}{l^U_x}\right) + \frac{T^J_{x+n}}{l^U_0}\left(\frac{l^U_x}{l^J_x} - \frac{l^U_{x+n}}{l^J_{x+n}}\right) $$

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_col(width=1,col=gray(.2),fill=gray(.2)) +
  labs(title = "Age-specific contributions of mortality differences\nto differences in life
       subtitle = "Arriaga method")
```




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

```
## [1] 4.780929
```

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

```
## [1] -4.780929
```

An extension of the Arriaga method decomposing life expectancy by age AND cause of death is also available (see (**preston2001demography?**)).

4.3 Generalized decomposition

A generalized decomposition method is one that will work for *any* deterministic function of parameters. I compared this family of methods in a talk in December 2021. You can find the tutorial on my github: https://github.com/timriffe/FDWG_decomp_code

In that tutorial, I demonstrate the usage of three different generalized decomposition methods, including:

- Horiuchi et al, with the `horiuchi()` function.
- Andreev et al, with the `stepwise_replacement()` function.
- Caswell, with the `ltre()` function.

There is unfortunately no space to demonstrate these methods in this lecture, but you should know that they exist and can be very convenient, especially when there is no *à propos* method

for the index you wish to decompose. If participants insist, then I will demonstrate this approach rather than the Arriaga method in the session

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.

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

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