

# Workshop on Decomposition Methods

KOSTAT, July 2023

Vladimir Cannudas-Romo and Wen Su

## Table of contents

<b>Forewords</b>	<b>4</b>
<b>Message on r preparation</b>	<b>5</b>
<b>Day 1. Standaridization</b>	<b>6</b>
Crude Death Rate . . . . .	6
Example: Korea and Japan . . . . .	7
Weighted Average . . . . .	20
Direct Standaridization . . . . .	21
Indirect Standaridization . . . . .	22
Direct Standardization Examples . . . . .	22
Mean Age of the Population . . . . .	22
General Fertility Rate . . . . .	23
Expected Number of Births . . . . .	25
Indirect Standardization Examples . . . . .	28
Crude Death Rate . . . . .	28
General Fertility Rate . . . . .	31
Expected Number of Deaths . . . . .	32
Crude Birth Rate . . . . .	35
Assignment 1 . . . . .	40
<b>Day 2. Kitagawa and Vaupel &amp; Canudas-Romo Decomposition</b>	<b>41</b>
Recap . . . . .	41
Kitagawa Decomposition . . . . .	41
Method Explained . . . . .	41
Examples of the Use of Kitagawa Method . . . . .	43

Vaupel & Canudas-Romo Decomposition . . . . .	50
Decomposition Method Explained . . . . .	50
Estimation . . . . .	51
Examples of the VCR method . . . . .	53
Assignment 2 . . . . .	66
<b>Day 3. Decomposition for Mortality Measures</b>	<b>67</b>
Recap . . . . .	67
Background on Life Expectancy ( $e_0$ ) . . . . .	67
Discrete Decomposition of Life Expectancy ( $e_0$ ) . . . . .	68
Method Explained . . . . .	68
Examples . . . . .	72
Age and Cause Contributions to Life Expectancy . . . . .	77
Vaupel & Canudas-Romo Decomposition of Life Expectancy ( $e_0$ ) . . . . .	82
Method Explained . . . . .	82
Examples . . . . .	83
Background on Cross-sectional Average Length of Life ( $CAL$ ) . . . . .	89
Decomposition of Cross-sectional Average Length of Life ( $CAL$ ) . . . . .	91
Method Explained . . . . .	91
Examples . . . . .	93
Assignment 3 . . . . .	98
<b>Day 4. Decomposition for Fertility Measures</b>	<b>99</b>
Recap . . . . .	99
Completed Cohort Fertility ( $CCF$ ) . . . . .	99
Method Explained . . . . .	99
Example . . . . .	100
Cross-sectional Average Length of Life Childless ( $CALC$ ) . . . . .	105
Method Explained . . . . .	105
Examples . . . . .	108
Net-Reproduction Rate ( $NRR$ ) . . . . .	118
Method Explained . . . . .	118
Examples . . . . .	120
Assignment 4 . . . . .	123
<b>Day 5. Decomposition for Alternative Measures</b>	<b>125</b>
Life years lost (LYL) . . . . .	125
Methods Explained . . . . .	125
Examples . . . . .	130
Variable-r . . . . .	133
Methods Explained . . . . .	133

Examples . . . . .	134
Assignment 5 . . . . .	142
<b>Assignment Answers</b>	<b>143</b>
Assignment 1 Answer . . . . .	143
Population Growth Rate ( $\bar{r}$ ) . . . . .	143
Assignment 2 Answer . . . . .	146
Crude Birth Rate (CBR) . . . . .	146
Assignment 3 Answer . . . . .	149
Part 1 . . . . .	149
Part 2 . . . . .	154
Assignments 4 Answer . . . . .	156

## Forewords

**This document is the intellectual property of Vladimir Canudas-Romo and Wen Su. Please under no circumstances share this document with anyone outside of the workshop. If you are copying and extending materials from this document, please acknowledge this document in the following way:**

Canudas-Romo, V and Su, W. (2023, July 24). Introduction to Decomposition Methods [Course handout]. KOSTAT Workshops.

This short course assumes some preliminary knowledge of calculus, Programming in R and demographic methods. We will aim to learn by replicating results and methods from previously published papers on demographic journals (*Demography*, *Population Studies*, *Demographic Research*, etc.).

We will be using data from the Human Mortality Database (HMD 2023) and the Human Fertility Database (HFD 2023). The computer labs will consist of components that you are expected to complete or understand during the course, and the components where you are expected to complete by yourself or within a group as a test of your comprehension of the course materials.

You are welcome to email me ([Vladimir.Canudas-Romo@anu.edu.au](mailto:Vladimir.Canudas-Romo@anu.edu.au)) or Wen Su ([Wen.Su@anu.edu.au](mailto:Wen.Su@anu.edu.au)) for any questions regarding computer labs. Thanks.

And let us do some demography magic!

## Message on r preparation

```
# We will be using extensively these packages
# So, please install them on your computer beforehand

# Making beautiful figures
library(ggplot2)

# Making tables
library(knitr)
# the package "knitr" does not have the option
# to add notes under the table. In this document
# the notes are added with LaTeX
# Alternatively, you can use the add_footnote()
# function from the "kableExtra" package to add the notes.

# Make the color palette pretty
library(colorRamps)
library(RColorBrewer)

# data manipulation package
# contains many small useful
# packages such as "magrittr"
# and "dplyr".
library(tidyverse)

# reshaping the data.frame and data manipulation
# e.g. from wide to long and from long to wide
library(data.table)

# Similar package to reshape data.frame
library(reshape2)

# This is to regulate the results regarding
# the rounding and the digits we will be working
# with.
options(scipen = 100000,digits=4)
```

# Day 1. Standardization

## Crude Death Rate

The Crude death rate is an informative measure of the overall mortality level of a population. Crude death rate includes deaths of all ages. It is written as the ratio of the total number of deaths over the population under exposure, usually the mid-year population, and written as

$$CDR(t) = D(t)/P(t),$$

where  $CDR(t)$  represents the crude death rate at time  $t$ , and the variable  $D(t)$  and  $P(t)$  represent the number of deaths and mid-year population at time  $t$ , respectively. We are going to explore this measure with a comparison between Korean and US data from Human Mortality Database (HMD 2023).

Since crude death rate includes deaths from all ages, we can also re-write the crude death rate as a weighted average of the age-specific death rates weighted by the population composition. This reads as:

$$CDR(t) = \frac{\sum_x m(x, t)P(x, t)}{\sum_x P(x, t)}.$$

The term  $m(x, t)$  represents the age-specific death rate at age  $x$  and time  $t$  and the term  $P(x, t)$  represents the population counts at age  $x$  and time  $t$ .

This new expression of the  $CDR(t)$  arises from the following simple derivation:

$$\begin{aligned} CDR(t) &= \frac{D(t)}{P(t)} \\ &= \frac{\sum_x D(x, t)}{P(t)} \\ &= \sum_x \frac{D(x, t)}{P(x, t)} \frac{P(x, t)}{P(t)} \\ &= \sum_x m(x, t)c(x, t). \end{aligned}$$

The notation  $c(x, t)$  represents the population structure.

Why are these two components significant? This can be seen with an example.

## Example: Korea and Japan

First the difference between the crude deaths rates between Korea and Japan during the period of 2020 is calculated.

```
JPN_Dx <-  
  read.table("data/Standard/JPN.Deaths_1x1.txt",  
            header = T, skip = 2)  
  
JPN_Dx <- JPN_Dx[JPN_Dx$Year==2020,c(1:2,5)]  
  
JPN_Ex <-  
  read.table("data/Standard/JPN.Population.txt",  
            header = T, skip = 2)  
  
JPN_Ex1 <- JPN_Ex[JPN_Ex$Year==2020,c(1:2,5)]  
JPN_Ex2 <- JPN_Ex[JPN_Ex$Year==2021,c(1:2,5)]  
  
KOR_Dx <-  
  read.table("data/Standard/KOR.Deaths_1x1.txt",  
            header = T, skip = 2)  
  
KOR_Dx <- KOR_Dx[KOR_Dx$Year==2020,c(1:2,5)]  
  
KOR_Ex <-  
  read.table("data/Standard/KOR.Population.txt",  
            header = T, skip = 2)  
  
KOR_Ex1 <- KOR_Ex[KOR_Ex$Year==2020,c(1:2,5)]  
KOR_Ex2 <- KOR_Ex[KOR_Ex$Year==2021,c(1:2,5)]  
  
JPN_CDR <- sum(JPN_Dx$Total)/  
  sum((JPN_Ex1$Total+JPN_Ex2$Total)/2)*1000  
  
KOR_CDR <- sum(KOR_Dx$Total)/  
  sum((KOR_Ex1$Total+KOR_Ex2$Total)/2)*1000  
  
CDR <- matrix(c(round(KOR_CDR,1),round(JPN_CDR,1)),  
              ncol = 2)  
colnames(CDR) <- c("Korea","Japan")
```

```
row.names(CDR) <- "CDR"
```

We can visualize our results in the form of a table:

```
kable(CDR,caption="CDR Comparison")
```

Table 1: CDR Comparison

	Korea	Japan
CDR	5.9	11.1

Note: Results are multiplied by 1000

The Japanese crude death rate is almost twice as high as the one for Korea. Can it be the case that mortality is double as high in Japan than in Korea?

The crude death rates is calculated, as we saw earlier, as the product of age-specific death rates multiplying the population composition. Let's calculate the population structure and visualize it as a population pyramid for Korea.

```
Pop <-
  read.table("data/Standard/KOR.Population.txt",
            header=TRUE,fill=TRUE, skip=2)

Pop$Age[Pop$Age=="110+"] <- "110"
Pop$Age <- as.numeric(Pop$Age)

Year1 <- 2020

Pop1 <- Pop[Pop$Year==Year1,]

Pop1$pctf <- Pop1$Female/sum(Pop1$Total)*100
Pop1$pctm <- Pop1$Male/sum(Pop1$Total)*100*-1

PopF <- Pop1[,c(1,2,6)]
PopF$Sex <- "Female"
names(PopF)[3] <- "percentage"

PopM <- Pop1[,c(1,2,7)]
```



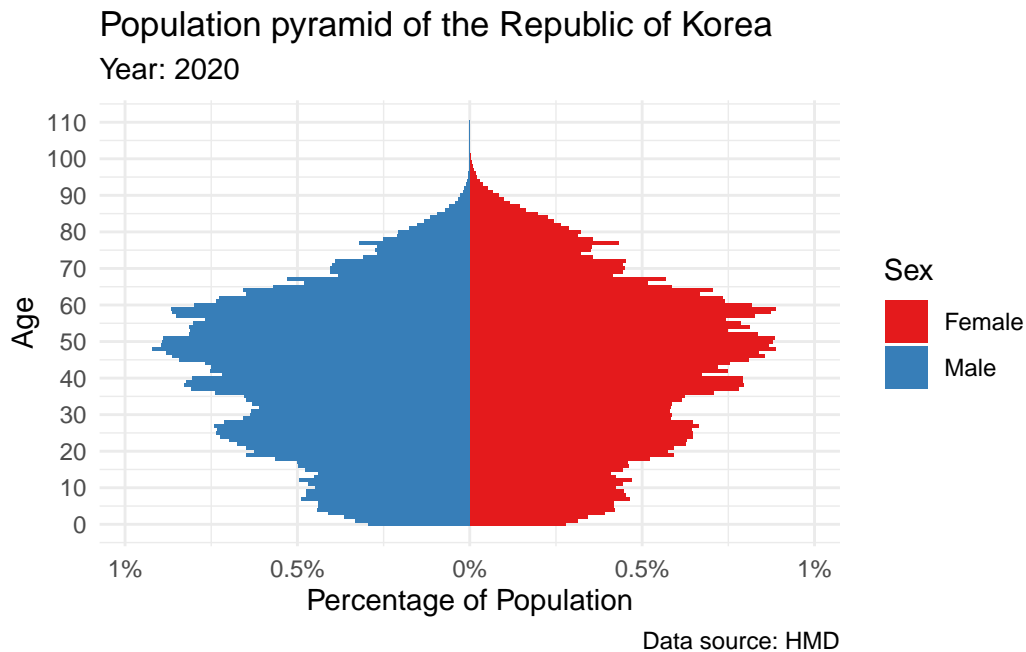
```

PopM$Sex <- "Male"
names(PopM)[3] <- "percentage"

Pop1 <- rbind(PopF,PopM)

ggplot(data = Pop1, aes(x = Age,
                        y = percentage,
                        fill = Sex)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(labels = function(x){paste0(abs(x),"%")},
                    limits = max(Pop1$percentage) * c(-1.1,1.1)) +
  scale_x_continuous(n.breaks = 10)+
  labs(x = "Age", y = "Percentage of Population",
       title = paste("Population pyramid of the","Republic of Korea"),
       subtitle = paste("Year:",Year1),
       caption = "Data source: HMD")+
  coord_flip()+ # flip x and y axis
  scale_fill_brewer(palette= "Set1")+
  theme_minimal()

```



We can also Visualize the age-specific death rates for Korea during the year 2020, for

both females and males.

```
Mx <- read.table("data/Standard/Kor.Mx_1x1.txt",
                 header = T, skip = 2)

Mx <- Mx[Mx$Year==2020,]
Male<-as.numeric(Mx$Male)
Female<-as.numeric(Mx$Female)
Mx$Female<-Mx$Male<-NULL

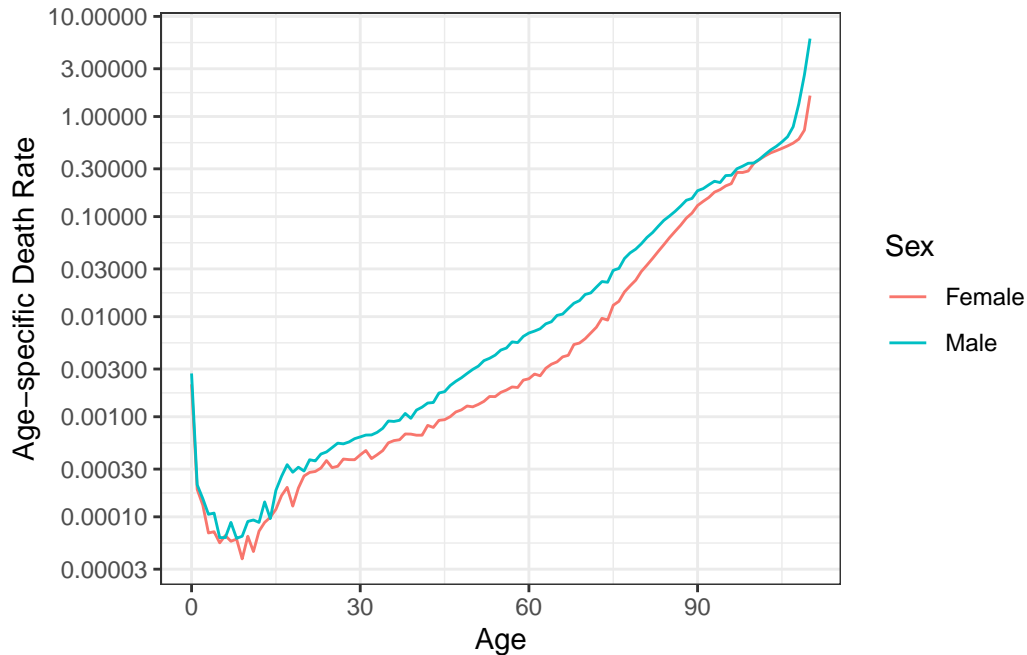
Mx<-rbind(Mx,Mx)

Mx$Total <- c(Male,Female)

Mx$Sex <- c(rep("Male",111),rep("Female",111))

Mx$Age[Mx$Age=="110+"] <- "110"
Mx$Age <- as.numeric(Mx$Age)

ggplot(Mx,aes(x=Age,y=Total,color=Sex))+
  geom_line()+
  scale_y_continuous(n.breaks = 10,trans = "log10")+
  theme_bw()+
  labs(x="Age",y="Age-specific Death Rate",
       color="Sex")
```



The crude death rate consist of two parts: the age-specific death rates and the population composition. We can compare these two components in Korea and Japan.

First we can compare the age-specific death rates between Korea and Japan. We can do this by plotting the trajectories of the mortality curve on a logged scale.

```
### Korea

Mx_KOR <- read.table("data/Standard/KOR.Mx_1x1.txt",
                    header = T, skip = 2)

Mx_KOR <- Mx_KOR[Mx_KOR$Year==2020,]

Mx_KOR$Age[Mx_KOR$Age=="110+"] <- "110"
Mx_KOR$Age <- as.numeric(Mx_KOR$Age)

Mx_KOR$Total <- as.numeric(Mx_KOR$Total)

Mx_KOR$Male <- Mx_KOR$Female <- NULL

Mx_KOR$pop <- "Korea"
```

```

### Japan

Mx_JPN <- read.table("data/Standard/JPN.Mx_1x1.txt",
                     header = T, skip = 2)

Mx_JPN <- Mx_JPN[Mx_JPN$Year==2020,]

Mx_JPN$Age[Mx_JPN$Age=="110+"] <- "110"
Mx_JPN$Age <- as.numeric(Mx_JPN$Age)

Mx_JPN$Total <- as.numeric(Mx_JPN$Total)

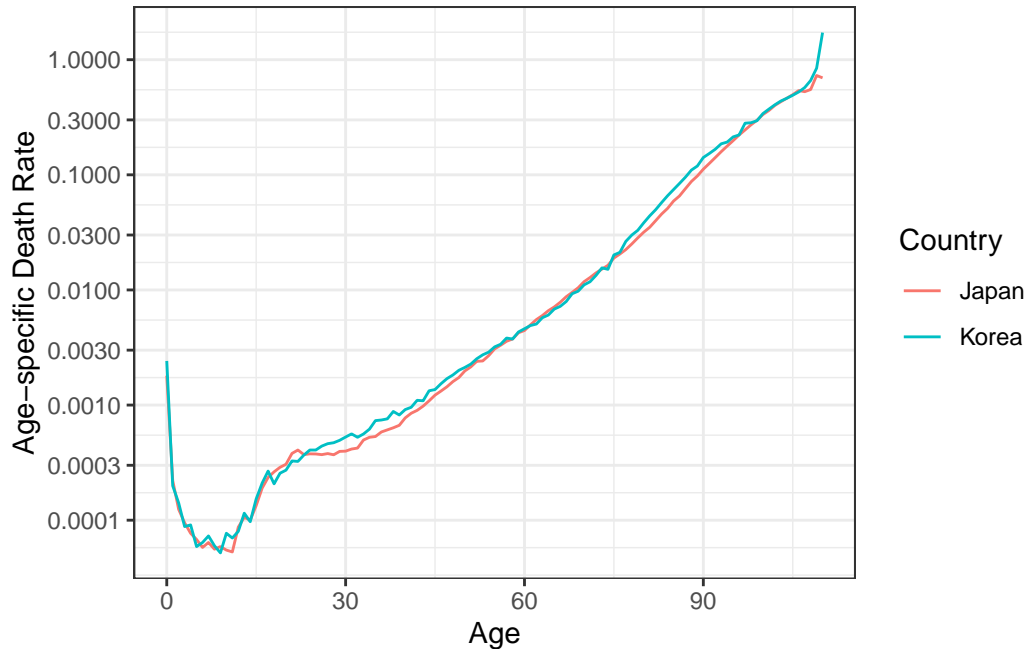
Mx_JPN$Male <- Mx_JPN$Female <- NULL

Mx_JPN$pop <- "Japan"

Mx <- rbind(Mx_KOR, Mx_JPN)

ggplot(Mx, aes(x=Age, y=Total, color=pop))+
  geom_line()+
  scale_y_continuous(n.breaks = 10, trans = "log10")+
  theme_bw()+
  labs(x="Age", y="Age-specific Death Rate",
       color="Country")

```



We can tell that the mortality curves between Korea and Japan have differences but they are not significant to the degree of Japan having twice the CDR as Korea.

Now we can compare the population structure, with the help of the population pyramids. We can compare the difference between the population structures between Japan and Korea.

```
### Korea

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE, fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

Year1 <- 2020

Pop_KOR1 <- Pop_KOR[Pop_KOR$Year==Year1,]

Pop_KOR1$pctf <- Pop_KOR1$Female/sum(Pop_KOR1$Total)*100
Pop_KOR1$pctm <- Pop_KOR1$Male/sum(Pop_KOR1$Total)*100*-1
```

```

Pop_KORF <- Pop_KOR1[,c(1,2,6)]
Pop_KORF$Sex <- "Female"
names(Pop_KORF)[3] <- "percentage"

Pop_KORM <- Pop_KOR1[,c(1,2,7)]
Pop_KORM$Sex <- "Male"
names(Pop_KORM)[3] <- "percentage"

Pop_KOR1 <- rbind(Pop_KORF,Pop_KORM)

Pop_KOR1$pop <- "Korea"

### Japan

Pop_JPN <-
  read.table("data/Standard/JPN.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_JPN$Age[Pop_JPN$Age=="110+"] <- "110"
Pop_JPN$Age <- as.numeric(Pop_JPN$Age)

Year1 <- 2020

Pop_JPN1 <- Pop_JPN[Pop_JPN$Year==Year1,]

Pop_JPN1$pctf <- Pop_JPN1$Female/sum(Pop_JPN1$Total)*100
Pop_JPN1$pctm <- Pop_JPN1$Male/sum(Pop_JPN1$Total)*100*-1

Pop_JPNF <- Pop_JPN1[,c(1,2,6)]
Pop_JPNF$Sex <- "Female"
names(Pop_JPNF)[3] <- "percentage"

Pop_JPNM <- Pop_JPN1[,c(1,2,7)]
Pop_JPNM$Sex <- "Male"
names(Pop_JPNM)[3] <- "percentage"

Pop_JPN1 <- rbind(Pop_JPNF,Pop_JPNM)

Pop_JPN1$pop <- "Japan"

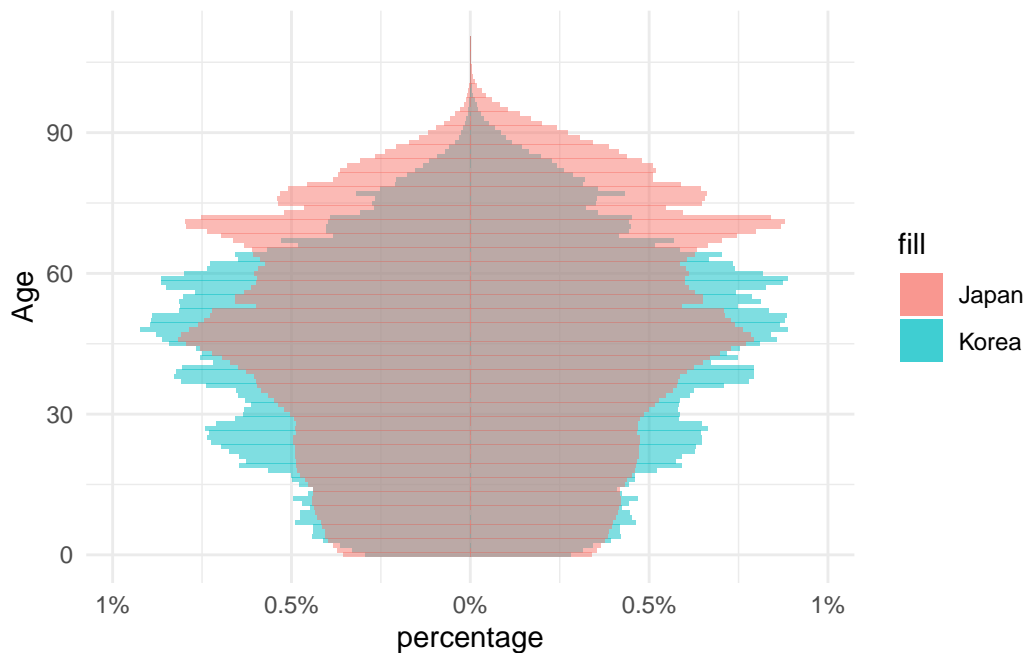
```

```

Pop <- rbind(Pop_KOR1,Pop_JPN1)

ggplot()+
  geom_col(data = Pop[Pop$pop=="Korea",],
           mapping = aes(x=Age,y=percentage,
                        fill = "Korea"),
           alpha = 0.5)+
  geom_col(data = Pop[Pop$pop=="Japan",],
           mapping = aes(x=Age,y=percentage,
                        fill = "Japan"),
           alpha = 0.5)+
  scale_y_continuous(labels = function(x){
    paste0(abs(x),"%")},
    limits = max(Pop$percentage) * c(-1.1,1.1))+
  coord_flip()+
  theme_minimal()

```



The Population structure is very different in Korea than in the more aged population of Japan. So the reason of the higher  $CDR(t)$  in Japan than in Korea is because of the age-specific death rates are much higher at older ages, where Japan has a much higher population also.

So what if we adjust these two populations with a standard population in the CDR calculation? What if we assume that the population of Japan and Korea have the same hypothetical population structure? For instance, WHO has published a standard population.

```
std <- read.table("data/Standard/std_pop.txt")

### Korea

Mx_KOR <- read.table("data/Standard/KOR.Mx_1x1.txt",
                     header = T, skip = 2)

Mx_KOR <- Mx_KOR[Mx_KOR$Year==2020,]

Mx_KOR$Age[Mx_KOR$Age=="110+"] <- "110"
Mx_KOR$Age <- as.numeric(Mx_KOR$Age)

Mx_KOR$Total <- as.numeric(Mx_KOR$Total)

Mx_KOR$Male <- Mx_KOR$Female <- NULL

Mx_KOR$pop <- "Korea"

### Japan

Mx_JPN <- read.table("data/Standard/JPN.Mx_1x1.txt",
                     header = T, skip = 2)

Mx_JPN <- Mx_JPN[Mx_JPN$Year==2020,]

Mx_JPN$Age[Mx_JPN$Age=="110+"] <- "110"
Mx_JPN$Age <- as.numeric(Mx_JPN$Age)

Mx_JPN$Total <- as.numeric(Mx_JPN$Total)

Mx_JPN$Male <- Mx_JPN$Female <- NULL

Mx_JPN$pop <- "Japan"

KOR_CDR_DS <- sum(Mx_KOR$Total*std$pop)*1000
```



```

JPN_CDR_DS <- sum(Mx_JPN$Total*std$pop)*1000

CDR_DS <-
  matrix(c(round(KOR_CDR,1),round(KOR_CDR_DS,1),
           round(JPN_CDR,1),round(JPN_CDR_DS,1)),
        ncol = 2)

colnames(CDR_DS) <- c("Korea","Japan")
row.names(CDR_DS) <- c("CDR","Standardized CDR")

kable(CDR_DS,
      caption = "CDR comparison with WHO standard")

```

Table 2: CDR comparison with WHO standard

	Korea	Japan
CDR	5.9	11.1
Standardized CDR	3.1	2.8

Note: Results are multiplied by 1000

As we can tell, when we use the WHO standard population, which is a much younger population composition than either Japan or Korea, the CDRs between the two populations are not far apart.

We could then try a better comparison using the population structure of Japan or Korea as our standard population.

```

### Korea

Mx_KOR <- read.table("data/Standard/KOR.Mx_1x1.txt",
                    header = T, skip = 2)

Mx_KOR <- Mx_KOR[Mx_KOR$Year==2020,]

Mx_KOR$Age[Mx_KOR$Age=="110+"] <- "110"
Mx_KOR$Age <- as.numeric(Mx_KOR$Age)

Mx_KOR$Total <- as.numeric(Mx_KOR$Total)

```

```

Mx_KOR$Male <- Mx_KOR$Female <- NULL

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

Year1 <- 2020
Year2 <- 2021

Pop_KOR1 <- Pop_KOR[Pop_KOR$Year==Year1,]
Pop_KOR2 <- Pop_KOR[Pop_KOR$Year==Year2,]

Pop_KOR1$Total <-
  Pop_KOR1$Total/sum(Pop_KOR1$Total)

Pop_KOR2$Total <-
  Pop_KOR2$Total/sum(Pop_KOR2$Total)

### Japan

Mx_JPN <- read.table("data/Standard/JPN.Mx_1x1.txt",
                     header = T,skip = 2)

Mx_JPN <- Mx_JPN[Mx_JPN$Year==2020,]

Mx_JPN$Age[Mx_JPN$Age=="110+"] <- "110"
Mx_JPN$Age <- as.numeric(Mx_JPN$Age)

Mx_JPN$Total <- as.numeric(Mx_JPN$Total)

Mx_JPN$Male <- Mx_JPN$Female <- NULL

Pop_JPN <-
  read.table("data/Standard/JPN.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

```

```

Pop_JPN$Age[Pop_JPN$Age=="110+"] <- "110"
Pop_JPN$Age <- as.numeric(Pop_JPN$Age)

Year1 <- 2020
Year2 <- 2021

Pop_JPN1 <- Pop_JPN[Pop_JPN$Year==Year1,]
Pop_JPN2 <- Pop_JPN[Pop_JPN$Year==Year2,]

Pop_JPN1$Total <-
  Pop_JPN1$Total/sum(Pop_JPN1$Total)

Pop_JPN2$Total <-
  Pop_JPN2$Total/sum(Pop_JPN2$Total)

###

KOR_CDR_DS_JPN <- sum(Mx_KOR$Total*
                      (Pop_JPN1$Total+
                       Pop_JPN2$Total)/2)*1000

JPN_CDR <- sum(Mx_JPN$Total*
               (Pop_JPN1$Total+
                Pop_JPN2$Total)/2)*1000

JPN_CDR_DS_KOR <- sum(Mx_JPN$Total*
                      (Pop_KOR1$Total+
                       Pop_KOR2$Total)/2)*1000

KOR_CDR <- sum(Mx_KOR$Total*
               (Pop_KOR1$Total+
                Pop_KOR2$Total)/2)*1000

CDR_table1 <-
  matrix(c(round(KOR_CDR_DS_JPN,1),
            round(JPN_CDR,1),
            round(KOR_CDR,1),
            round(JPN_CDR_DS_KOR,1)),
         ncol = 2)

```

```
colnames(CDR_table1) <- c("JPN as Standard",
                          "KOR as Standard")
row.names(CDR_table1) <- c("Korea", "Japan")

kable(CDR_table1, caption = "Comparison of Standardized CDR")
```

Table 3: Comparison of Standardized CDR

	JPN as Standard	KOR as Standard
Korea	12.7	5.9
Japan	11.1	5.3

Note: Results are multiplied by 1000

And all these options were examples of the direct standardization technique, which we will go and explain into details in the following section, relating to the concept of “weighted average”.

## Weighted Average

Many demographic measures can be written as a weighted average of a variable of interest weighting by another variable [Vaupe and Canudas-Romo \(2002\)](#).

In discrete form, the average can be written as the  $CDR(t)$  that we saw above, as:

$$\bar{v}(t) = \frac{\sum_x v(x, t)w(x, t)}{\sum_x w(x, t)},$$

where the notation  $\bar{v}(t)$  corresponds to the average of a demographic variable  $v(x, t)$  at age  $x$  and time  $t$ , and  $w(x, t)$  is the weighting variable at age  $x$  and time  $t$ .

In continuous form, the average can be written as:

$$\bar{v}(t) = \frac{\int_0^\omega v(x, t)w(x, t)dx}{\int_0^\omega w(x, t)dx}, \quad (1)$$

where  $\omega$  is the last age value of our variable of interest.

As done for the  $CDR(t)$  it is easy to see that demographic measures that can be seen as averages, as  $\bar{v}(t)$ , can then be written as the product of the two elements also:

$$\bar{v}(t) = \int_0^\omega v(x, t) c(x, t) dx, \quad (2)$$

where  $c(x, t) = \frac{w(x, t)}{\int_0^\omega w(x, t) dx}$  is the composition of the weighting variable. In many demographic examples, the latter variable is population counts  $P(x, t)$  as in the  $CDR$ .

By controlling for the composition component, we are performing a essential demographic technique called “standardization”. Standardization can be performed between two populations or the changes across time for one population. This concept is closely related to the demographic decomposition techniques which will be studied in the rest of this workshop.

## Direct Standardization

When we are trying to compare the variable of interest  $v(x, t)$  across two populations without the effect of the population structure or composition, we adjust the composition component to be a common or an average. This procedure is called “direct standardization”. In the equations, we do this by changing the component  $\frac{w(x, t)}{\int_0^\omega w(x, t) dx}$ . In the example of Korea and Japan, we compared the crude death rate by standardizing the composition component (population structure) for that of the WHO, Japan and Korea, as:

$$\bar{v}^s(t) = \frac{\int_0^\omega v(x, t) w^s(x, t) dx}{\int_0^\omega w^s(x, t) dx}, \quad (3)$$

where the notation  $w^s(x, t)$  represents the standard population composition. This standardization procedure assigns the same population structure that is selected, to both populations. Although the procedure is well established, researchers have to take care that situations of unclear population standard, as the one for Japan, Korea, or WHO in the example above, do not arise.

## Indirect Standardization

Similar standardization procedures can be done for the variable of interest, if the desire is to observe comparison attributable to the population composition across two populations. In this case, the variable  $v(x, t)$  is assigned a standard value  $v^s(x, t)$  for both populations. This difference can be interpreted as the difference in “expected outcome” given a set of standard values  $v^s(x, t)$ .

$$\bar{v}^s(t) = \frac{\int_0^\omega v^s(x, t)w(x, t)dx}{\int_0^\omega w(x, t)dx}. \quad (4)$$

In the following section some different examples of demographic variables that can be “standardized” are presented. For all the examples, the standard values the average of the two populations will be used.

## Direct Standardization Examples

### Mean Age of the Population

The mean age of the population, denoted as  $\bar{a}$ , measures the population structure with a single value. The mathematical equation using now  $a$  for age writes as:

$$\bar{a} = \frac{\int_0^\omega aP(a, t)da}{\int_0^\omega P(a, t)da}.$$

We are going to compare the difference in the mean age of the population for both Korea and Japan.

```
Pop_KOR <-  
  read.table("data/Standard/KOR.Population.txt",  
            header=TRUE, fill=TRUE, skip=2)  
  
Pop_KOR <- Pop_KOR[Pop_KOR$Year=="2020",]  
  
Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"  
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)  
Pop_KOR$Year <- as.numeric(Pop_KOR$Year)
```

```

Pop_KOR$Total <- Pop_KOR$Total/sum(Pop_KOR$Total)

Pop_JPN <-
  read.table("data/Standard/JPN.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_JPN <- Pop_JPN[Pop_JPN$Year=="2020",]

Pop_JPN$Age[Pop_JPN$Age=="110+"] <- "110"
Pop_JPN$Age <- as.numeric(Pop_JPN$Age)
Pop_JPN$Year <- as.numeric(Pop_JPN$Year)

Pop_JPN$Total <- Pop_JPN$Total/sum(Pop_JPN$Total)

MAP_JPN <- weighted.mean(seq(0,110,1),Pop_JPN$Total)

MAP_KOR <- weighted.mean(seq(0,110,1),Pop_KOR$Total)

table <- matrix(c(round(MAP_KOR,1),
                  round(MAP_JPN,1)),
               ncol=2)

row.names(table) <- "MAP"
colnames(table) <- c("Korea","Japan")

kable(table,caption = "MAP comparison")

```

Table 4: MAP comparison

	Korea	Japan
MAP	42	47.2

## General Fertility Rate

General fertility rate (GFR) is simply the number of babies divided by the number of women at reproductive ages. The mathematical equation is written as:

$$\bar{g}(t) = \frac{\int_{12}^{55} f(x, t) P_F(x, t) dx}{\int_{12}^{55} P_F(x, t) dx},$$

where the notation  $f(x, t)$  denotes the age-specific fertility rate at age  $x$  at time  $t$ , and the notation  $P_F(x, t)$  represents the population of women at age  $x$  at time  $t$ . It should be noted that the limits of integration correspond to the window of reproductive ages, 12 to 50 or 55.

```
ASFR_KOR <-
  read.table("data/Standard/KORasfrRR.txt",
             header=TRUE, fill=TRUE, skip=2)

ASFR_KOR$Age[ASFR_KOR$Age=="12-"] <- "12"
ASFR_KOR$Age[ASFR_KOR$Age=="55+"] <- "55"
ASFR_KOR$Age <- as.numeric(ASFR_KOR$Age)

ASFR_KOR1 <- ASFR_KOR[ASFR_KOR$Year==2000,]
ASFR_KOR2 <- ASFR_KOR[ASFR_KOR$Year==2020,]

Ex_KOR <-
  read.table("data/Standard/KORexposRR.txt",
             header=TRUE, fill=TRUE, skip=2)

Ex_KOR$Age[Ex_KOR$Age=="12-"] <- "12"
Ex_KOR$Age[Ex_KOR$Age=="55+"] <- "55"
Ex_KOR$Age <- as.numeric(Ex_KOR$Age)

Ex_KOR1 <- Ex_KOR[Ex_KOR$Year==2000,]
Ex_KOR1$Exposure <- Ex_KOR1$Exposure/sum(Ex_KOR1$Exposure)

Ex_KOR2 <- Ex_KOR[Ex_KOR$Year==2020,]
Ex_KOR2$Exposure <- Ex_KOR2$Exposure/sum(Ex_KOR2$Exposure)

Exposure_avg <-
  (Ex_KOR1$Exposure+Ex_KOR2$Exposure)/
  sum(Ex_KOR1$Exposure+Ex_KOR2$Exposure)

CBR1_NS <-
  weighted.mean(ASFR_KOR1$ASFR,
```



```

      Ex_KOR1$Exposure)*1000

CBR1_DS <-
  weighted.mean(ASFR_KOR1$ASFR,
    Exposure_avg)*1000

CBR2_NS <-
  weighted.mean(ASFR_KOR2$ASFR,
    Ex_KOR2$Exposure)*1000

CBR2_DS <-
  weighted.mean(ASFR_KOR2$ASFR,
    Exposure_avg)*1000

table <- matrix(c(round(CBR1_NS,1),
  round(CBR1_DS,1),
  round(CBR2_NS,1),
  round(CBR2_DS,1)),
  nrow = 2)

row.names(table) <- c("GFR",
  "Standardized GFR")

colnames(table) <- c("2000", "2020")

kable(table,caption = "GFR comparison, Korea")

```

Table 5: GFR comparison, Korea

	2000	2020
GFR	40.1	18.0
Standardized GFR	35.8	20.4

Note: Results are multiplied by 1000

## Expected Number of Births

Expected number of births (ENB) measures the number of baby expected to be born with a set of age-specific fertility rates and the female population. We write this as:

$$ENB(t) = \int_{12}^{55} f(x, t) P_F(x, t) dx.$$

The notation  $f(x, t)$  denotes the age-specific fertility rate at age  $x$  at time  $t$ . The notation  $P_F(x, t)$  represents the population of women at age  $x$  at time  $t$ . For example, it could be asked what would be the number of children for Korea if they had the population of the USA or inversely for the USA with Korea's population.

```
ASFR_KOR <-
  read.table("data/Standard/KORasfrRR.txt",
             header=TRUE, fill=TRUE, skip=2)

ASFR_KOR$Age[ASFR_KOR$Age=="12-"] <- "12"
ASFR_KOR$Age[ASFR_KOR$Age=="55+"] <- "55"
ASFR_KOR$Age <- as.numeric(ASFR_KOR$Age)

ASFR_KOR <- ASFR_KOR[ASFR_KOR$Year==2019, "ASFR"]

ASFR_USA <-
  read.table("data/Standard/USAasfrRR.txt",
             header=TRUE, fill=TRUE, skip=2)

ASFR_USA$Age[ASFR_USA$Age=="12-"] <- "12"
ASFR_USA$Age[ASFR_USA$Age=="55+"] <- "55"
ASFR_USA$Age <- as.numeric(ASFR_USA$Age)

ASFR_USA <- ASFR_USA[ASFR_USA$Year==2019, "ASFR"]

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE, fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

Exp_KOR <-
  (Pop_KOR[Pop_KOR$Year==2019, "Female"] +
   Pop_KOR[Pop_KOR$Year==2020, "Female"])/2
```

```

Pop_USA <-
  read.table("data/Standard/USA.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_USA$Age[Pop_USA$Age=="110+"] <- "110"
Pop_USA$Age <- as.numeric(Pop_USA$Age)

Exp_USA <-
  (Pop_USA[Pop_USA$Year==2019,"Female"]+
   Pop_USA[Pop_USA$Year==2020,"Female"])/2

ENB_KOR_NS <-
  sum(c(rep(0,12),ASFR_KOR,rep(0,55))*
      Exp_KOR)/100000

ENB_KOR_DS <-
  sum(c(rep(0,12),ASFR_KOR,rep(0,55))*
      Exp_USA)/100000

ENB_USA_NS <-
  sum(c(rep(0,12),ASFR_USA,rep(0,55))*
      Exp_USA)/100000

ENB_USA_DS <-
  sum(c(rep(0,12),ASFR_USA,rep(0,55))*
      Exp_KOR)/100000

table <- matrix(c(round(ENB_KOR_NS,1),
                  round(ENB_USA_DS,1),
                  round(ENB_KOR_DS,1),
                  round(ENB_USA_NS,1)
                  ),
                ncol = 2)

row.names(table) <- c("Korea",
                     "USA")

colnames(table) <- c("KOR standard", "USA standard")

```

```
kable(table,caption = "ENB Comparison, Korea & USA, * 100,000")
```

Table 6: ENB Comparison, Korea & USA, \* 100,000

	KOR standard	USA standard
Korea	3.0	20.4
USA	5.6	37.5

Note: Results are divided by 100000

## Indirect Standardization Examples

### Crude Death Rate

The crude death rate can also be compared with indirect standardization. In this case the average of the age-specific death rates between Korea and Japan is used as the standard and different population structure will lead to differences in crude death rates.

```
### Korea

Mx_KOR <- read.table("data/Standard/KOR.Mx_1x1.txt",
                     header = T, skip = 2)

Mx_KOR <- Mx_KOR[Mx_KOR$Year==2020,]

Mx_KOR$Age[Mx_KOR$Age=="110+"] <- "110"
Mx_KOR$Age <- as.numeric(Mx_KOR$Age)

Mx_KOR$Total <- as.numeric(Mx_KOR$Total)

Mx_KOR$Male <- Mx_KOR$Female <- NULL

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE, fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)
```

```

Year1 <- 2020
Year2 <- 2021

Pop_KOR1 <- Pop_KOR[Pop_KOR$Year==Year1,]
Pop_KOR2 <- Pop_KOR[Pop_KOR$Year==Year2,]

Pop_KOR1$Total <-
  Pop_KOR1$Total/sum(Pop_KOR1$Total)

Pop_KOR2$Total <-
  Pop_KOR2$Total/sum(Pop_KOR2$Total)

Exposure1 <- (Pop_KOR1$Total+Pop_KOR2$Total)/2

### Japan

Mx_JPN <- read.table("data/Standard/JPN.Mx_1x1.txt",
                     header = T, skip = 2)

Mx_JPN <- Mx_JPN[Mx_JPN$Year==2020,]

Mx_JPN$Age[Mx_JPN$Age=="110+"] <- "110"
Mx_JPN$Age <- as.numeric(Mx_JPN$Age)

Mx_JPN$Total <- as.numeric(Mx_JPN$Total)

Mx_JPN$Male <- Mx_JPN$Female <- NULL

Pop_JPN <-
  read.table("data/Standard/JPN.Population.txt",
             header=TRUE, fill=TRUE, skip=2)

Pop_JPN$Age[Pop_JPN$Age=="110+"] <- "110"
Pop_JPN$Age <- as.numeric(Pop_JPN$Age)

Year1 <- 2020
Year2 <- 2021

Pop_JPN1 <- Pop_JPN[Pop_JPN$Year==Year1,]

```

```

Pop_JPN2 <- Pop_JPN[Pop_JPN$Year==Year2,]

Pop_JPN1$Total <-
  Pop_JPN1$Total/sum(Pop_JPN1$Total)

Pop_JPN2$Total <-
  Pop_JPN2$Total/sum(Pop_JPN2$Total)

Exposure2 <- (Pop_JPN1$Total+Pop_JPN2$Total)/2

###

KOR_CDR_IS <- sum((Mx_KOR$Total+Mx_JPN$Total)/2*Exposure1)*1000

KOR_CDR <- sum(Mx_KOR$Total*Exposure1)*1000

JPN_CDR_IS <- sum((Mx_KOR$Total+Mx_JPN$Total)/2*Exposure2)*1000

JPN_CDR <- sum(Mx_JPN$Total*Exposure2)*1000

table <- matrix(c(round(KOR_CDR,1),
                    round(KOR_CDR_IS,1),
                    round(JPN_CDR,1),
                    round(JPN_CDR_IS,1)),
                 ncol = 2)

row.names(table) <- c("CDR",
                     "Indirect Standardized CDR")

colnames(table) <- c("Korea", "Japan")

kable(table, caption = "CDR comparison, Korea vs Japan")

```

Table 7: CDR comparison, Korea vs Japan

	Korea	Japan
CDR	5.9	11.1
Indirect Standardized CDR	5.6	11.9

Note: Results are multiplied by 1000

As expected the difference between CDRs original and standardized is not different from each other, since both countries had similar levels of mortality.

## General Fertility Rate

Similarly we would ask about the indirect standardization on the General Fertility Rate. The example below shows this result for Korea comparing 2000 vs 2020.

```
ASFR_KOR <-  
  read.table("data/Standard/KORasfrRR.txt",  
             header=TRUE, fill=TRUE, skip=2)  
  
ASFR_KOR$Age[ASFR_KOR$Age=="12-"] <- "12"  
ASFR_KOR$Age[ASFR_KOR$Age=="55+"] <- "55"  
ASFR_KOR$Age <- as.numeric(ASFR_KOR$Age)  
  
ASFR_KOR1 <- ASFR_KOR[ASFR_KOR$Year==2000,]  
ASFR_KOR2 <- ASFR_KOR[ASFR_KOR$Year==2020,]  
  
Ex_KOR <-  
  read.table("data/Standard/KORexposRR.txt",  
             header=TRUE, fill=TRUE, skip=2)  
  
Ex_KOR$Age[Ex_KOR$Age=="12-"] <- "12"  
Ex_KOR$Age[Ex_KOR$Age=="55+"] <- "55"  
Ex_KOR$Age <- as.numeric(Ex_KOR$Age)  
  
Ex_KOR1 <- Ex_KOR[Ex_KOR$Year==2000,]  
Ex_KOR1$Exposure <- Ex_KOR1$Exposure/sum(Ex_KOR1$Exposure)  
  
Ex_KOR2 <- Ex_KOR[Ex_KOR$Year==2020,]  
Ex_KOR2$Exposure <- Ex_KOR2$Exposure/sum(Ex_KOR2$Exposure)  
  
CBR1_NS <-  
  weighted.mean(ASFR_KOR1$ASFR,  
                Ex_KOR1$Exposure)*1000
```

```

CBR1_IS <-
  weighted.mean((ASFR_KOR1$ASFR+ASFR_KOR2$ASFR)/2,
                Ex_KOR1$Exposure)*1000

CBR2_NS <-
  weighted.mean(ASFR_KOR2$ASFR,
                Ex_KOR2$Exposure)*1000

CBR2_IS <-
  weighted.mean((ASFR_KOR1$ASFR+ASFR_KOR2$ASFR)/2,
                Ex_KOR2$Exposure)*1000

table <- matrix(c(round(CBR1_NS,1),
                    round(CBR1_IS,1),
                    round(CBR2_NS,1),
                    round(CBR2_IS,1)),
                nrow = 2)

row.names(table) <- c("GFR",
                     "Indirect Standardized GFR")

colnames(table) <- c("2000", "2020")

kable(table, caption = "GFR comparison, Korea 2000 and 2020")

```

Table 8: GFR comparison, Korea 2000 and 2020

	2000	2020
GFR	40.1	18.0
Indirect Standardized GFR	31.4	24.8

Note: Results are multiplied by 1000

## Expected Number of Deaths

Expected number of deaths (END) measures the number of people that are expected to die with a set of age-specific death rates in a given population. Its expression is:



$$END(t) = \int_0^{\omega} m(x, t)P(x, t)dx.$$

The notation  $m(x, t)$  denotes the age-specific death rate at age  $x$  at time  $t$ . The notation  $P(x, t)$  represents the population at age  $x$  at time  $t$ . It is easy to see the use of this measure with an indirect standardization procedure, for example to ask what would the number of deaths be if the 2020 death rates of the USA had applied to Korea.

```
Mx_KOR <-
  read.table("data/Standard/KOR.Mx_1x1.txt",
             header=TRUE, fill=TRUE, skip=2)

Mx_KOR$Age[Mx_KOR$Age=="110+"] <- "110"
Mx_KOR$Age <- as.numeric(Mx_KOR$Age)

Mx_KOR <- Mx_KOR[Mx_KOR$Year==2020, "Total"]

Mx_USA <-
  read.table("data/Standard/USA.Mx_1x1.txt",
             header=TRUE, fill=TRUE, skip=2)

Mx_USA$Age[Mx_USA$Age=="110+"] <- "110"
Mx_USA$Age <- as.numeric(Mx_USA$Age)

Mx_USA <- Mx_USA[Mx_USA$Year==2020, "Total"]

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE, fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

Exp_KOR <-
  (Pop_KOR[Pop_KOR$Year==2020, "Total"] +
   Pop_KOR[Pop_KOR$Year==2021, "Total"])/2

Pop_USA <-
  read.table("data/Standard/USA.Population.txt",
```

```

      header=TRUE,fill=TRUE, skip=2)

Pop_USA$Age[Pop_USA$Age=="110+"] <- "110"
Pop_USA$Age <- as.numeric(Pop_USA$Age)

Exp_USA <-
  (Pop_USA[Pop_USA$Year==2020,"Total"]+
   Pop_USA[Pop_USA$Year==2021,"Total"])/2

ENB_KOR_NS <-
  sum(Mx_KOR*Exp_KOR)/100000

ENB_KOR_IS <-
  sum(Mx_USA*Exp_KOR)/100000

ENB_USA_NS <-
  sum(Mx_USA*Exp_USA)/100000

ENB_USA_IS <-
  sum(Mx_KOR*Exp_USA)/100000

table <- matrix(c(round(ENB_KOR_NS,0),
                    round(ENB_USA_IS,0),
                    round(ENB_KOR_IS,0),
                    round(ENB_USA_NS,0)
                  ),
                ncol = 2)

row.names(table) <- c("Korea",
                     "USA")

colnames(table) <- c("KOR standard", "USA standard")

kable(table,caption = "END Comparison with indirect standard, Korea & USA, * 100,000")

```

Table 9: END Comparison with indirect standard, Korea & USA, \* 100,000

	KOR standard	USA standard
Korea	3	5
USA	21	34

Note: Results are divided by 100,000

## Crude Birth Rate

The crude birth rate has similar interpretation as the crude death rate, in this case the number of babies per 1000 individuals. The mathematical notation of CBR can be written as:

$$\begin{aligned}
 CBR(t) &= \frac{B(t)}{P(t)} \\
 &= \frac{\sum_x B(x, t)}{P(t)} \\
 &= \sum_x \frac{B(x, t)}{P_F(x, t)} \frac{P_F(x, t)}{P(t)} \\
 &= \sum_x f(x, t) c_F(x, t).
 \end{aligned}$$

As before, the notation  $P(x, t)$  and  $P_F(x, t)$  denote the total population and the population of females at age  $x$  at time  $t$ , respectively, and  $P(t)$  is the overall population at time  $t$ .  $B(x, t)$  stands for the number of births from women age  $x$  and time  $t$ . The notation  $f(x, t)$  denotes the age-specific fertility rate at age  $x$  within the fertility window (usually ages 12 to 55) and time  $t$ . It should be noted that here the weighting function is the number of women divided by the total population  $c_F(x, t) = \frac{P_F(x, t)}{P(t)}$ . To compare the changes across time in crude birth rate both populations, female and total, have to be used in the standardization.

```

ASFR_KOR <-
  read.table("data/Standard/KORasfrRR.txt",
             header=TRUE, fill=TRUE, skip=2)

ASFR_KOR$Age[ASFR_KOR$Age=="12-"] <- "12"
ASFR_KOR$Age[ASFR_KOR$Age=="55+"] <- "55"

```

```

ASFR_KOR$Age <- as.numeric(ASFR_KOR$Age)

ASFR_KOR1 <- ASFR_KOR[ASFR_KOR$Year==2003,]
ASFR_KOR2 <- ASFR_KOR[ASFR_KOR$Year==2019,]

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

totpop_KOR1 <-
  (Pop_KOR[Pop_KOR$Year==2003,"Total"]+
   Pop_KOR[Pop_KOR$Year==2004,"Total"])/2

totpop_KOR2 <-
  (Pop_KOR[Pop_KOR$Year==2019,"Total"]+
   Pop_KOR[Pop_KOR$Year==2020,"Total"])/2

fpop_KOR1 <-
  (Pop_KOR[Pop_KOR$Year==2003,"Female"]+
   Pop_KOR[Pop_KOR$Year==2004,"Female"])/2

fpop_KOR2 <-
  (Pop_KOR[Pop_KOR$Year==2019,"Female"]+
   Pop_KOR[Pop_KOR$Year==2020,"Female"])/2

Exposure1 <- fpop_KOR1/sum(totpop_KOR1)

Exposure2 <- fpop_KOR2/sum(totpop_KOR2)

Exposure_avg <-
  (fpop_KOR1+fpop_KOR2)*0.5/
  sum((totpop_KOR1+totpop_KOR2)*0.5)

CBR1_NS <-
  weighted.mean(c(rep(0,12),ASFR_KOR1$ASFR,rep(0,55)),
                Exposure1)*1000

```

```

CBR1_DS <-
  weighted.mean(c(rep(0,12),ASFR_KOR1$ASFR,rep(0,55)),
                Exposure_avg)*1000

CBR2_NS <-
  weighted.mean(c(rep(0,12),ASFR_KOR2$ASFR,rep(0,55)),
                Exposure2)*1000

CBR2_DS <-
  weighted.mean(c(rep(0,12),ASFR_KOR2$ASFR,rep(0,55)),
                Exposure_avg)*1000

table <- matrix(c(round(CBR1_NS,1),
                    round(CBR1_DS,1),
                    round(CBR2_NS,1),
                    round(CBR2_DS,1)),
                ncol = 2)

row.names(table) <- c("CBR",
                     "Standard CBR")

colnames(table) <- c("2003", "2019")

kable(table,caption = "CBR Comparison, Korea")

```

Table 10: CBR Comparison, Korea

	2003	2019
CBR	20.6	11.8
Standard CBR	17.6	14.0

Note: Results are multiplied by 1000

It is also possible to apply the indirect standardization technique to the Crude Birth Rate. The following example does that.

```

ASFR_KOR <-
  read.table("data/Standard/KORasfrRR.txt",
             header=TRUE,fill=TRUE, skip=2)

ASFR_KOR$Age[ASFR_KOR$Age=="12-"] <- "12"
ASFR_KOR$Age[ASFR_KOR$Age=="55+"] <- "55"
ASFR_KOR$Age <- as.numeric(ASFR_KOR$Age)

ASFR_KOR1 <- ASFR_KOR[ASFR_KOR$Year==2003,]
ASFR_KOR2 <- ASFR_KOR[ASFR_KOR$Year==2019,]

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

totpop_KOR1 <-
  (Pop_KOR[Pop_KOR$Year==2003,"Total"]+
   Pop_KOR[Pop_KOR$Year==2004,"Total"])/2

totpop_KOR2 <-
  (Pop_KOR[Pop_KOR$Year==2019,"Total"]+
   Pop_KOR[Pop_KOR$Year==2020,"Total"])/2

fpop_KOR1 <-
  (Pop_KOR[Pop_KOR$Year==2003,"Female"]+
   Pop_KOR[Pop_KOR$Year==2004,"Female"])/2

fpop_KOR2 <-
  (Pop_KOR[Pop_KOR$Year==2019,"Female"]+
   Pop_KOR[Pop_KOR$Year==2020,"Female"])/2

Exposure1 <- fpop_KOR1/sum(totpop_KOR1)

Exposure2 <- fpop_KOR2/sum(totpop_KOR2)

CBR1_NS <-

```

```

weighted.mean(c(rep(0,12),ASFR_KOR1$ASFR,rep(0,55)),
              Exposure1)*1000

CBR1_IS <-
  weighted.mean((c(rep(0,12),ASFR_KOR1$ASFR,rep(0,55))+
                  c(rep(0,12),ASFR_KOR2$ASFR,rep(0,55)))/2,
                Exposure1)*1000

CBR2_NS <-
  weighted.mean(c(rep(0,12),ASFR_KOR2$ASFR,rep(0,55)),
                Exposure2)*1000

CBR2_IS <-
  weighted.mean((c(rep(0,12),ASFR_KOR1$ASFR,rep(0,55))+
                  c(rep(0,12),ASFR_KOR2$ASFR,rep(0,55)))/2,
                Exposure2)*1000

table <- matrix(c(round(CBR1_NS,1),
                    round(CBR1_IS,1),
                    round(CBR2_NS,1),
                    round(CBR2_IS,1)),
                ncol = 2)

row.names(table) <- c("No Standardization",
                     "Indirect Standardization")

colnames(table) <- c("Year 2003", "Year 2019")

kable(table,caption = "CBR Comparison, Korea")

```

Table 11: CBR Comparison, Korea

	Year 2003	Year 2019
No Standardization	20.6	11.8
Indirect Standardization	18.5	13.3

Note: Results are multiplied by 1000

## Assignment 1

Select one of the suggestions below and submit in ONE page: one Figure (or Table) and a brief paragraph describing the results that you find.

- Select one of the measures in the examples of this section and try to apply a comparison between Korea and a different population in HMD or HFD that we have not used in class.
- For example, you could compare the crude death rates for Korea and US adjusting for the population structure of each country.



## Day 2. Kitagawa and Vaupel & Canudas-Romo Decomposition

### Recap

Please refresh your memory with the materials from the previous section, and answer the following questions:

- What is direct standardization?
- What is indirect standardization.
- What are the differences between the two?

### Kitagawa Decomposition

#### Method Explained

Kitagawa (1955) proposed a method to disentangle differences in demographic averages into the contributions of its components: the variable of interest and the weighting variable. The underlying assumption is that the difference can be separated arithmetically, or that there is a linear relationship between the two compared measures. This assumption enable us to decompose the difference in crude death rates into two components:

$$CDR(JPN) - CDR(KOR) = \sum_x \Delta m_x \bar{c}_x + \sum_x \Delta c_x \bar{m}_x, \quad (5)$$

the first term on the right of the equation refers to the contribution of the differences in death rates between the Japanese and Korean females, calculated as

$$\Delta m_x = m_x(JPN) - m_x(KOR),$$

and it is multiplied by the population composition, which in this case is the average of the two populations as

$$\bar{c}_x = \frac{c_x(JPN) + c_x(KOR)}{2}.$$

This first term will be the Direct component of the difference.

Similarly the second term on the right corresponds to the difference in population compositions between Japan and Korea or

$$\Delta c_x = c_x(JPN) - c_x(KOR),$$

multiplying the average death rate between the two populations calculated as

$$\overline{m}_x = \frac{m_x(JPN) + m_x(KOR)}{2}.$$

The first component on the right side of the equation is called the direct component, which uses direct standardization (standard population composition), and the second component is called the indirect component, which uses indirect standardization (standard rate). The Kitagawa method combines the direct and indirect standardization in one expression to observe the difference in demographic averages and it can be applied to all demographic measures that can be expressed as a weighted average.

This method can also be applied in a regression setting (Fortin, Lemieux, and Firpo 2011). Can you think of how that would be done?

This decomposition involves two variables, the rate of interest and the population composition by age. However, it can be further extended to include more terms, for example if information was available for the rate and population by age, and also by sex, education and many other characteristics. With arithmetic additions and subtractions Gupta (1993) shows how to do this procedure when information of the rate of interest and the population is available for all those elements. Nevertheless, the procedure can easily become cumbersome since the user will need to decide some hierarchy of those characteristics of the population, e.g. is it age which is more important than sex in the population composition, or education, etc. and the researcher will also have to decide what to do with interaction terms. The complexity is further increased with the number of variables included.

Changing from this discrete approach (arithmetic: differences and additions), as the Kitagawa method, to a continuous approach (calculus: derivatives and integrals) avoids having to deal with hierarchy nor interaction terms. In the rest of the workshop we study both type of methods, those that look at demographic measures as if they changed on a continuous basis, and those concerned with differences. The first method studied is the [Vaupel and Canudas-Romo \(2002\)](#) method, which is an extension of the specific perspective applied for studying changes in the mean age of the population (S. H. Preston (1976); Kim and Strobino (1984)).

## Examples of the Use of Kitagawa Method

As examples of the use of the Kitagawa method, the CDR and other measure illustrations will be repeated, and the researcher will be asked to assess any advantage or disadvantage of the technique used.

### Crude Death Rate (CDR)

```
### Korea

Mx_KOR <- read.table("data/Standard/KOR.Mx_1x1.txt",
                    header = T, skip = 2)

Mx_KOR <- Mx_KOR[Mx_KOR$Year==2020,]

Mx_KOR$Age[Mx_KOR$Age=="110+"] <- "110"
Mx_KOR$Age <- as.numeric(Mx_KOR$Age)

Mx_KOR$Total <- as.numeric(Mx_KOR$Total)

Mx_KOR$Male <- Mx_KOR$Female <- NULL

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
            header=TRUE, fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

Year1 <- 2020
Year2 <- 2021

Pop_KOR1 <- Pop_KOR[Pop_KOR$Year==Year1,]
Pop_KOR2 <- Pop_KOR[Pop_KOR$Year==Year2,]

Pop_KOR1$Total <-
  Pop_KOR1$Total/sum(Pop_KOR1$Total)

Pop_KOR2$Total <-
```

```

    Pop_KOR2$Total/sum(Pop_KOR2$Total)

Exposure_KOR <- (Pop_KOR1$Total+Pop_KOR2$Total)/2

### Japan

Mx_JPN <- read.table("data/Standard/JPN.Mx_1x1.txt",
                    header = T, skip = 2)

Mx_JPN <- Mx_JPN[Mx_JPN$Year==2020,]

Mx_JPN$Age[Mx_JPN$Age=="110+"] <- "110"
Mx_JPN$Age <- as.numeric(Mx_JPN$Age)

Mx_JPN$Total <- as.numeric(Mx_JPN$Total)

Mx_JPN$Male <- Mx_JPN$Female <- NULL

Pop_JPN <-
  read.table("data/Standard/JPN.Population.txt",
            header=TRUE, fill=TRUE, skip=2)

Pop_JPN$Age[Pop_JPN$Age=="110+"] <- "110"
Pop_JPN$Age <- as.numeric(Pop_JPN$Age)

Year1 <- 2020
Year2 <- 2021

Pop_JPN1 <- Pop_JPN[Pop_JPN$Year==Year1,]
Pop_JPN2 <- Pop_JPN[Pop_JPN$Year==Year2,]

Pop_JPN1$Total <-
  Pop_JPN1$Total/sum(Pop_JPN1$Total)

Pop_JPN2$Total <-
  Pop_JPN2$Total/sum(Pop_JPN2$Total)

Exposure_JPN <- (Pop_JPN1$Total+Pop_JPN2$Total)/2

```

```

CDR_JPN <- sum(Mx_JPN$Total*Exposure_JPN)*1000

CDR_KOR <- sum(Mx_KOR$Total*Exposure_KOR)*1000

CDR_diff <- CDR_JPN-CDR_KOR

CDR_direct <-
  sum((Mx_JPN$Total-Mx_KOR$Total)*
      (Exposure_JPN+Exposure_KOR)/2)*1000

CDR_comp <-
  sum((Exposure_JPN-Exposure_KOR)*
      (Mx_JPN$Total+Mx_KOR$Total)/2)*1000

CDR_decomp <- CDR_direct+CDR_comp

table <-
matrix(round(c(CDR_JPN,CDR_KOR,
              CDR_diff,CDR_direct,
              CDR_comp,CDR_decomp),2),
       ncol=1)

colnames(table) <- ""

row.names(table) <- c(paste0("CDR in Japan"),
                     paste0("CDR in Korea"),
                     "Difference",
                     "Direct component",
                     "Indirect component",
                     "Total estimated difference")

kable(table,caption = "CDR decomposition (Kitagawa)")

```

Table 12: CDR decomposition (Kitagawa)

CDR in Japan	11.08
CDR in Korea	5.94
Difference	5.14
Direct component	-1.11
Indirect component	6.26

Table 12: CDR decomposition (Kitagawa)

Total estimated difference	5.14
----------------------------	------

Note: Results are multiplied by 1000

### Population Growth ( $\bar{r}$ )

Population growth rate at the population level  $\bar{r}(t)$  can be written as the weighted average of age-specific growth rates  $r(x, t)$  and the population structure  $c(x, t)$  Horiuchi (1991), as

$$\bar{r}(t) = \int_0^{\omega} r(x, t) c(x, t) dx.$$

The difference in population growth rates between females and males can be written as:

$$\Delta\bar{r}(t) = \bar{r}_M(t) - \bar{r}_F(t),$$

Which can be further expanded with the Kitagawa method into

$$\begin{aligned} \Delta\bar{r}(t) &= \bar{r}_M(t) - \bar{r}_F(t) \\ &= \int_0^{\omega} [r_M(x, t) - r_F(x, t)] \frac{c_M(x, t) + c_F(x, t)}{2} dx + \\ &\quad \int_0^{\omega} [c_M(x, t) - c_F(x, t)] \frac{r_M(x, t) + r_F(x, t)}{2} dx. \end{aligned}$$

The example below applies this difference in growth rates to compare the female and male Japanese populations.

```
Year1 = 2010

Year2 = 2020

table_t1 <- read.table("data/Standard/JPN.Population.txt",
                      header = T, skip=2)

table_t1 <- table_t1[table_t1$Year==Year1,
```

```

      c("Age", "Female", "Male")]

table_t1$Cx_f <- table_t1$Female/sum(table_t1$Female)

table_t1$Cx_m <- table_t1$Male/sum(table_t1$Male)

table_t2 <- read.table("data/Standard/JPN.Population.txt",
                      header = T, skip=2)

table_t2 <- table_t2[table_t2$Year==Year2,
                    c("Age", "Female", "Male")]

table_t2$Cx_f <- table_t2$Female/sum(table_t2$Female)

table_t2$Cx_m <- table_t2$Male/sum(table_t2$Male)

r_Pm <- log(table_t2$Male/table_t1$Male)/10

r_Pm <- ifelse(is.infinite(r_Pm), 0, r_Pm)

r_Pf <- log(table_t2$Female/table_t1$Female)/10

r_Pf <- ifelse(is.infinite(r_Pf), 0, r_Pf)

Cx_m <- (table_t1$Cx_m+table_t2$Cx_m)/2

Cx_f <- (table_t1$Cx_f+table_t2$Cx_f)/2

rm <- sum(ifelse(is.infinite(r_Pm), 0, r_Pm)*Cx_m, na.rm = T)
rf <- sum(ifelse(is.infinite(r_Pf), 0, r_Pf)*Cx_f, na.rm = T)

r_diff <- rm-rf

r_direct <- sum((r_Pm-r_Pf)*(Cx_m+Cx_f)/2)

r_comp <- sum((Cx_m-Cx_f)*(r_Pm+r_Pf)/2)

r_decomp <-
  sum((r_Pm-r_Pf)*(Cx_m+Cx_f)/2+

```

```

      (Cx_m-Cx_f)*(r_Pm+r_Pf)/2,
      na.rm = T)

table <-
  matrix(round(c(rm*100,rf*100,r_diff*100,
    r_direct*100,r_comp*100,r_decomp*100),2),
    ncol=1)

row.names(table) <- c(paste0("Growth of Male Population ",
    Year1,"-",Year2),
    paste0("Growth of Female Population ",
    Year1,"-",Year2),
    "Observed difference",
    "Direct Component",
    "Indirect Component",
    "Estimated Difference")

colnames(table) <- "Japan"

kable(table, caption = "Growth Difference between Male and Female, Japan")

```

Table 13: Growth Difference between Male and Female, Japan

	Japan
Growth of Male Population 2010-2020	-0.22
Growth of Female Population 2010-2020	-0.18
Observed difference	-0.04
Direct Component	0.24
Indirect Component	-0.28
Estimated Difference	-0.04

Note: Results are multiplied by 1000

Methods of decomposition can also be presented in interesting plots, highlighting how different parts contribute to the overall difference. For example, we can calculate the age-specific contribution of both the direct and indirect components. Actually, instead of using the direct and indirect terms it is better to directly call them the growth rate and the population composition effects of the difference.

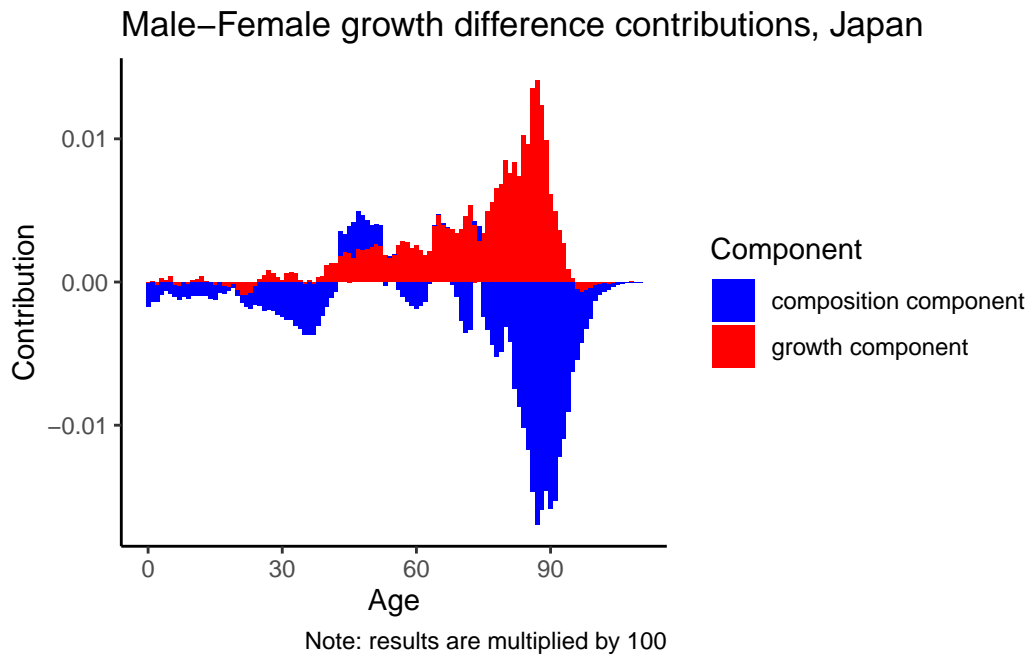


```

table <-
  data.frame(
    age = rep(0:110,2),
    variable =
      rep(c("growth component",
            "composition component"),
          each = 111),
    value =
      c((r_Pm-r_Pf)*(Cx_m+Cx_f)/2*100,
        (Cx_m-Cx_f)*(r_Pm+r_Pf)/2*100)
  )

ggplot(data = table,aes(x=age,y=value,fill=variable))+
  geom_col()+
  scale_fill_manual(values = c("blue","red"))+
  labs(title = "Male-Female growth difference contributions, Japan",
       x="Age",y="Contribution",fill="Component",
       caption = "Note: results are multiplied by 100")+
  theme_classic()

```



## Vaupel & Canudas-Romo Decomposition

### Decomposition Method Explained

As in the first day, let a demographic measure that can be expressed as an average be denoted by a bar on top of the variable of interest  $\bar{v}(t)$  and mathematically written as:

$$\bar{v}(t) = \frac{\int_0^\omega v(x, t)w(x, t)dx}{\int_0^\omega w(x, t)dx}. \quad (6)$$

In this case the average  $\bar{v}(t)$  is an age-aggregated measure, since all the values of  $x$  are included in the integral from age 0 to  $\omega$ .

The change over time in this variable can be studied by applying derivatives respect to time. Let a dot on top of the variable denote the derivative of the variable respect to time (notation which derives from work by Newton (Newton 1704)). For example the derivative of the variable  $v(x, t)$  with respect to time  $t$  is:

$$\dot{v}(x, t) = \frac{\partial}{\partial t}v(x, t).$$

Another useful notation is the acute accent to denoted the relative derivative, or the intensity with respect to time  $t$ . For example, the relative derivative of the weighing function  $w(x, t)$  is written as

$$\acute{w}(x, t) = \frac{\frac{\partial}{\partial t}w(x, t)}{w(x, t)}.$$

The main equation of the VCR method is that the change over time of the demographic variable  $\bar{v}(t)$  is equal to a direct and compositional component as

$$\dot{\bar{v}} = \bar{\dot{v}} + Cov(v, \acute{w}). \quad (7)$$

The notation of  $\dot{\bar{v}}$  represents the change in the weighted average, where the notation  $\bar{\dot{v}}$  represents the weighted average of changes in the variable of interest  $v(x, t)$ . This is the Direct effect of the variable of interest in the demographic measure of study. At the same time, the covariance measures the structural or composition change, as well as the effect in population heterogeneity. This term will be positive when both  $v(x, t)$  and  $\acute{w}(x, t)$  have similar age pattern (declining/increasing at the same ages), and negative when they oppose in trend.

The covariance term can be expressed as a function of the two variables within the parenthesis and the weighting function:

$$\begin{aligned}
Cov(v, \dot{w}) &= E[(v - \bar{v})(\dot{w} - \bar{\dot{w}})] \\
&= E(v\dot{w}) - E(v)E(\dot{w}) \\
&= \frac{\int_0^\omega v(x, t)\dot{w}(x, t)w(x, t)dx}{\int_0^\omega w(x, t)dx} - \frac{\int_0^\omega v(x, t)w(x, t)dx}{\int_0^\omega w(x, t)dx} \frac{\int_0^\omega \dot{w}(x, t)w(x, t)dx}{\int_0^\omega w(x, t)dx}
\end{aligned}$$

## Estimation

Equations of the VCR method are in continuous but our demographic information is only available by year or specific age groups. As such this decomposition method will rely on approximations of derivatives and integrals. The mathematical estimations of the method developed by Vaupel and Canudas-Romo in 2002 are here introduced.

For estimating derivatives, is usually useful to first calculate relative derivatives or

$$\dot{v}(x, t) = \frac{\frac{\partial}{\partial t}v(x, t)}{v(x, t)}.$$

Relative derivatives can also be expressed as the derivatives of the logarithm of the function.

$$\dot{v}(x, t) = \frac{\partial}{\partial t} \ln[v(x, t)],$$

which during approximation of two values of  $v(x, t)$  at time  $t$  and  $t + h$  can be written as:

$$\dot{v}(x, t) \approx \ln \left[ \frac{v(x, t + h)}{v(x, t)} \right] / h,$$

dividing by  $h$  produces an annual relative derivative.

Demographic variables are usually assumed to change uniformly across time (S. Preston, Heuveline, and Guillot 2001). Therefore we can approximate the derivatives with respect to time  $t$  as the product of a relative derivative with respect to time  $t$  multiplying the mid-point value of the variable. The mid-point value can be written as

$$v\left(x, t + \frac{h}{2}\right) \approx v(x, t) \exp\left(\frac{h}{2} \dot{v}\left(x, t + \frac{h}{2}\right)\right) \\ \approx [v(x, t)v(x, t + h)]^{\frac{1}{2}}.$$

Finally, the derivative is written as:

$$\dot{v}(x, t) = \dot{v}(x, t)v\left(x, t + \frac{h}{2}\right).$$

Readers should be reminded that the logarithmic function is not defined for negative numbers. Additionally, ratios with denominators “0” or “NA” are not defined for logarithms:

```
# We have 0 on the denominator
```

```
log(1/0)
```

```
[1] Inf
```

```
# or we have NA on the denominator
```

```
log(1/NA)
```

```
[1] NA
```

Therefore we have written a function for the logarithm function that can be used in the decomposition.

```
log_func <- function(v2,v1,time){
  v1 <- ifelse(v1==0|is.na(v1),1,v1)
  div <- ifelse(v2/v1==0|is.na(v2/v1),1,v2/v1)
  result <- log(div)/time
  return(result)
}
```

Now we can try out the function and see if it works

```
log_func(1,0,1)
```

[1] 0

Although the presentation above refer to changes over time  $t$ , the same method can also be written as between-population comparisons, as shown later in this workshop (Canudas-Romo and Guillot 2015).

## Examples of the VCR method

### Crude Death Rate (CDR)

The decomposition for the crude death rate, and using the same notation of the first day can be written as:

$$\dot{\bar{m}} = \bar{m} + Cov(m, r),$$

where  $\bar{m}$  is the crude death rate,  $m(x, t)$  is the age-specific death rate, and  $r(x, t)$  the age-specific growth rate of the population. So why is the crude death rate changing, partially because the age-specific death rates are changing, captured by the  $\dot{\bar{m}}$  term. The covariance captures then the common (or not) age pattern in the age-specific death rates and growth rates.

```
Year1 <- 2010

Year2 <- 2020

KOR_pop <- read.table("data/standard/KOR.Population.txt",
                      header = T, skip = 2)

KOR_pop1 <-
  (KOR_pop[KOR_pop$Year==Year1, "Female"] *
   KOR_pop[KOR_pop$Year==Year1+1, "Female"])^0.5

KOR_pop2 <-
  (KOR_pop[KOR_pop$Year==Year2, "Female"] *
   KOR_pop[KOR_pop$Year==Year2+1, "Female"])^0.5

KOR_Mx <- read.table("data/standard/KOR.Mx_1x1.txt",
                     header = T, skip = 2)

KOR_Mx1 <- KOR_Mx[KOR_Mx$Year==Year1, "Female"]
```

```

KOR_Mx2 <- KOR_Mx[KOR_Mx$Year==Year2,"Female"]

KOR_comp1 <- (KOR_pop1)/sum(KOR_pop1)

KOR_comp2 <- (KOR_pop2)/sum(KOR_pop2)

KOR_comp <-
  (KOR_pop1*KOR_pop2)^0.5/
  sum((KOR_pop1*KOR_pop2)^0.5)

CDR_Y1 <- sum(KOR_Mx1*KOR_comp1)*1000

CDR_Y2 <- sum(KOR_Mx2*KOR_comp2)*1000

CDR_diff <-
  log(CDR_Y2/CDR_Y1)/(Year2-Year1)*
  (CDR_Y1*CDR_Y2)^0.5

CDR_direct <-
  sum(log_func(KOR_Mx2,KOR_Mx1,Year2-Year1)*
    (KOR_Mx2*KOR_Mx1)^0.5*KOR_comp,na.rm = T)*1000

CDR_comp <-
  (sum((KOR_Mx2*KOR_Mx1)^0.5*
    log_func(KOR_pop2,KOR_pop1,
      Year2-Year1)*KOR_comp)-
    sum((KOR_Mx2*KOR_Mx1)^0.5*KOR_comp)*
    sum(log_func(KOR_pop2,KOR_pop1,
      Year2-Year1)*KOR_comp))*1000

CDR_decomp <- (CDR_direct+CDR_comp)

table <-
matrix(round(c(CDR_Y1,CDR_Y2,
  CDR_diff,CDR_direct,
  CDR_comp,CDR_decomp),2),
  ncol=1)

colnames(table) <- "Korea"

```

```

row.names(table) <- c(paste0("CDR in ",Year1),
                      paste0("CDR in ",Year2),
                      "CDR annualized difference",
                      "Direct component",
                      "Compositional component",
                      "Total estimated difference")

kable(table,caption = "CDR decomposition (VCR)")

```

Table 14: CDR decomposition (VCR)

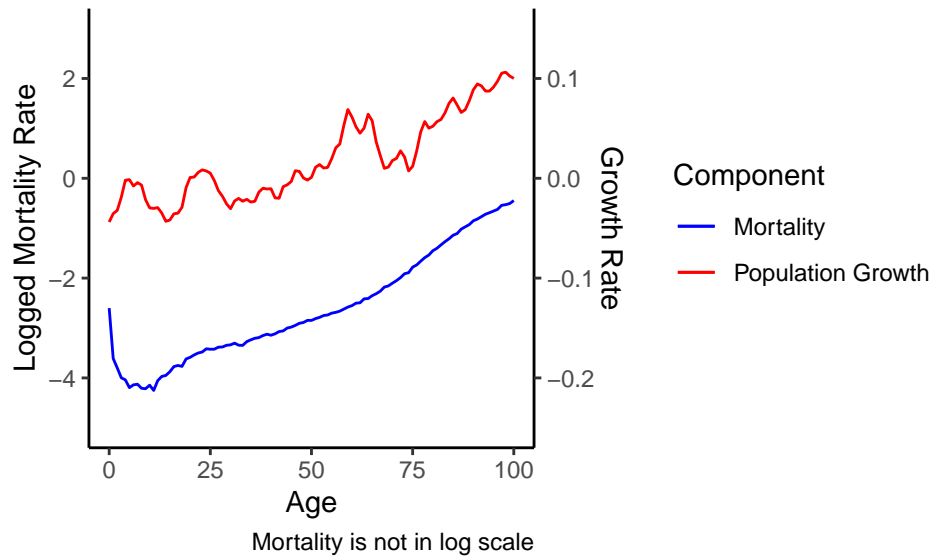
	Korea
CDR in 2010	4.54
CDR in 2020	5.42
CDR annualized difference	0.09
Direct component	-0.16
Compositional component	0.25
Total estimated difference	0.09

Note: Results are multiplied by 1000

```

ggplot()+
  geom_line(aes(x=0:110,
                y=log10((KOR_Mx2*KOR_Mx1)^0.5),
                color = "Mortality"))+
  geom_line(aes(x=0:110,
                y=log_func(KOR_pop2,KOR_pop1,Year2-Year1)*20,
                color = "Population Growth"))+
  scale_x_continuous(limits = c(0,100))+
  scale_y_continuous(limits = c(-5,3),
                    sec.axis = sec_axis(~./20,
                                         name = "Growth Rate"))+
  scale_color_manual(values = c("blue","red"))+
  labs(x="Age",y="Logged Mortality Rate",color="Component",
       caption = "Mortality is not in log scale")+
  theme_classic()

```



```
# Please notice the second axis of the figure,
# in this case we are including two lines in the
# ggplot system which have different scale,
# into one plot region. We will need to change
# the scale of one line by multiplying a factor
# or adding or subtracting some numbers.
```

## General Fertility Rate

The decomposition of the general fertility rate, and using the same notation of the first day can be written as:

$$\dot{\bar{f}} = \bar{\dot{f}} + Cov(f_x, r_f),$$

where  $r_f(x, t) = \dot{P}_f(x, t)$  is the age-specific population growth rates for females only.

```
Year1 <- 2010
Year2 <- 2020

ASFR_KOR <-
  read.table("data/Standard/KORasfrRR.txt",
             header=TRUE, fill=TRUE, skip=2)
```



```

ASFR_KOR$Age[ASFR_KOR$Age=="12-"] <- "12"
ASFR_KOR$Age[ASFR_KOR$Age=="55+"] <- "55"
ASFR_KOR$Age <- as.numeric(ASFR_KOR$Age)

ASFR_KOR1 <- ASFR_KOR[ASFR_KOR$Year==Year1,"ASFR"]
ASFR_KOR2 <- ASFR_KOR[ASFR_KOR$Year==Year2,"ASFR"]

Pop_KOR <-
  read.table("data/Standard/KOR.Population.txt",
             header=TRUE,fill=TRUE, skip=2)

Pop_KOR$Age[Pop_KOR$Age=="110+"] <- "110"
Pop_KOR$Age <- as.numeric(Pop_KOR$Age)

Pop_KOR <- Pop_KOR[Pop_KOR$Age%in%c(12:55),]

Pop_KOR1 <-
  (Pop_KOR[Pop_KOR$Year==Year1,"Female"]*
   Pop_KOR[Pop_KOR$Year==Year1+1,"Female"])^0.5

Pop_KOR2 <-
  (Pop_KOR[Pop_KOR$Year==Year2,"Female"]*
   Pop_KOR[Pop_KOR$Year==Year2+1,"Female"])^0.5

Comp_KOR1 <- Pop_KOR1/sum(Pop_KOR1)

Comp_KOR2 <- Pop_KOR2/sum(Pop_KOR2)

Comp_avg <-
  (Pop_KOR1*Pop_KOR2)^0.5/
  sum((Pop_KOR1*Pop_KOR2)^0.5)

GFR_Y1 <- sum(ASFR_KOR1*Comp_KOR1)*1000

GFR_Y2 <- sum(ASFR_KOR2*Comp_KOR2)*1000

GFR_diff <-
  log(GFR_Y2/GFR_Y1)/(Year2-Year1)*
  (GFR_Y1*GFR_Y2)^0.5

```

```

GFR_direct <-
  sum(log_func(ASFR_KOR2,ASFR_KOR1,Year2-Year1)*
      (ASFR_KOR2*ASFR_KOR1)^0.5*
      Comp_avg,na.rm = T)*1000

GFR_comp <-
  (sum((ASFR_KOR2*ASFR_KOR1)^0.5*
      log_func(Pop_KOR2,Pop_KOR1,Year2-Year1)*Comp_avg)-
   sum((ASFR_KOR2*ASFR_KOR1)^0.5*Comp_avg)*
   sum(log_func(Pop_KOR2,Pop_KOR1,Year2-Year1)*Comp_avg))*1000

GFR_decomp <- (GFR_direct+GFR_comp)

table <-
matrix(round(c(GFR_Y1,GFR_Y2,
              GFR_diff,GFR_direct,
              GFR_comp,GFR_decomp),2),
       ncol=1)

colnames(table) <- "Korea"

row.names(table) <- c(paste0("GFR in ",Year1),
                      paste0("GFR in ",Year2),
                      "GFR annualized difference",
                      "Direct component",
                      "Compositional component",
                      "Total estimated difference")

kable(table,caption = "GFR decomposition (VCR)")

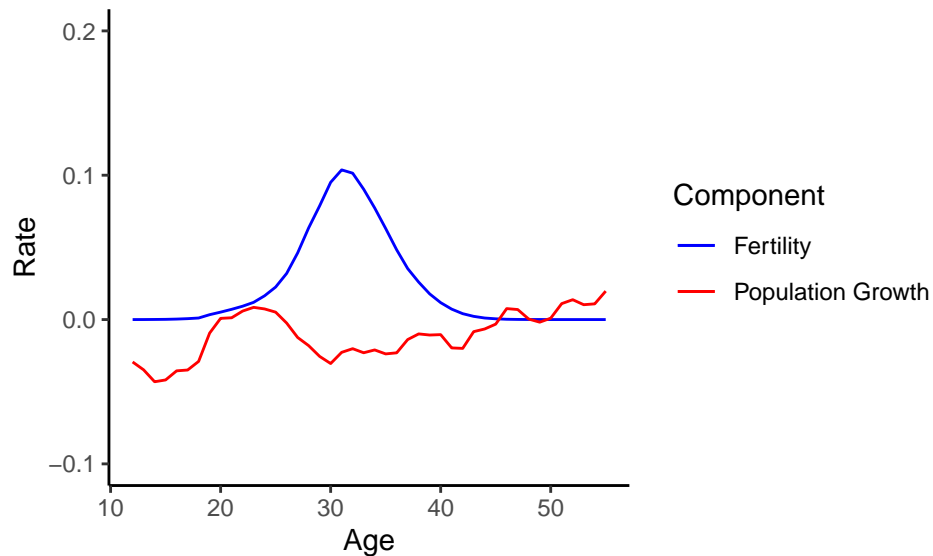
```

Table 15: GFR decomposition (VCR)

	Korea
GFR in 2010	28.24
GFR in 2020	18.06
GFR annualized difference	-1.01
Direct component	-0.79
Compositional component	-0.20
Total estimated difference	-0.99

Note: Results are multiplied by 1000

```
ggplot()+  
  geom_line(aes(x=12:55,  
                y=(ASFR_KOR1*ASFR_KOR2)^0.5,  
                color = "Fertility"))+  
  geom_line(aes(x=12:55,  
                y=log_func(Pop_KOR2,Pop_KOR1,Year2-Year1),  
                color = "Population Growth"))+  
  scale_x_continuous(limits = c(12,55))+  
  scale_y_continuous(limits = c(-0.1,0.2))+  
  scale_color_manual(values = c("blue","red"))+  
  labs(x="Age",y="Rate",color="Component")+  
  theme_classic()
```



### Mean Age of the Population (MAP)

The decomposition for mean age of the population (MAP), which was seen on the first day and can be expressed as

$$\bar{a} = \frac{\int_0^{\omega} aP(a, t)da}{\int_0^{\omega} P(a, t)da}$$

can be written as:

$$\dot{\bar{a}} = Cov(a, r),$$

where  $r(x, t)$  is, as before, the age-specific growth rate, or relative derivative of the population counts, or  $r(x, t) = \dot{P}(x, t)$ . There is only one term in the change of the MAP since age does not change with time and its derivative is equal to zero, or  $\dot{\bar{a}} = 0$ .

```
Year <- 2020

KOR_pop <- read.table("data/standard/KOR.Population.txt",
                      header = T, skip = 2)

KOR_pop <- KOR_pop[KOR_pop$Year==Year, "Female"]

JPN_pop <- read.table("data/standard/JPN.Population.txt",
                      header = T, skip = 2)

JPN_pop <- JPN_pop[JPN_pop$Year==Year, "Female"]

MAP_KOR <- sum(0:110*KOR_pop)/sum(KOR_pop)

MAP_JPN <- sum(0:110*JPN_pop)/sum(JPN_pop)

MAP_diff <- log(MAP_JPN/MAP_KOR)*(MAP_JPN*MAP_KOR)^0.5

comp_pop <- (JPN_pop*KOR_pop)^0.5/sum((JPN_pop*KOR_pop)^0.5)

MAP_cov <-
  sum(0:110*log_func(JPN_pop, KOR_pop, 1)*comp_pop) -
  sum(0:110*comp_pop)*
  sum(log_func(JPN_pop, KOR_pop, 1)*comp_pop)

table <- matrix(round(c(MAP_KOR, MAP_JPN,
                      MAP_diff,
                      MAP_cov), 2), ncol = 1)

row.names(table) <- c("MAP for Korea",
                     "MAP for Japan",
```

```

      "Difference in MAP",
      "Estimated Difference")

colnames(table) <- "Korea"

kable(table,caption = "MAP decomposition (VCR)")

```

Table 16: MAP decomposition (VCR)

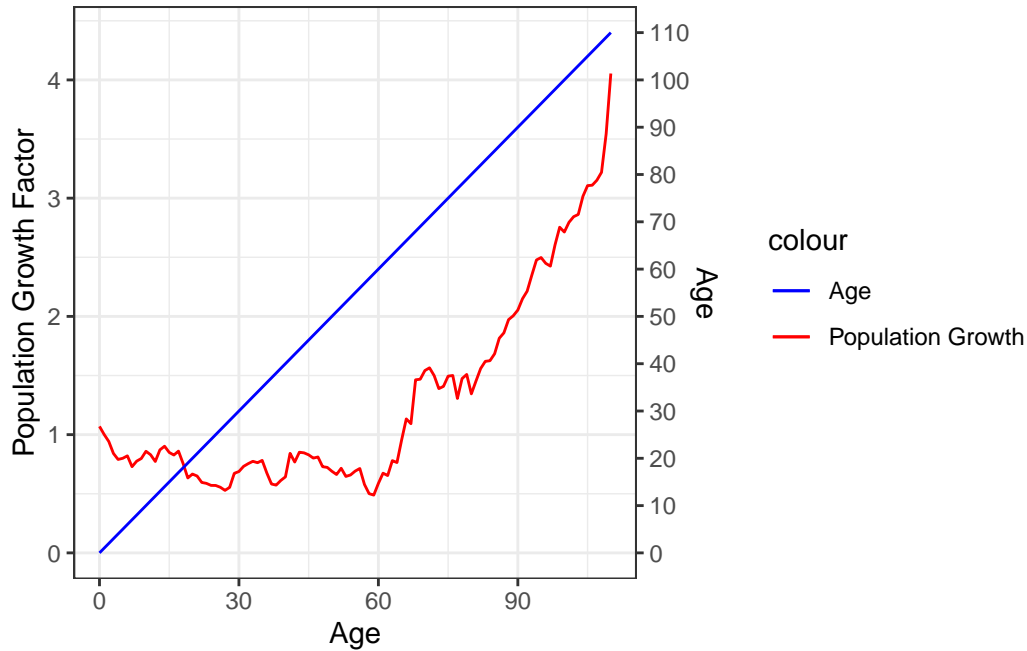
	Korea
MAP for Korea	43.14
MAP for Japan	48.72
Difference in MAP	5.58
Estimated Difference	5.52

Note: Results are multiplied by 1000

```

ggplot()+
  geom_line(aes(x=0:110,
                y=log_func(JPN_pop,KOR_pop,1),
                color = "Population Growth"))+
  geom_line(aes(x=0:110,
                y=(0:110)/25,
                color = "Age"))+
  scale_y_continuous(sec.axis =
                     sec_axis(trans = ~.*25,
                               breaks=seq(0,110,10),
                               name = "Age"))+
  scale_color_manual(values = c("blue","red"))+
  labs(x= "Age", y = "Population Growth Factor")+
  theme_bw()

```



### Mean Age at Childbearing (MAB)

The decomposition for the mean age at childbearing (MAB) or  $\bar{a}_B$  which can be expressed as the product of ages of mothers by their age-specific fertility rates  $f(a, t)$  at age  $a$  and time  $t$ , as

$$\bar{a}_B = \frac{\int_{\alpha}^{\beta} a f(a, t) da}{\int_{\alpha}^{\beta} f(a, t) da}.$$

Its change over time can be written as:

$$\dot{\bar{a}}_B = Cov(a, \dot{f}).$$

```
Year1 <- 2010

Year2 <- 2020

KOR_fx <- read.table("data/standard/KORasfrRR.txt",
                     header = T, skip = 2)
```

```

KOR_fx1 <- KOR_fx[KOR_fx$Year==Year1,"ASFR"]

KOR_fx2 <- KOR_fx[KOR_fx$Year==Year2,"ASFR"]

MAB_Y1 <- sum(12.5:55.5*KOR_fx1)/sum(KOR_fx1)

MAB_Y2 <- sum(12.5:55.5*KOR_fx2)/sum(KOR_fx2)

MAB_diff <-
  (log(MAB_Y2/MAB_Y1)/(Year2-Year1))*
  (MAB_Y2*MAB_Y1)^0.5

comp <- (KOR_fx2*KOR_fx1)^0.5/
  sum((KOR_fx2*KOR_fx1)^0.5)

MAB_decomp <-
  sum(12.5:55.5*log_func(KOR_fx2,KOR_fx1,
                        Year2-Year1)*comp)-
  sum(12.5:55.5*comp)*
  sum(log_func(KOR_fx2,KOR_fx1,
                Year2-Year1)*comp)

table <- matrix(round(c(MAB_Y1,MAB_Y2,
                        MAB_diff,
                        MAB_decomp),2),ncol = 1)

row.names(table) <- c(paste0("MAB for Korea ",Year1),
                      paste0("MAB for Korea ",Year2),
                      "Difference in MAB",
                      "Estimated Difference")

colnames(table) <- "Korea"

kable(table,caption = "MAB decomposition (VCR)")

```

Table 17: MAB decomposition (VCR)

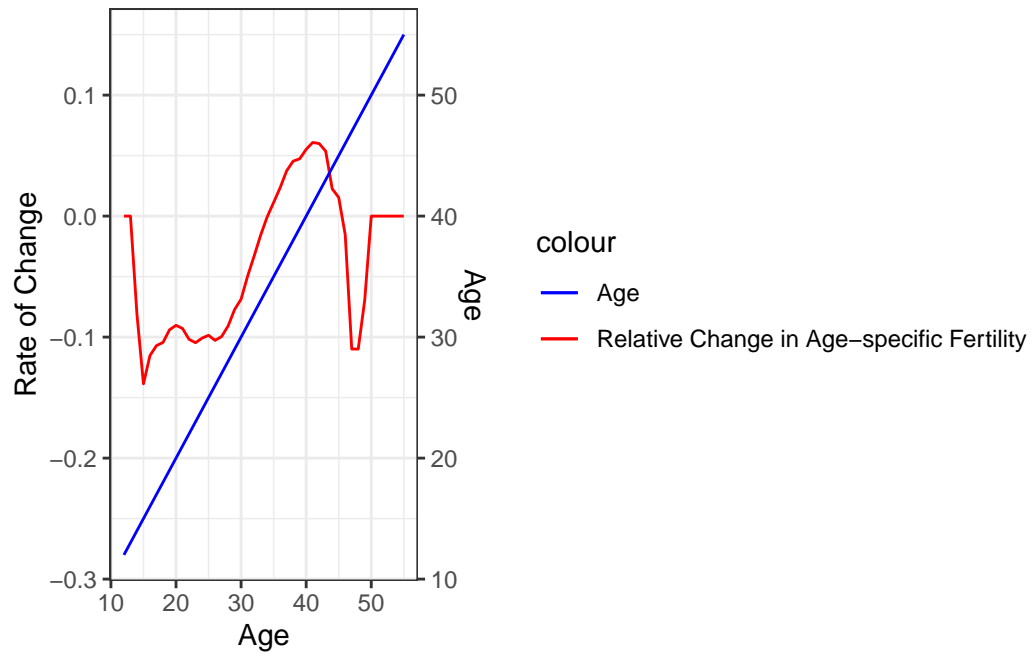
	Korea
MAB for Korea 2010	30.99

	Korea
MAB for Korea 2020	32.88
Difference in MAB	0.19
Estimated Difference	0.19

Note: Results are multiplied by 1000

```
ggplot()+
  geom_line(aes(x=12:55,
                y=log_func(KOR_fx2,KOR_fx1,
                          Year2-Year1),
                color = "Relative Change in Age-specific Fertility"))+
  geom_line(aes(x=12:55,
                y=(12:55)/100-0.4,
                color = "Age"))+
  scale_y_continuous(sec.axis =
                    sec_axis(trans = ~(.+0.4)*100,
                              breaks=seq(0,110,10),
                              name = "Age"))+
  scale_color_manual(values = c("blue","red"))+
  labs(x= "Age", y = "Rate of Change")+
  theme_bw()
```





## Assignment 2

Select one of the suggestions below and submit in ONE page: one Figure (or Table) and a brief paragraph describing the results that you find.

- Apply the Kitagawa or the VCR decomposition on a measure we introduced in class to a different population from HMD or HFD.
- Perform the VCR decomposition on the changes in growth rates in Japan between 2010 and 2020.
- Try decomposing the population growth rate  $\bar{r}(t)$  at time  $t$  as defined earlier as the average of the age-specific growth rates  $r(x, t)$  weighted by the population counts  $P(x, t)$ .

## Day 3. Decomposition for Mortality Measures

### Recap

We have introduced two decomposition methods on weighted averages. The Kitagawa method uses a linear assumption that disentangles the difference in weighted average into a rate component and composition component. At the same time, we also introduced the Vaupel & Canudas-Romo method which disentangles the changes in a demographic variable into a direct effect (changes in the variable of interest) and a covariance term between the variable of interest and the relative derivative of the weighting function.

We will be learning how to decompose two mortality measures that have wide applications in demography.

### Background on Life Expectancy ( $e_0$ )

For the mortality measures, we have introduced the crude death rate and expected number of deaths in our previous sessions. Life expectancy at birth (denote as  $e_0$ ), for example, measures the average years lived by a population if they are subjected to the mortality rates of a given time, this could be in a cohort or period perspective. Let the life table survival function at age  $x$  and time  $t$  be denoted as  $\ell(x, t)$ . Assuming the radix of the population equal to one,  $\ell(0, t) = 1$ , then life expectancy can be expressed as:

$$\begin{aligned} e_0(t) &= \int_0^{\omega} \ell(x, t) dx \\ &= \int_0^{\omega} e^{-\int_0^x \mu(a, t) da} dx, \end{aligned}$$

where  $\mu(a, t)$  stands for the instantaneous age-specific death rates, or also called force of mortality, at age  $a$  and time  $t$ . So life expectancy is the average survival, or expected years lived, by a population at a given time. But, life expectancy can also be expressed as the life table average age at death, which can be expressed as:

$$e_0(t) = \int_0^{\omega} x f(x, t) dx,$$

where the notation  $f(x, t)$  represents the life table death distribution. Note that there is no denominator since the total deaths in the population equal the radix of the population  $\ell(0, t) = 1 = \int_0^{\omega} f(x, t) dx$ .

Life tables are the preferred method of mortality researchers, because by working with death rates the confounding effect of the structure of the population is removed. Life expectancy itself is a standardized mortality measure that enables the comparison between different populations and changes across time. However, we also want to know what exactly is influencing the changes/differences between life expectancies. We can relate the importance of this question with some hypothetical scenarios. For example, if we avert deaths for all infants, would life expectancy be any different? Or, if we successfully extend the life of everyone who is 65+ within the population with 1 extra year of life, would life expectancy be any different? The answers to these questions combined with real-life comparisons with contextual factors (e.g. public health interventions, periodic events) will tell stories of factors contributing to mortality reduction and longevity extension. For the past 50 years, there has been a considerable interest in the age pattern of contributions to the changes/differences in life expectancy. We are going to introduce two of these methodologies: the discrete method of Arriaga (1984), and a continuous method from [Vaupel and Romo \(2003\)](#).

## Discrete Decomposition of Life Expectancy ( $e_0$ )

### Method Explained

The Arriaga decomposition is a discrete decomposition exploring the age pattern contributing the changes/differences in life expectancy. The equation developed by Arriaga is based on the discrete notation using the functions of two life tables and summarized below.

For the discrete notation for the life table, we use  $e_0$  to represent the life expectancy at birth. The difference between two life expectancies from population 1 and population 2 is expressed as  $\Delta e_0 = e_0^2 - e_0^1$ . Survival functions from either population  $i$  can be expressed as  $\ell_x^i$  for age  $x$ . The person-years lived in each age interval can be written as  $L_x^i$ , and the total person-years lived from age  $x$  onward for population  $i$  is denoted as  $T_x^i$ . With this notation in hand the difference in life expectancies can be decomposed by age as

$$\begin{aligned}\Delta e_0 &= \sum_x \Delta_x \\ &= \sum_0^{\omega-1} l_x^1 \left[ \frac{L_x^2}{l_x^2} - \frac{L_x^1}{l_x^1} \right] + \sum_0^{\omega-1} T_{x+1}^2 \left[ \frac{l_x^1}{l_x^2} - \frac{l_{x+1}^1}{l_{x+1}^2} \right] + \Delta_\omega.\end{aligned}\tag{8}$$

In his original paper, Arriaga also separated the effect from mortality changes within each age group which reacts on life expectancy (the direct effect), and the effect of those

changes in later ages (the indirect and interaction effect). The first term on the right side of the equation represents the direct effect, and the second term represents the indirect and interaction effect. The last term  $\Delta_\omega$  denotes the effect from the last age group and consists of the direct effect only as:

$$\Delta_\omega = \ell_x^1 \left[ \frac{T_x^2}{\ell_x^2} - \frac{T_x^1}{\ell_x^1} \right],$$

as before the radix of the populations of interest is equal to 1 for both populations,  $\ell^1(0) = 1 = \ell^2(0)$

In this section of the R code we use R-functions. We first look at the functions to perform the decomposition. The life table function uses the standard from the Coale and Demeny (1983) West model life tables taken from the Preston, Heuveline, and Guillot book (S. Preston, Heuveline, and Guillot 2001).

```
# This is a very simple life table function
# The input values are
# 1. a vector of single age-specific death rates
#    with a length of 111 (age 0 to age 110+).
# 2. an indicator for the sex of either Male ("m") or
#    Female ("f") you are computing.

life.table<-function(mx,sex){

  N<-length(mx)

  ax<-rep(0.5,N)
  # We assume the people who died during the age
  # interval lived half of the year. A very common
  # assumption.

  if(sex=="m"){
    ax[1]<-ifelse(mx[1]<0.107,0.045+mx[1]*2.684,0.330)}
  if(sex=="f"){
    ax[1]<-ifelse(mx[1]<0.107,0.053+2.800*mx[1],0.350)
  }
  # a "ifelse" function to compute the ax for the infant
  # age since the distribution of death for infants are
  # very different from adult years
```

```

qx<-mx/(1+(1-ax)*mx)
# Chiang's conversion from age-specific mortality
# to age-specific probability of death

qx[N] <- 1
# Everyone dies in a life table eventually
# with the last value of qx = 1

px<-1-qx
lx<-100000
for(y in 1:(N-1)){
  lx[y+1]<-lx[y]*px[y]
}
lx <- ifelse(lx<0,0,lx)

dx<-lx*qx
# Calculating the death distribution of the
# life table

Lx<-lx[-1]+ax[-N]*dx[-N]
Lx[N]<-ifelse(mx[N]>0,lx[N]/mx[N],0)
# Person-year lived within each age interval.

Tx<-c()
for(y in 1:N){
  Tx[y]<-sum(Lx[y:N])
}

ex<-Tx/lx
# Calculate life expectancy at each age

Age<-0:110
ALL<-data.frame(Age,mx,lx,dx,Lx,Tx,ex)
return(ALL)
}

# The Arriaga decomposition function
# The input consists of two sets of vector containing
# age-specific mortality rates.

```

```

# At the same time, an indicator of sex of
# either Male ("m") or Female ("f")
# needs to be specified.
# A separate indicator of whether the results comes in
# contributions by age (breakdown = F)
# or contributions by age and direct & indirect+interaction effect
# (breakdown = T).

arriaga <- function(nmx1,nmx2,sex,breakdown=F){

  LT1 <- life.table(nmx1,sex)
  LT2 <- life.table(nmx2,sex)

  # Creating the two life tables

  lx1 <- LT1$lx
  lx2 <- LT2$lx

  Lx1 <- LT1$Lx
  Lx2 <- LT2$Lx

  Tx1 <- LT1$Tx
  Tx2 <- LT2$Tx

  # Specifying the life table statistics we
  # need to perform the Arriaga decomposition.

  if(breakdown==FALSE){
    delta <- rep(0,111)
    for (i in 1:110){
      delta[i] <-
        (lx1[i]/lx1[1])*(Lx2[i]/lx2[i]-Lx1[i]/lx1[i])+
        (Tx2[i+1]/lx1[1])*(lx1[i]/lx2[i]-lx1[i+1]/lx2[i+1])
      delta[111] <-
        (lx1[111]/lx1[1])*(Tx2[111]/lx2[111]-Tx1[111]/lx1[111])
    }
  }

  if(breakdown==T){

```

```

direct <- rep(0,111)
indirect <- rep(0,111)
for (i in 1:110){
  direct[i] <-
    (lx1[i]/lx1[1])*(Lx2[i]/lx2[i]-Lx1[i]/lx1[i])
  direct[111] <-
    (lx1[111]/lx1[1])*(Tx2[111]/lx2[111]-Tx1[111]/lx1[111])
  indirect[i] <-
    (Tx2[i+1]/lx1[1])*(lx1[i]/lx2[i]-lx1[i+1]/lx2[i+1])
}
delta <- data.frame(
  age=0:110,
  direct = direct,
  indirect = indirect
)
}

# depends on whether you want to separate the direct and indirect
# effect, the "delta" results can either be a vector of contributions
# or a data.frame containing the direct and indirect components of
# the decomposition.
return(delta)
}

```

## Examples

We are going to illustrate how to perform the Arriaga decomposition with the function and interpret the results. We are going to use the comparison between Korea and USA in 2019 as an example.

```

USA_Mx <- read.table("data/MortDecomp/USA.Mx_1x1.txt",
  header = T, skip = 2, fill = T)

KOR_Mx <- read.table("data/MortDecomp/KOR.Mx_1x1.txt",
  header = T, skip = 2, fill = T)

ex_diff <-
  life.table(KOR_Mx[KOR_Mx$Year==2019,3], "f")$ex[1]-

```



```

life.table(USA_Mx[USA_Mx$Year==2019,3], "f")$ex[1]

ex_decomp <-
arriaga(as.numeric(USA_Mx[USA_Mx$Year==2019,3]),
        as.numeric(KOR_Mx[KOR_Mx$Year==2019,3]),
        sex = "f",
        breakdown = F)

table <-
matrix(round(
  c(life.table(KOR_Mx[KOR_Mx$Year==2019,3], "f")$ex[1],
    life.table(USA_Mx[USA_Mx$Year==2019,3], "f")$ex[1],
    ex_diff,
    sum(ex_decomp)),
  1), ncol = 1)

row.names(table) <-
c("Life expectancy at birth for Korea",
  "Life expectancy at birth for USA",
  "Life expectancy difference between Korea and USA",
  "Estimated difference from decompositon")

colnames(table) <- "Arriaga"

kable(table, caption = "Female life expectancy gap, Korea - USA, 2019 ")

```

Table 18: Female life expectancy gap, Korea - USA, 2019

	Arriaga
Life expectancy at birth for Korea	86.3
Life expectancy at birth for USA	81.5
Life expectancy difference between Korea and USA	4.9
Estimated difference from decompositon	4.9

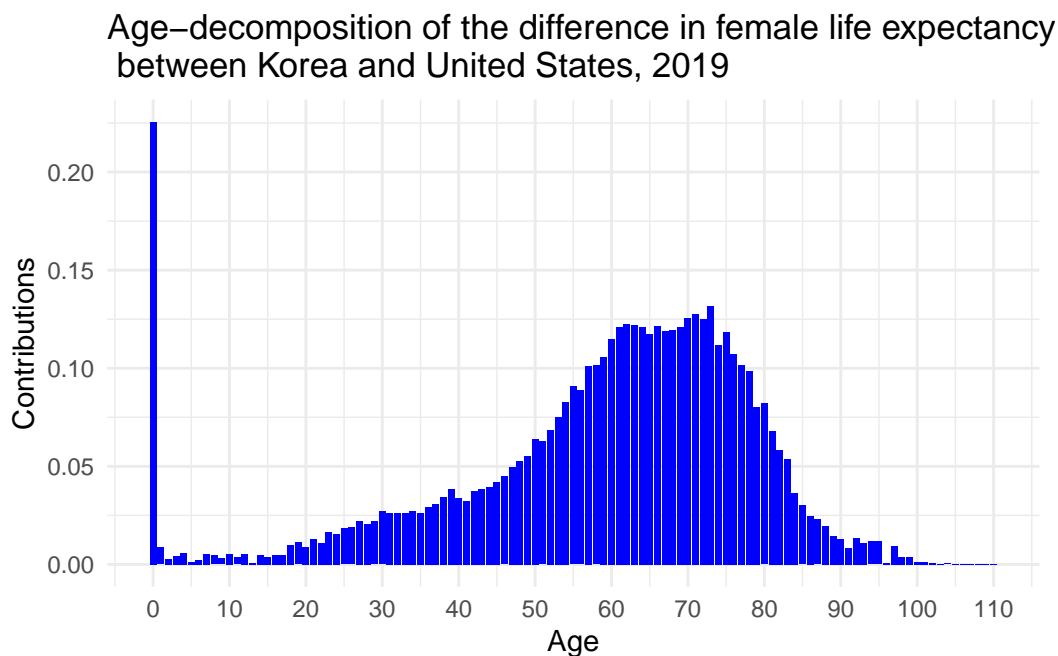
We can also visualize the age contribution with a figure.

```

ggplot()+
  geom_col(aes(x=0:110, y=ex_decomp), fill="blue")+

```

```
scale_x_continuous(breaks = seq(0,110,10))+
theme_minimal()+
labs(title = "Age-decomposition of the difference in female life expectancy \n bet
      x="Age",y="Contributions")
```



Infant mortality and older age mortality contribute substantially to this gap. Apart from cross-country comparisons, we can also compare across time for one population.

At the same time, remember Arriaga decomposition can also separate the age contributions to life expectancy differences into two components: the direct component and indirect component. We will illustrate this with the same example between Korea and USA.

```
USA_Mx <- read.table("data/MortDecomp/USA.Mx_1x1.txt",
                    header = T,skip = 2,fill = T)

KOR_Mx <- read.table("data/MortDecomp/KOR.Mx_1x1.txt",
                    header = T,skip = 2,fill = T)

ex_diff <-
  life.table(KOR_Mx[KOR_Mx$Year==2019,3],"f")$ex[1]-
```

```

life.table(USA_Mx[USA_Mx$Year==2019,3], "f")$ex[1]

ex_decomp <-
arriaga(as.numeric(USA_Mx[USA_Mx$Year==2019,3]),
        as.numeric(KOR_Mx[KOR_Mx$Year==2019,3]),
        sex = "f",
        breakdown = T)

table <-
matrix(round(
  c(life.table(KOR_Mx[KOR_Mx$Year==2019,3], "f")$ex[1],
    life.table(USA_Mx[USA_Mx$Year==2019,3], "f")$ex[1],
    ex_diff,
    sum(ex_decomp$direct),
    sum(ex_decomp$indirect),
    sum(ex_decomp$direct)+sum(ex_decomp$indirect)),
  1), ncol = 1)

row.names(table) <-
c("Life expectancy at birth for Korea",
  "Life expectancy at birth for USA",
  "Life expectancy difference between Korea and USA",
  "Direct component",
  "Indirect and interaction component",
  "Estimated difference from decompositon")

colnames(table) <- "Arriaga"

kable(table, caption = "Female life expectancy gap, Korea - USA, and its components 2019")

```

Table 19: Female life expectancy gap, Korea - USA, and its components 2019

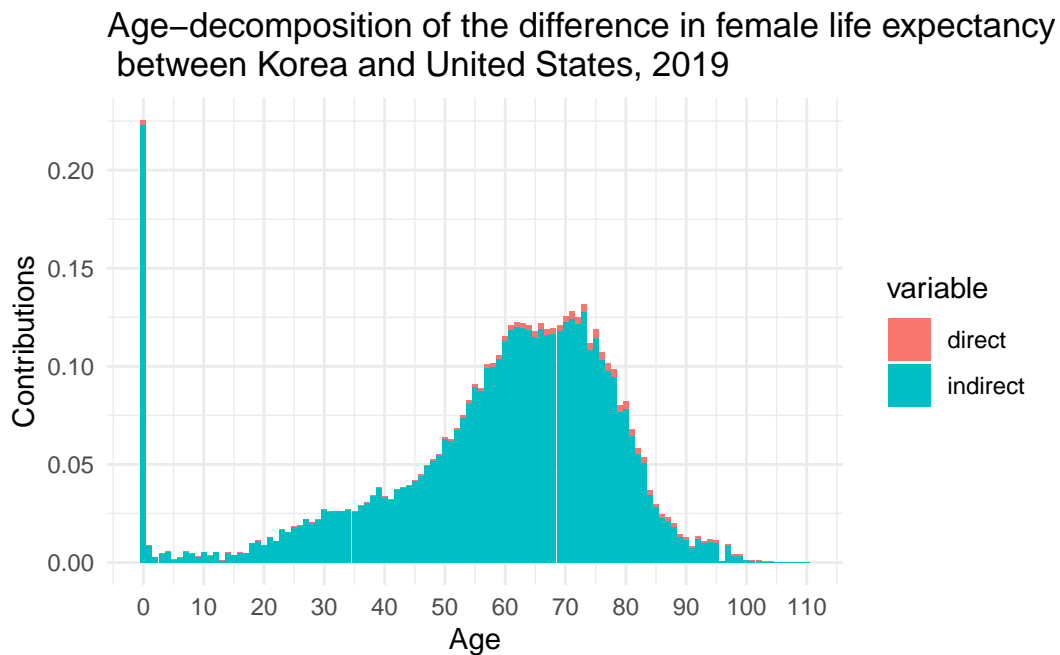
	Arriaga
Life expectancy at birth for Korea	86.3
Life expectancy at birth for USA	81.5
Life expectancy difference between Korea and USA	4.9
Direct component	0.1
Indirect and interaction component	4.7
Estimated difference from decompositon	4.9

We can also visualize the age contribution with a figure.

```
# in order to visualize the data with both direct and indirect
# components, we are going to turn this data.frame from wide to
# long format. We are using the functions from the package
# data.table.

ex_decomp <- melt.data.table(setDT(ex_decomp),
                             id.vars = "age",
                             measure.vars = c("direct","indirect"))

ggplot(ex_decomp,aes(x=age,y=value,fill=variable))+
  geom_col()+
  scale_x_continuous(breaks = seq(0,110,10))+
  theme_minimal()+
  labs(title = "Age-decomposition of the difference in female life expectancy \n bet
        x="Age",y="Contributions")
```



Notice how small the direct component is compared to indirect component? Why is this?

## Age and Cause Contributions to Life Expectancy

### Method Explained

There is one fate we all will face someday and that is death, however, the ages and causes of death differ from person to person. On a population level, understanding these causes of death, and how they affect life expectancy is extremely valuable for set targeted health interventions and other public health policies.

$$\begin{aligned}\Delta e_0 &= \sum_x \Delta_x \\ &= \sum_x \sum_c \Delta_x^c\end{aligned}\tag{9}$$

where the component  $\Delta_x^c$  is calculated based on the age- and cause-specific death rates as well as the overall mortality death rates as

$$\begin{aligned}\Delta_x^c &= \frac{m_x^c(2) - m_x^c(1)}{m_x(2) - m_x(1)} \\ &= \Delta_x \times \frac{R_x^c(2)m_x(2) - R_x^c(1)m_x(1)}{m_x(2) - m_x(1)},\end{aligned}$$

where  $m_x^c(i)$  is the age- and cause-specific death rate at age  $x$ , cause  $c$  and time  $t$ , for population  $i$ ;  $R_x^c(i)$  is the fraction of deaths at age  $x$  due to cause  $c$  in population  $i$ ;  $m_x(i)$  is the age-specific death rate in population  $i$ ; and  $\Delta_x$  is the age-contribution calculated before.

### Data Description

The data for this example comes from the [Human Causes of Death database \(HCoD\)](#). The causes of death here presented are the underlying cause of death, which are the causes directly contributing to a person's passing. This cause is then identified by a medical doctor and subsequently recorded in the "International Classification of Diseases", or "ICD".

The data comes with smoothed counts of deaths from each causes from age 0 to 110 for US from the year 2010 to 2018, which we will be using to illustrate the decomposition.

There are 16 categories of causes within this data set which accounts for all of the deaths occurred within the year. In reality there are more causes to breakdown, but the most common causes of death have been selected and the rest are included in the group of "Other causes". This is particularly useful for visualization purposes: 1. Neoplasms

(column 2) 2. Cardiovascular Diseases (column 6-7) 3. External causes (column 16) 4. Other causes (column 1,3-5,8-15)

For more information on the causes included in this data, you can click [here](#)

## Examples

```
USA_Mx <- read.table("data/MortDecomp/USA.Mx_1x1.txt",
                    header = T, skip = 2, fill = T)

Mx1 <- USA_Mx[USA_Mx$Year==2010,3]

Mx2 <- USA_Mx[USA_Mx$Year==2018,3]

ex_diff <-
  life.table(Mx2,"f")$ex[1]-
  life.table(Mx1,"f")$ex[1]

ex_decomp <-
  arriaga(as.numeric(Mx1),
          as.numeric(Mx2),
          sex = "f",
          breakdown = F)

cause_count <- read.csv("data/MortDecomp/USA_Cause16_counts.csv")

cause_count1 <- cause_count[cause_count$Year==2010&
                           cause_count$Sex==2,]

cause_count1 <- matrix(c(cause_count1$Count),
                      ncol = length(unique(cause_count1$Cause)),
                      nrow = length(unique(cause_count1$Age)))

row.names(cause_count1) <- 0:110

colnames(cause_count1) <- 1:16

cause_total1 <- rowSums(cause_count1)

cause_prop1 <- apply(cause_count1, 2, function(x){x/cause_total1})
```

```

cause_count2 <- cause_count[cause_count$Year==2018&
                           cause_count$Sex==2,]

cause_count2 <- matrix(c(cause_count2$Count),
                      ncol = length(unique(cause_count2$Cause)),
                      nrow = length(unique(cause_count2$Age)))
# Each row is an age and each column is a cause.

row.names(cause_count2) <- 0:110

colnames(cause_count2) <- 1:16

cause_total2 <- rowSums(cause_count2)

cause_prop2 <- apply(cause_count2, 2, function(x){x/cause_total2})

## Now we construct the cause-specific factor.

cause_fac1 <-
  (cause_prop1*Mx1)/
  ifelse((Mx2-Mx1)==0,1,Mx2-Mx1)*ex_decomp

cause_fac2 <-
  (cause_prop2*Mx2)/
  ifelse((Mx2-Mx1)==0,1,Mx2-Mx1)*ex_decomp

# sum(cause_fac2-cause_fac1)
# This get you the same results as the Arriaga method by age

cause_mat <- cause_fac2 - cause_fac1

table <- matrix(round(
  c(life.table(Mx1,"f")$ex[1],
    life.table(Mx2,"f")$ex[1],
    sum(ex_decomp),
    sum(cause_mat[,6:7]),
    sum(cause_mat[,2]),
    sum(cause_mat[,16]),
    sum(cause_mat[,c(1,3:5,8:15)])),

```

```

sum(cause_mat)),2))

# The column number is coded by the number of causes it
# represents in the short list provided in the link above.
# You can crate your own list by changing the column number
# included and create your own list of cause contributions.

row.names(table) <- c(
  "Life expectancy at birth for USA, 2010",
  "Life expectancy at birth for USA, 2018",
  "Life expectancy difference between 2010 and 2018",
  "Cardiovascular disease component",
  "Neoplasms component",
  "External causes component",
  "Other causes component",
  "Estimated total difference from decompositon")

colnames(table) <- "USA"

kable(table, caption = "Arriaga Decomposition by age and cause")

```

Table 20: Arriaga Decomposition by age and cause

	USA
Life expectancy at birth for USA, 2010	81.06
Life expectancy at birth for USA, 2018	81.35
Life expectancy difference between 2010 and 2018	0.28
Cardiovascular disease component	0.01
Neoplasms component	0.35
External causes component	-0.20
Other causes component	0.11
Estimated total difference from decompositon	0.28

The interpretation of the table is that there is a 0.28 difference between the life expectancy in 2010 and 2018 for the USA. Of the 0.28 years difference, the reduction in mortality of neoplasms, and cardiovascular diseases contribute to the increase in life expectancy by 0.37 years, while increasing mortality related to external causes contributed to -0.2 years of life expectancy, offset progress made by reduction in mortality from other causes.



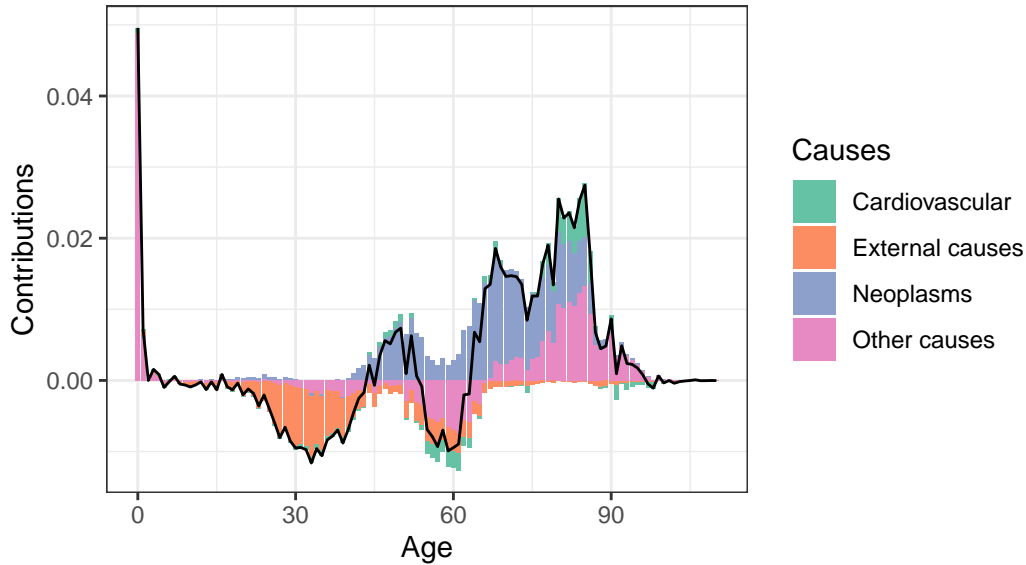
We can also visualize the results with a figure that highlights the age pattern of the cause contributions by age.

```
total_dat <- data.frame(
  Age = 0:110,
  Value = ex_decomp
)

cause_dat <- data.frame(
  Age = rep(0:110,4),
  Cause = rep(c("Cardiovascular",
                "Neoplasms",
                "External causes",
                "Other causes"),each=111),
  Value = c(rowSums(cause_mat[,6:7]),cause_mat[,2],cause_mat[,16],
            rowSums(cause_mat[,c(1,3:5,8:15)]))
)

ggplot()+
  geom_col(data=cause_dat,
           mapping = aes(x=Age,y=Value,fill=Cause))+
  geom_line(data=total_dat,
            mapping = aes(x=0:110,y=Value))+
  scale_fill_brewer(type = "qual",palette = 7)+
  labs(title = "Age- and cause-decomposition of the change in the American \n female",
       x="Age",y="Contributions",fill="Causes")+
  theme_bw()
```

Age– and cause–decomposition of the change in the America female life expectancy between 2010 and 2018.



The solid line is the total age contributions and different color represents contributions from different causes.

## Vaupel & Canudas-Romo Decomposition of Life Expectancy ( $e_0$ )

### Method Explained

The Arriaga decomposition was extended to a continuous framework, first by Pollard (1988). [Vaupel and Romo \(2003\)](#) extended this decomposition to a new dimension, we refer to this method as the VCR decomposition. Additionally to dis-aggregating life expectancy differences/changes into age & cause contributions, it also separates the differences/changes into several meaningful demographic components. There are several levels of the mathematical derivation of this decomposition that are here briefly highlighted.

The decomposition of the change over time in life expectancy can be written as the combination of three components: (1) the improvements in mortality, or relative change in the death rates  $\rho(x) = -\frac{\dot{m}_x}{m_x}$ , (2) remaining life expectancy at age  $x$  or  $e(x, t)$ , and (3) the life tables proportion of deaths  $f(x, t)$ . In mathematical form, this writes:

$$\dot{e}_0 = \int_0^{\omega} \rho(x, t) f(x, t) e(x, t) dx,$$

as before, a dot on top of a variable represents the derivative with respect to changes in time (or it can also be adapted to be comparisons across population).

## Examples

Comparing continuous change versus discrete, year to year change, results in small changes gaps when looking at life expectancy comparisons. An example of this follows using the difference in the American life expectancies in two years, and comparing those with results of the estimation procedures of the continuous derivations.

```
LT1 <- life.table(
  ifelse(is.na(as.numeric(USA_Mx[USA_Mx$Year==2014,4])),
    0, as.numeric(USA_Mx[USA_Mx$Year==2014,4])), sex = "m")
LT2 <- life.table(
  ifelse(is.na(as.numeric(USA_Mx[USA_Mx$Year==2019,4])),
    0, as.numeric(USA_Mx[USA_Mx$Year==2019,4])), sex = "m")

e0_diff <- (log(LT2$ex[1]/LT1$ex[1])/5)*((LT2$ex[1]*LT1$ex[1])^0.5)
e0_diff2 <- (LT2$ex[1]-LT1$ex[1])/5

# Not much of a difference between the two
e0_diff-e0_diff2
```

```
[1] 0.0000000001924
```

```
ex_diff <-
  (-log(LT2$mx/LT1$mx)/5)*
  ((LT1$dx/100000*LT2$dx/100000)^0.5)*
  (LT1$ex+LT1$ex)/2

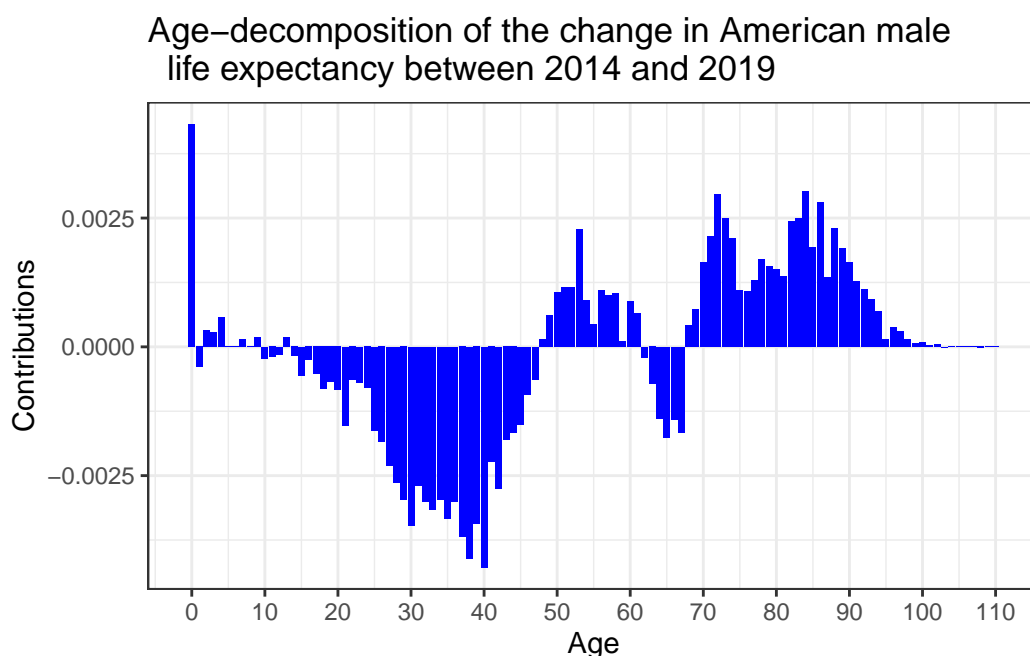
table <- matrix(c(e0_diff, sum(ex_diff)),
  ncol=1)
row.names(table) <- c("observed gap", "estimated gap")
kable(table, caption = "Comparing discrete vs continuous changes in the USA life exp")
```

Table 21: Comparing discrete vs continuous changes in the USA life expectancy.

observed gap	-0.0103
estimated gap	-0.0100

The results are interpreted as the contributions to annual change in life expectancy, which enables comparisons across time and populations in similar ways. We can then visualize the age patterns from above.

```
ggplot()+
  geom_col(aes(x=0:110,y=ex_diff),fill="blue")+
  scale_x_continuous(breaks = seq(0,110,10))+
  theme_bw()+
  labs(title = "Age-decomposition of the change in American male \n life expectancy",
       x="Age",y="Contributions")
```



The age-specific results are similar to the discrete age contributions calculated from the Arriaga decomposition method from the previous section.

The second main relevant results from the VCR decomposition is that the changes in life expectancy can be expressed as the product of lifespan variation and improvements

in mortality, and a covariance term referring to the distribution of improvements in mortality across life expectancy levels. Since life expectancy declines with age this last component reacts if improvements have a greater concentration at young ages (positive covariance) or at old ages (negative covariance), as

$$\dot{e}_0(t) = \bar{\rho}(t)e^\dagger(t) + Cov_f(\rho, e), \quad (10)$$

where  $\rho(a, t) = -\frac{\dot{\mu}(a, t)}{\mu(a, t)}$  is interpreted as the age-specific improvements in mortality, and the  $\bar{\rho}(t) = \int_0^\omega \rho(a, t)f(a, t)da$  is interpreted as the average improvements in mortality.

The  $e^\dagger$  term is calculated from the life expectancies and distribution of deaths as:

$$e^\dagger(t) = \int_0^\omega e(a, t)f(a, t)da.$$

This measure, is called “e-dagger” or “life disparity”, and has two demographic meanings: (1) it measures the average years of life lost due to death; (2) it also measures the degree of variation in mortality, or uncertainty in age at death, in a life table. Finally the  $Cov_f(\rho, e)$  is calculated as:

$$Cov_f(\rho, e) = \int_0^\omega [\rho(a, t) - \bar{\rho}(t)] [e(a, t) - e^\dagger(t)] f(a, t)da.$$

The following R code replicates the results presented in the paper using the VCR decomposition. The replication focused on Table 1 from Vaupel and Romo (2003). Changing years in Y1 and Y2 in the R code, yields similar results to other columns of Table 1.

```
Y1 <- 1950
Y2 <- 1955

# The period of time, this is to calculate annualized change
tp <- Y2-Y1

# For life table, I think they used the both sex one from Sweden so this is the one

LT1 <- read.table('data/MortDecomp/SWE.bltper_1x1.txt',
                  header=T, skip=2)
LT1 <- LT1[LT1$Year==Y1,]
```

```

LT2 <- read.table('data/MortDecomp/SWE.bltpcr_1x1.txt',
                  header=T,skip=2)
LT2 <- LT2[LT2$Year==Y2,]

gap <-
  (life.table(LT2$mx,sex="m")$ex[1]-
   life.table(LT1$mx,sex="m")$ex[1])+
  (life.table(LT2$mx,sex="f")$ex[1]-
   life.table(LT1$mx,sex="f")$ex[1])

# Since it's life table by both sexes,
# in order to minimize the error that
# could arise from the infant age group
# due to differential ax values, we therefore
# do an average of both the female and male
# life table procedure.

gap <- gap/(tp*2)

```

The decomposition is performed as below. Because of the mathematical complexity we will use line by line code for this one.

```

ex <- (LT1$ex+LT2$ex)/2

# from Alyson van Raalte to calculate e-dagger
ineq_edag <- function(age, dx, lx, ex, ax){
  age_length_equal <- all.equal(length(age),length(dx),
                                length(lx),length(ex),
                                length(ax))

  stopifnot(age_length_equal)

  # length of the age interval
  n <- c(diff(age),1)
  explusone <- c(ex[-1],ex[length(age)])
  # the average remaining life expectancy in each age interval
  # (as opposed to the beginning of the interval)
  # ends up being roughly half of the ex between ages
  ex_average <- ex + ax / n * (explusone - ex)

```

```

  rev(cumsum(rev(dx * ex_average))) / lx
}

edag <- (ineq_edag(0:110,LT1$dx,LT1$lx,LT1$ex,LT1$ax)*
        ineq_edag(0:110,LT2$dx,LT2$lx,LT2$ex,LT2$ax))^0.5

rho <- -log(LT2$mx/LT1$mx)/tp

fx <- (LT1$dx/100000*LT2$dx/100000)^0.5

main <- sum(rho*fx)*edag[1]
# direct component, as equation shows is the product
# between average improvements in mortality and
# life disparity or e-dagger at birth.

covar <- sum((rho-sum(rho*fx))*(ex-edag[1])*fx)

gap2 <- main+covar

table <-
  matrix(round(c(LT1$ex[1],LT2$ex[1],gap,
               sum(rho*fx),edag[1],
               main,covar,gap2),3),ncol=1)

row.names(table) <- c(paste0("Life expectancy at ",Y1),
                     paste0("Life expectancy at ",Y2),
                     paste0("Annualized change between ",
                             Y1,"-",Y2),
                     "Average improvements in mortality",
                     "Life disparity at birth",
                     "Direct component",
                     "Covariance component",
                     "Total estimated difference")

colnames(table) <- "SWE"

kable(table,caption = "Decomposition of life expectancy (VCR)")

```

Table 22: Decomposition of life expectancy (VCR)

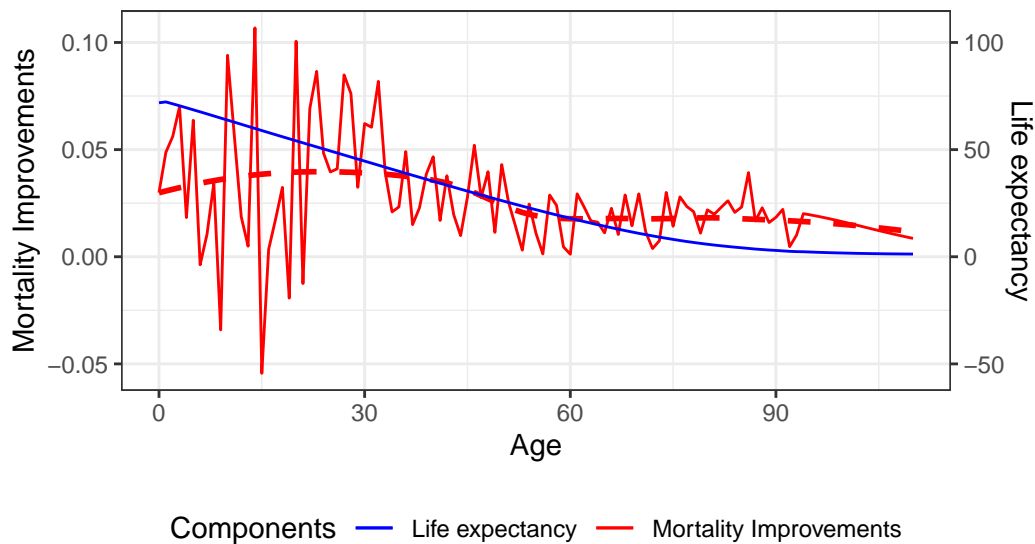
	SWE
Life expectancy at 1950	71.140
Life expectancy at 1955	72.590
Annualized change between 1950-1955	0.290
Average improvements in mortality	0.021
Life disparity at birth	11.995
Direct component	0.248
Covariance component	0.041
Total estimated difference	0.289

The gap observed in the life expectancy difference: 0.2896 and the estimated difference: 0.2892 is 0.0004.

We can also visualize this as a covariance between improvements in mortality and life expectancy.

```
ggplot()+
  geom_line(mapping = aes(x=0:110,y=rho,
                        color="Mortality Improvements"))+
  geom_smooth(mapping = aes(x=0:110,y=rho,
                        color="Mortality Improvements"),se=F,linetype=2,show.legend=F)+
  geom_line(mapping = aes(x=0:110,y=ex/1000,
                        color="Life expectancy"))+
  scale_y_continuous(sec.axis =
    sec_axis(trans=~.*1000,
              name = "Life expectancy"))+
  scale_colour_manual(values = c("blue","red"))+
  theme_bw()+
  labs(x="Age",y="Mortality Improvements", color = "Components",
       caption = "Note: red dotted line is the smoothed curve of
       mortality improvements to help see the trend more
       clearly.")+
  theme(legend.position = "bottom")
```





Note: red dotted line is the smoothed curve of mortality improvements to help see the trend more clearly.

## Background on Cross-sectional Average Length of Life ( $CAL$ )

There are traditionally two perspectives to view human mortality, period and cohort perspective. So far we have been working primarily with the period mortality measures. A period measure refers to the mortality data collected for one specific period, say the year of 2023.

However, we each have different ages and were born in different years, that is we each belong to a birth cohort. Recording data for one specific period is easy, however, tracking the mortality information for cohorts born 100 and more years ago is hard for most populations. Therefore, using period information has its perks of inferring much needed demographic information with some assumptions.

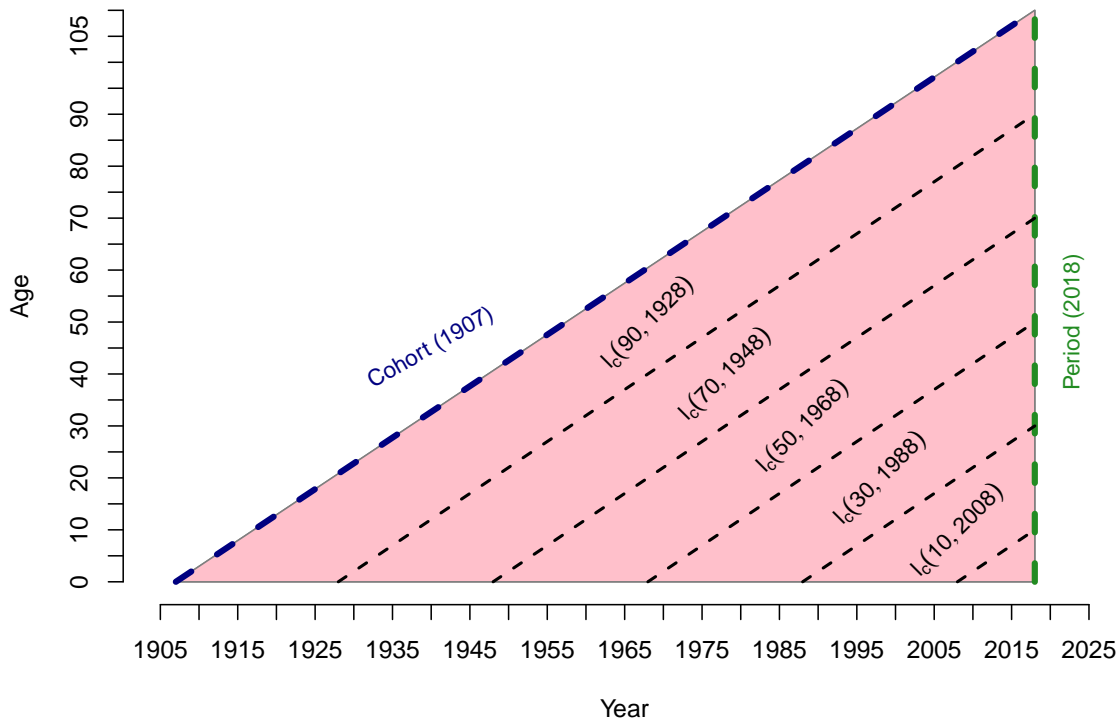
With the period data, there is an assumption that there is a hypothetical cohort living through all these observed conditions of a given period. For example, period life expectancy for 2023 refers to a hypothetical cohort that lives through all the observed age-specific death rates from age 0 to age 110 observed in 2023.

The assumption of the period measure has its merits. However, it also brings challenges of interpretation. For example, life expectancy during the COVID era, say the year of 2020. For that year, the assumption of the period life table is of a hypothetical cohort

living perpetually through COVID-19 pandemic. This scenario is not realistic and say little about the mortality in the long term.

Cohort data tracks the mortality condition of an actual cohort. This cohort can be born 110 years ago and even more. This approach helps us understand how mortality changes from cohort to cohort across time. However, using cohort data requires waiting for the cohort in observation has nearly completed its mortality schedule and all members have died out. However, a complete cohort returns little information to mortality conditions of cohorts born more recently.

A new perspective that tries to consolidate the two perspectives, the period and cohort, by calculating a period life table statistic that at each age is calculated from cohort mortality information pertain to that age. This measure is called the “Cross-Sectional Average Length of Life”, or *CAL*. The idea of *CAL* was first presented by Brouard (1986) and then formally defined by Guillot (2003), with many applications. We will use an illustration to show how to define and calculate *CAL*.



The green line represents the period life expectancy for the year of 2018, calculated

with the period mortality information, while the dark blue line represents the cohort life expectancy for the year of 1907, calculated with the cohort mortality information. Meanwhile, the *CAL* is calculated as the summary of all cohort mortality information presented at the last period, which in this figure is 2018. Each dark dotted line represents the cohort survival function present in year 2018. The interpretation of *CAL* as a mortality measure is that it summarizes the mortality experience of all the cohorts present at time  $t$ , in the figure's case is 2018.

The mathematical expression for *CAL* is written as:

$$\begin{aligned} CAL(t) &= \int_0^{\omega} \ell_c(x, t) dx \\ &= \int_0^{\omega} e^{-\int_0^x \mu(a, t-x+a) da} dx, \end{aligned}$$

where  $\ell_c(x, t)$  is the cohort life table survival function at age  $x$  and time  $t$ , and  $\mu(a, t-x+a)$  is the force of mortality at age  $a$  for the cohort born in year  $t-x$ . The similarity between *CAL*( $t$ ) and the expressions for life expectancy at birth for period and cohort perspectives should be noted.

The decomposition for *CAL* for example by comparing two populations is based on the work by [Canudas-Romo and Guillot \(2015\)](#).

## Decomposition of Cross-sectional Average Length of Life (*CAL*)

### Method Explained

Let  $p_a(t-x, i)$  be the cohort probability of surviving from age  $a$  to  $a+1$  for the cohort born in year  $t-x$  in population  $i$ . The cohort survival function to age  $x$  is then calculated as the product of the age-specific probabilities of surviving as  $\ell_c(x, t, i) = p_0(t-x, i)p_1(t-x, i)\dots p_{x-1}(t-x, i)$ . As the previous decompositions in continuous, a dot on top of *CAL* will correspond to comparisons between populations. The comparison of *CAL* can be separated as:

$$\dot{CAL}(t, \xi) = \int_0^{\omega} \ell_c^*(x, t, \xi) \left[ \sum_{a=0}^x \frac{\dot{p}_a(t-x, \xi)}{p_a(t-x, \xi)} \right] dx.$$

This equation decomposes the *CAL* comparison between two populations regarding a hypothetical variable of inter-population difference  $\xi$ . The notation  $\ell_c^*(x, t, \xi)$  denotes

the average survival function between the two populations. At the same time, the notation  $p_a(t-x, \xi)$  represents the average age-specific probability of surviving to the next age between the two populations.

The following code corresponds to the *CAL* decomposition.

```
# The functions require two input matrix of
# age-specific probability of survival of more
# than 111 years. And the name of the two populations
# we are comparing.

CALDecompFunction<-function(Mx1,Mx2,Y,Name1,Name2){

  CALlx<-c()
  CALlx1<-c()
  CALlx2<-c()
  PxCh<-c()

  YM<-Y-Y1

  for (x in 1:111){
    if (x <(YM+1)){
      px1<-c()
      px2<-c()
      for (z in 1:x){
        px1<-c(px1,Mx1[z,YM-x+z])
        px2<-c(px2,Mx2[z,YM-x+z])
      }

      pxCH<-c(log(px2/px1),rep(0,111-x))

      lx1<-prod(px1)
      lx2<-prod(px2)
    }
    if (x >(YM)){
      px1<-c()
      px2<-c()
      for (z in (x-YM+1):x){
        px1<-c(px1,Mx1[z,YM-x+z])
        px2<-c(px2,Mx2[z,YM-x+z])
      }
    }
  }
}
```

```

    px1<-c(rep(1,(x-YM)),px1)
    px2<-c(rep(1,(x-YM)),px2)
    pxCH<-c(log(px2/px1),rep(0,111-x))

    lx1<-prod(px1)
    lx2<-prod(px2)
  }
  CALlx1<-c(CALlx1,lx1)
  CALlx2<-c(CALlx2,lx2)

  PxCh<-cbind(PxCh,pxCH)

}
CALlx<- t(matrix(rep((CALlx1+ CALlx2)/2,111),111))

PxCh[is.na(PxCh)]<-0

## as Guillot calculates this plus a one for l(0)
A1<-sum(c(1,CALlx1))+.5
A2<-sum(c(1,CALlx2))+.5
A3<-sum(CALlx2)-sum(CALlx1)
A4<-sum(PxCh*CALlx)

#print()
return(
  list(overall = rbind(c(paste("CAL-",Name2),
                          paste("CAL-",Name1),
                          "Diff","est-Diff"),
                      round(c(A2,A1,A3,A4),2)),
        detailed = PxCh*CALlx))
}

# The results returns a list of both age-cohort specific
# information and the aggregated information.

```

## Examples

We are going to illustrate the decomposition with data from Italy and France within HMD which have abundant mortality information from 1816.

```

LT1 <- read.table("data/MortDecomp/ITA.mltper_1x1.txt",
                  header = T, skip=2)

Mx1 <- matrix(c(1-LT1$qx), nrow = 111)

row.names(Mx1) <- 0:110
colnames(Mx1) <- unique(LT1$Year)

LT2 <- read.table("data/MortDecomp/FRATNP.mltper_1x1.txt",
                  header = T, skip=2)

Mx2 <- matrix(c(1-LT2$qx), nrow = 111)

row.names(Mx2) <- 0:110
colnames(Mx2) <- unique(LT2$Year)

Y <- 2020

Y1 <- 1909
# This is the year where we start doing the
# decomposition minus one.

result <- CALDecompFunction(Mx1, Mx2, 2020, "Italy", "France")
# If this return "You need longer series of data, you
# input data series is too small.
# Then, you need to have a good look on the data you
# prepared and try it again.

kable(result$overall,
      caption = "CAL decomposition between France & Italy")

```

Table 23: CAL decomposition between France & Italy

CAL- France	CAL- Italy	Diff	est-Diff
43.33	61.77	-18.44	-18.83

```

library(RColorBrewer)
mypalette<-rev(brewer.pal(8,"YlGnBu"))
mypalette2<-rev(brewer.pal(8,"YlOrRd"))

```

```

WildColors<-c(mypalette[1:4],"white","white",mypalette2[c(6,4,2,1)])
WildColors<-c(WildColors[1:4],"white","white",WildColors[7:10])
levels<-c(-1,-0.1,-0.01,-0.001,-0.0001,0,.0001,.001,.01,.1,1)

options(scipen=10)

customAxis <- function() {
n <- length(levels)
y <- seq(min(levels), max(levels), length.out=n)
rect(0, y[1:(n-1)], 1, y[2:n], col=WildColors)
axis(4, at=y, labels=levels)
}

CALlxDecomp <- result$detailed

# The correct assignment of contributions and the cummulative changes
CALlxD<-matrix(0,111,111)
CALlxDS<-CALlxD

Age<-c(0:110)

YEARS<-c((Y-110):Y)

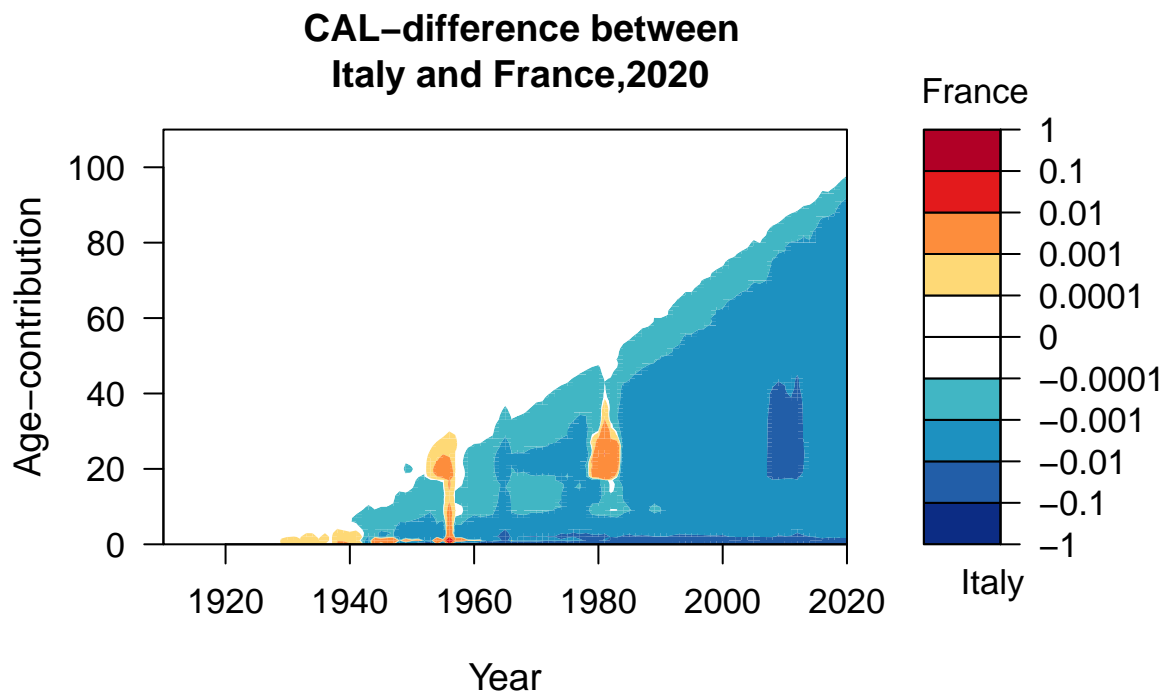
# The correct assignment of contributions and the cumulative changes

for (y in 1:111){
  for (x in 1:y){
    CALlxD[x,(111-y+x)]<-CALlxDecomp[x,y]
    CALlxDS[x,(111-y+x)]<-sum(CALlxDecomp[(1:x),y])
  }
}

par(cex.axis=1.1)
filled.contour(YEARS, Age, t(CALlxD),
               levels=levels,col=WildColors,
               key.axes=customAxis(),ylab="Age-contribution",
               xlab="Year",cex.lab=1.2)
title("CAL-difference between \n Italy and France,2020",
      adj=0.3)

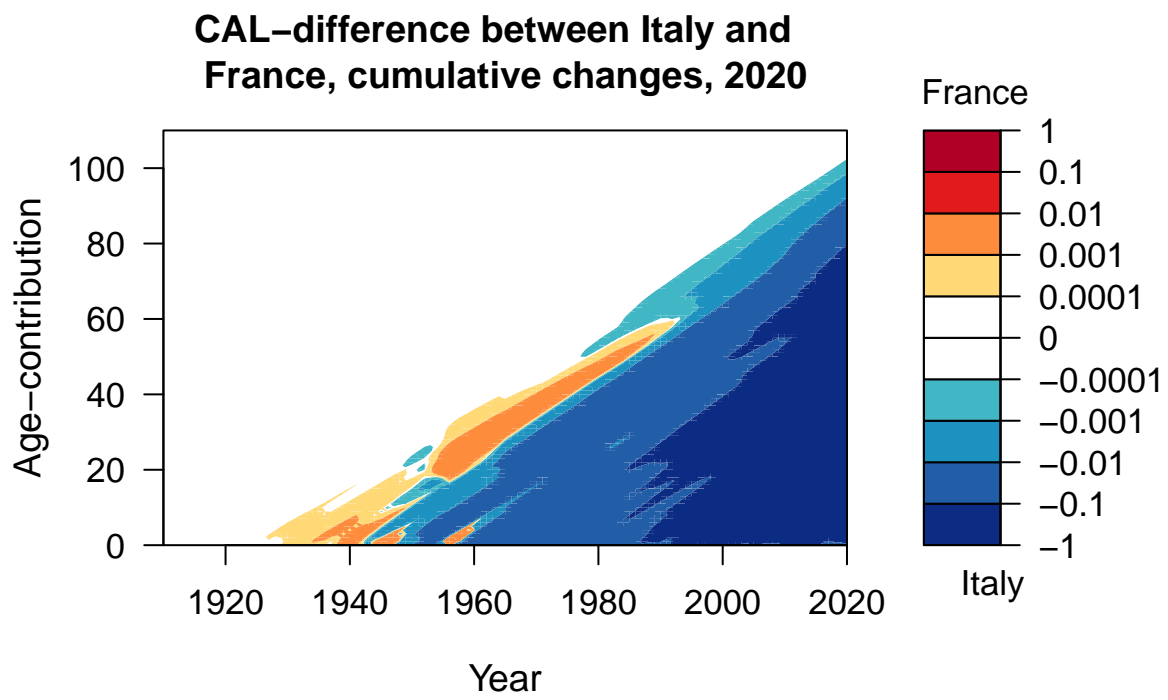
```

```
mtext("Italy",1,0.5,adj=.9,cex=1.1)
mtext("France",3,0.5,adj=.9,cex=1.1)
```



```
par(cex.axis=1.1)
filled.contour(YEARS, Age, t(CALLxDS),
               levels=levels, col=WildColors,
               key.axes=customAxis(), ylab="Age-contribution",
               xlab="Year", cex.lab=1.2)
title("CAL-difference between Italy and \n France, cumulative changes, 2020", adj=0.1)
mtext("Italy",1,0.5,adj=.9,cex=1.1)
mtext("France",3,0.5,adj=.9,cex=1.1)
```





## Assignment 3

Select one of the suggestions below and submit in ONE page: one Figure (or Table) and a brief paragraph describing the results that you find.

- Apply the measures in the examples of this section to a different population from HMD or HFD.
- Compute the age contributions to life expectancy changes between the year 2000 and 2010, and 2010 to 2019 using the Arriaga decomposition (you can choose whether to include the cause component or not). What is happening there? Describe your findings with a Figure and a short summary
- Perform the Vaupel & Canudas-Romo decomposition on life expectancy for both-sex Swedish population for a more recent period (2010-2020).
- (**Theoretical**) The calculation of CAL requires a huge amount of data. So, if we need to compute values similar to CAL for countries with not so abundant mortality data (e.g. Japan, which had mortality information since 1954), what can be done?

## Day 4. Decomposition for Fertility Measures

### Recap

Last session the decomposition for mortality was presented. The decomposition of mortality measures not only include common measures such as crude deaths rate, but also measures where the confounding effect of the structure of the population is removed, such as life expectancy. Comparing such measures and disentangling their age- and cause- contributions can be done for life expectancy differences. Additionally, other dimensions such as average mortality improvements and life disparity can be calculated as elements of the life expectancy comparisons. More importantly, alternative measures that include period and cohort perspectives such as the “cross-average length of life” were introduced. *CAL* aggregates the cohort survival histories of all cohorts present at a given time. Its decomposition illustrates age- and cohort- contributions to population comparisons.

In this session, decomposition methods for fertility measures are introduced: the complete cohort fertility (*CCF*), or *TFR* for cohorts, *CAL* perspective with fertility, and the net reproduction rate (*NRR*).

### Completed Cohort Fertility (*CCF*)

#### Method Explained

The completed cohort fertility (*CCF*) is defined as the average number of babies born by a cohort of females born in year  $t$  who have completed their reproductive years, denoted as  $CCF(t)$ . The cohort *TFR* can only be calculated when all surviving women in the birth cohort have reached the end of their reproductive years, say at age 55. Assuming negligible mortality before this age, this is written as:

$$CCF(t) = \frac{B(t)}{W(t)},$$

where  $B(t)$  is the number of births to mothers born in year  $t$ , and  $W(t)$  is the number of women born in year  $t$ .

We can include other dimensions that affect fertility such as education. The numerator from the previous equation can be partitioned by the education level of the mothers,  $B(t) = \sum_{e=L}^H B^e(t)$ , where the notation  $B^e(t)$  corresponds to the number of births to mothers with education level  $e$ , denoted as  $W^e(t)$ , and education denoted as  $e$  ( $L =$

low;  $M$  = Medium; and  $H$  = high). Therefore, the equation can be further expressed as a product between the education composition ( $E^e(t) = \frac{W^e(t)}{W(t)}$ ) for women at time  $t$ , and the education-specific fertility ( $F^e(t) = \frac{B^e(t)}{W^e(t)}$ ).

$$\begin{aligned} CCF(t) &= \sum_{e=L}^H \frac{B^e(t)}{W(t)} \\ &= \sum_{e=L}^H \frac{B^e(t)}{W^e(t)} \frac{W^e(t)}{W(t)} \\ &= \sum_{e=L}^H F^e(t) E^e(t). \end{aligned}$$

As in previous decompositions, continuous change is assumed in the cohort fertility measure  $CCF(t)$  and the comparisons between different cohorts is decomposed as the effect from changing education distribution and education-specific fertility as

$$C\dot{C}F = \sum_{e=L}^H [\dot{E}^e(t) F^e(t) + E^e(t) \dot{F}^e(t)]. \quad (11)$$

This is the main decomposition in the paper by [Lazzari, Mogi, and Canudas-Romo \(2021\)](#).

## Example

In this example, the data from the Cohort Fertility and Education (CEF) [database](#) is used. The database compiles fertility data from Census information from worldwide populations.

The data used is from South Korea with the information on education composition and education-specific fertility, cleaned by Dr. Ryohei Mogi. The original data included the parity-specific information but we are going to focus only on cohort fertility.

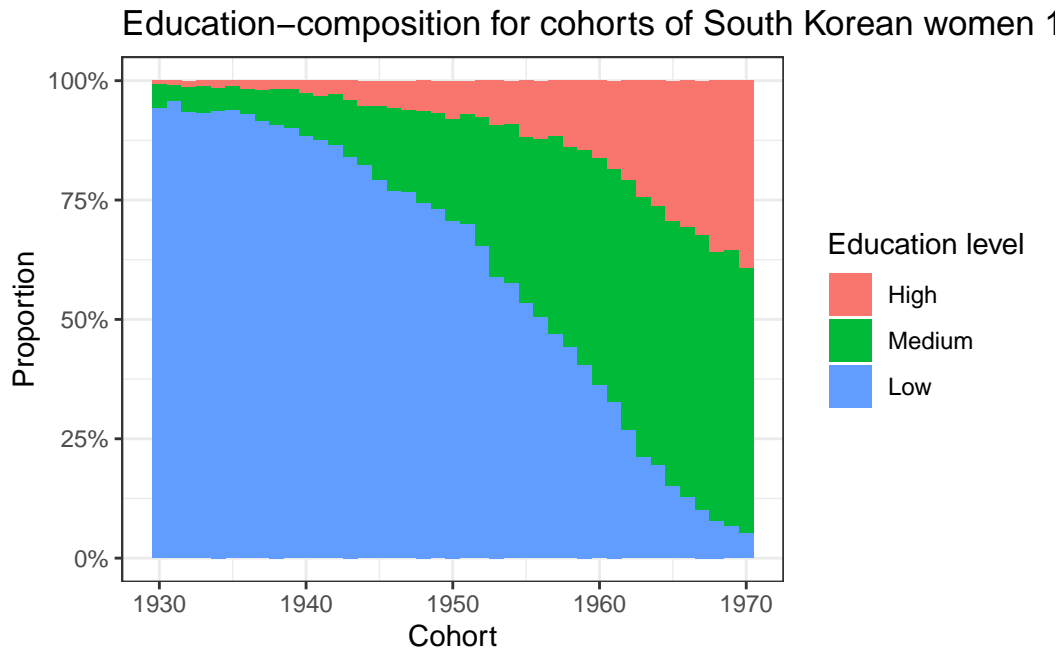
```
data <- read.csv("data/FertDecomp/data_base_prop.csv")

data$edu <- factor(data$edu,
                  levels = c("High", "Medium", "Low"))

# We can first visualize the education composition
```

```
# change across time.
```

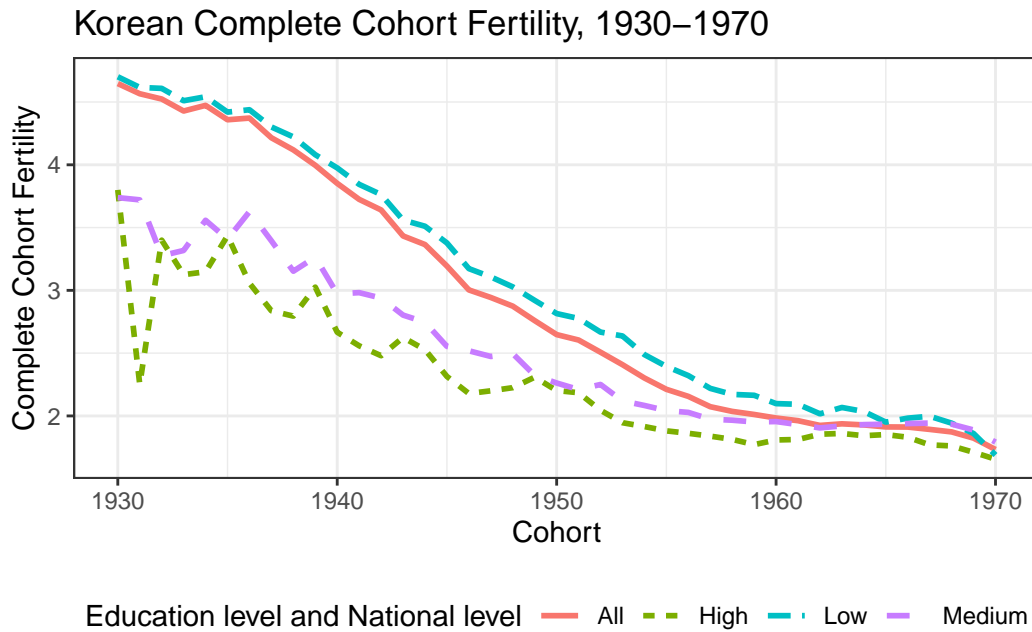
```
ggplot(data,aes(x=cohort,y=Prop_edu,fill=edu))+  
  geom_col(width = 1)+  
  scale_y_continuous(labels = function(x){paste0(x*100,"%")})+  
  theme_bw()+  
  labs(title ="Education-composition for cohorts of South Korean women 1930 to 1970.
```



Korea has gone through substantial improvements in education expansion with later cohorts of people having high education. For example, for the cohort of women born in 1970 over half of them achieved medium education and more than 40% high education. We can also check the trajectories of the *CCF* for each education group over cohorts.

```
data <- read.csv("data/FertDecomp/data_base_fert.csv")  
  
ggplot(data,aes(x=cohort,y=CCF,  
               linetype=edu,color=edu))+  
  geom_line(linewidth=1)+  
  theme_bw()+  
  labs(title = "Korean Complete Cohort Fertility, 1930-1970",
```

```
x="Cohort",y="Complete Cohort Fertility",
color = "Education level and National level",
linetype = "Education level and National level")+
theme(legend.position = "bottom")
```



What would these fertility and education dynamics do to the South Korean *CCF*, for example when comparing changes from cohorts born in 1930s to those born in 1960s and 1970? the decomposition returns precisely the contribution of those components.

```
Year1 <- 1940

Year2 <- 1950

data <- read.csv("data/FertDecomp/data_base_decomp.csv")

data <- data[,c("cohort","edu", "B",
               "We", "W", "E", "Fe")]

data1 <- data[data$cohort==Year1,]
```

```

data2 <- data[data$cohort==Year2,]

mat1 <- as.matrix(data1[,3:7])
row.names(mat1) <- unique(data1$edu)

mat2 <- as.matrix(data2[,3:7])
row.names(mat2) <- unique(data2$edu)

mid_E <- (mat1[,4] * mat2[,4]) ^0.5

mid_F <- (mat1[,5] * mat2[,5]) ^ 0.5

deriv_E <- (log(mat2[,4] / mat1[,4]) / (Year2-Year1)) * mid_E

deriv_F <- (log(mat2[,5] / mat1[,5]) / (Year2-Year1)) * mid_F

Results_E <- deriv_E*mid_F

Results_F <- deriv_F*mid_E

Results <-
  data.frame(Edu = rep(row.names(mat1),3),
             variable = c(rep(c("Fertility",
                                "Education-composition"),
                           each=3),
                           rep("Total",3)),
             value = c(Results_F*100,Results_E*100,
                       Results_E*100+Results_F*100))

table <- matrix(c(round(sum(mat1[,4]*mat1[,5]),2),
                  round(sum(mat2[,4]*mat2[,5]),2),
                  round(sum(c(Results_F+Results_E)),3),
                  round(sum(Results_F),3),
                  round(sum(Results_E),3)),
               ncol = 1)

row.names(table) <-
  c(paste0("CCF in ",Year1),
    paste0("CCF in ",Year2),

```

```

      "Total annualized change",
      "Fertility component",
      "Education-composition component")

colnames(table) <- "Korea"

kable(table,caption = "CCF decomposition")

```

Table 24: CCF decomposition

	Korea
CCF in 1940	3.850
CCF in 1950	2.650
Total annualized change	-0.119
Fertility component	-0.103
Education-composition component	-0.016

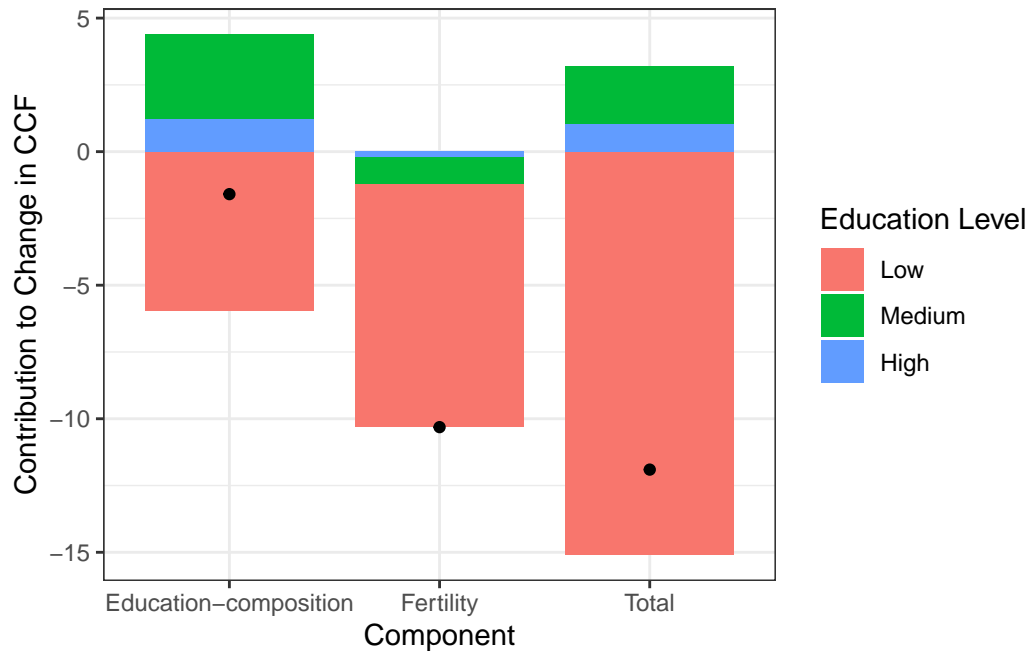
```

Results$Edu <- factor(Results$Edu,
                      levels = c("Low","Medium","High"))

ggplot(Results,aes(x=variable,y=value))+
  geom_col(aes(fill=Edu),width = 0.8)+
  stat_summary(geom = "point",fun = sum,show.legend = F)+
  labs(x = "Component",y = "Contribution to Change in CCF",
       fill = "Education Level")+
  theme_bw()

```





More details on the decomposition and how to write a function to do the decomposition calculation more easily can be found in the webpage of [Ryo Mogi](#). In his code he also has more information on making comparisons between populations.

## Cross-sectional Average Length of Life Childless (*CALC*)

### Method Explained

A life table is essentially the calculation of the duration of a population in the state of being alive before they die. Fertility, in some sense can also be seen as the transition between states, in this case from childlessness to motherhood. As such a life table can be applied to first births, following a cohort of young childless women, and letting their first birth be the movement out of the childless state, to the new state of motherhood. The resultant life expectancy is the “expected years without children (*EYWC*)”. Other uses of life tables can be applied to marriage, disability, and other life course events.

As suspected applying a life table perspective can be done to period or cohort data. However, an alternative option is to combine them as done in the previous study of the *CAL* perspective.

In this case the measure will be referred to as the “Cross-Sectional Average Length of Life Childless” (*CALC*) a measure introduced by [Mogi, Nisén, and Canudas-Romo](#)

(2021).  $CALC(t)$  is a period measure that includes age-specific first-birth histories of all female cohorts at reproductive ages 12–50 at a given time  $t$ . Mathematically it is written as:

$$CALC(t) = \int_{12}^{50} \ell_c(x, t-x) dx,$$

where  $\ell_c(x, t-x)$  is the probability of remaining childless from age 12 to the age attained in year  $t$ —that is,  $x$  for the cohort born in year  $t-x$ . As for the  $CAL(t)$  in mortality,  $CALC(t)$  includes the cohort fertility history for all women present at time  $t$ .

The decomposition of  $CALC$  follows closely that of  $CAL$ , comparing the cohort fertility histories between two populations. This is also the main finding in the paper by Mogi, Nisén, and Canudas-Romo (2021), and a perfect example of the application of mortality measures into fertility and vice-versa.

$$\begin{aligned} CALC(t, \xi) &= \int_{12}^{50} \dot{\ell}_c(x, t-x, \xi) dx \\ &= \int_{12}^{55} \ell_c(x, t-x, \xi) \left[ \sum_{a=12}^{x-1} \frac{\dot{p}_a(t-x, \xi)}{p_a(t-x, \xi)} \right] dx. \end{aligned}$$

This method has reproducible r code developed by Ryohei Mogi. Using these code to illustrate the method follows. The functions needed are included below.

```
# calculate CALC and decomposition
CALCDecompFunction <- function(px1, px2, lxLx,
                                Name1, Name2){

  CALClx <- c()
  CALClx1 <- c()
  CALClx2 <- c()
  PxCh <- c()

  PxCh <- log(px2 / px1)
  PxCh <- ifelse(is.na(PxCh), 0, PxCh)
  colnames(PxCh) <- rownames(PxCh) <- NULL

  # change the order: 1st column (the youngest cohort) ->
  # the last column (the oldest cohort)
```

```

PxCh <- PxCh[, ncol(PxCh):1]

px2CALlx <- function(px){
  # px matrix to lx
  lx <- apply(px, 2, cumprod)
  lx <- rbind(rep(1, ncol(px)), lx)

  # lx to CAL lx
  CALlx <- c()
  for(i in 1:ncol(px)){
    order <- 38:1
    CALlx[i] <- lx[order[i], i]
  }
  CALlx <- rev(CALlx)

  return(CALlx)
}

CALClx1 <- px2CALlx(px1)
CALClx2 <- px2CALlx(px2)

CALClx_mid <- t(matrix(rep((CALClx1 + CALClx2)/2, 38), length(CALClx1)))

# calculate CALC
CALC <- function(lx, type){

  # CALC using lx
  CALC_lx <- sum(lx[-1]) + 0.5

  # CALC using Lx
  Lx <- (lx[1:37] + lx[2:38]) / 2
  Lx <- c(Lx, lx[38])
  CALC_Lx <- sum(Lx)

  out <- ifelse(type == "lx", CALC_lx, CALC_Lx)

  return(out)
}

```

```

# final output
A1 <- CALC(lx = CALClx1, type = lxLx)
A2 <- CALC(lx = CALClx2, type = lxLx)
A3 <- sum(CALClx2) - sum(CALClx1)
A4 <- sum(PxCh * CALClx_mid)

print(rbind(c(paste("CALC-", Name1),
                 paste("CALC-", Name2),
                 "Diff", "est-Diff"),
            round(c(A1, A2, A3, A4), 2)))
return(PxCh * CALClx_mid)
}

```

## Examples

To illustrate the *CALC* comparisons between populations, Spain and Sweden are taken, their *CALC*s are calculated and the age- and cohort-contribution to their differences presented.

```

# We are comparing the two populations of Spain and Sweden.

Names <- c("SWE", "ESP")

Names2 <- c("Sweden", "Spain")

# The original code is written with tidyverse.

library(tidyverse)

# We are not considering the ages above 50.
over50 <- c("50", "51", "52", "53", "54", "55+")

# Here we are creating our own function that instead of
# matching contents (the overlap in the Venn diagram)
# we are finding the inverse. (the non-overlapping part)
`%out%` = Negate(`%in%`)

# We are using our own color palette

```

```

library(RColorBrewer)

mypalette <- rev(brewer.pal(8, "YlGnBu"))
mypalette2 <- rev(brewer.pal(8, "YlOrRd"))

WildColors <- c(mypalette[1:4], "white",
               "white", mypalette2[c(6, 4, 2, 1)])

Mycolor <- c(mypalette[1:4], "white", "white", "#FED976",
            "#FFEDA0", "#D4B9DA", "#FD8D3C")

# We are specifying the levels, here we are
# creating a logged scale manually. This illustrates the
# differences more vividly.

levels <- c(-1, -0.1, -0.01, -0.001, -0.0001, 0,
           .0001, .001, .01, .1, 1)

# This function is to accompany the previous
# logged scale and to make it show in the base r plot.
customAxis <- function() {
  n <- length(levels)
  y <- seq(min(levels), max(levels), length.out = n)
  rect(0, y[1:(n-1)], 1, y[2:n], col = WildColors)
  axis(4, at = y, labels = levels)
}

# A function to prepare the data, since for the fertility
# data we are combining completed cohort fertility and period
# fertility information from cohorts not yet reaching their
# fertility window.

lxxpx <- function(Names, lxxpx){

  # target life table function: lx or px
  target <- lxxpx

  ## Data from country 1
  A1 <- read.table(paste("data/FertDecomp/", Names,

```

```

      "cft.txt", sep = ""),
      header = TRUE, fill = TRUE, skip = 2)

## Function to create a matrix of lx or px
widelxpx <- function(data){
  px <- data %>%
    as.data.frame() %>%
    #filter(Cohort >= Y1) %>%
    # select year from the same year
    filter(x %out% over50) %>%
    # select age (12- to 49)
    mutate(q1x = as.numeric(as.character(q1x)),
           px = 1 - q1x) %>%
    select(Cohort, x, px)

  # create a matrix of px
  px_wide <- px %>%
    mutate(px = ifelse(x %in%
                        c("12-", "13", "14", "15") &
                        is.na(px), 1, px)) %>%
    spread(key = Cohort, value = px) %>%
    select(-x) %>%
    as.matrix()

  lx_wide <- matrix(NA,
                    ncol = ncol(px_wide),
                    nrow = nrow(px_wide))

  lx_wide[1, ] <- 1
  for(i in 1:(nrow(lx_wide)-1)){
    lx_wide[i+1, ] <- lx_wide[i, ] * px_wide[i, ]
  }

  colnames(lx_wide) <- colnames(px_wide)

  if(target == "lx"){
    outcome <- lx_wide
  } else {
    outcome <- px_wide
  }
}

```

```

    return(outcome)
  }

  ## For country 1
  lf_bc <- widelxpx(data = A1)

  if(any(colnames(lf_bc) == "1966")){
    lf_1966 <- lf_bc[, "1966"]
  } else {
    lf_1966 <- NA
  }

  ## the position of the maximum completed birth cohort
  if(length(lf_1966[!is.na(lf_1966)]) == 38){
    min1 <- which(colnames(lf_bc) == "1966")

    lf_bc <- lf_bc[, min1:ncol(lf_bc)]

    # extract data in a triangle format
    lf_triangle <- c()
    for(k in 1:ncol(lf_bc)){
      lf_triangle <- cbind(lf_triangle,
                           c(lf_bc[1:(38 - k + 1), k],
                             rep(NA, k - 1)))
    }

    colnames(lf_triangle) <- colnames(lf_bc)
  } else {
    min1 <- lf_bc[nrow(lf_bc), ]
    min1 <- length(min1[!is.na(min1)])

    ## select data from the maximum completed birth cohort
    lf_triangle <- lf_bc[, min1:ncol(lf_bc)]
  }

  ### Prepare period data to create hypothetical data

```

```

# create Age:Year matrix contains lx or px
data_select <- function(Pdata, lf_bc){

  startY <-
    as.numeric(as.character(colnames(lf_bc)))[ncol(lf_bc)] + 12
  lastBC <- lf_bc[, ncol(lf_bc)]
  endY    <- startY + length(lastBC[!is.na(lastBC)]) - 1

  px_wide <- Pdata %>%
    as.data.frame() %>%
    filter(x %out% over50) %>%
    mutate(q1x = as.numeric(as.character(q1x)),
           px = 1 - q1x) %>%
    select(Year, x, px) %>%
    filter(Year >= startY & Year <= endY) %>%
    spread(key = Year, value = px) %>%
    select(-x) %>%
    as.matrix()

  lx_wide <- matrix(NA,
                    ncol = ncol(px_wide),
                    nrow = nrow(px_wide))
  lx_wide[1, ] <- 1
  for(i in 1:(nrow(lx_wide)-1)){
    lx_wide[i+1, ] <- lx_wide[i, ] * px_wide[i, ]
  }

  colnames(lx_wide) <- colnames(px_wide)

  if(target == "lx"){
    outcome <- lx_wide
  } else {
    outcome <- px_wide
  }

  return(outcome)
}

```

*## Data from country A using period fertility table*



```

A1_per <- read.table(paste("data/FertDecomp/", Names,
                          "pft.txt", sep = ""),
                    header = TRUE, fill = TRUE, skip = 2)
A1_per <- data_select(Pdata = A1_per,
                     lf_bc = lf_triangle)

# make new data strage: hypthetical cohort
period2cohort <- function(data){

  n <- ncol(data)

  bc <- c()
  for(i in 2:n){
    row <- c()
    row <- c(data[i, - c(1:(i - 1))], rep(NA, i-1))
    bc <- rbind(bc, row)
  }
  bc <- rbind(data[1,], bc)

  years <- as.numeric(as.character(colnames(bc)))
  colnames(bc) <- years - 12
  rownames(bc) <- NULL

  return(bc)
}

A1_hypbc <- period2cohort(data = A1_per)

## combine cohort data and hypothetical data
A1_hypbc <- rbind(A1_hypbc,
                  matrix(NA,
                        dim(lf_triangle)[1] -
                          dim(A1_hypbc)[1],
                          ncol(A1_hypbc)))
lf_A1_bc <- cbind(lf_triangle, A1_hypbc[, -1])

return(lf_A1_bc)
}

```

```
px1 <- lxpx(Names = Names[1], lxpx = "px")
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `q1x = as.numeric(as.character(q1x))`.
Caused by warning:
! NAs introduced by coercion
There was 1 warning in `mutate()`.
i In argument: `q1x = as.numeric(as.character(q1x))`.
Caused by warning:
! NAs introduced by coercion
```

```
px2 <- lxpx(Names = Names[2], lxpx = "px")
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `q1x = as.numeric(as.character(q1x))`.
Caused by warning:
! NAs introduced by coercion
```

```
# Calculate CALC and decomposition
```

```
CALlxDecompBC <- CALCDecompFunction(px1, px2,
                                     "Lx", Names[1], Names[2])
```

	[,1]	[,2]	[,3]	[,4]
[1,]	"CALC- SWE"	"CALC- ESP"	"Diff"	"est-Diff"
[2,]	"19.84"	"22.5"	"2.66"	"3"

```
# The correct assignment of contributions and the cummulative changes
```

```
CALlxD <- matrix(0, 38, 38)
```

```
CALlxDS <- CALlxD
```

```
Age <- c(12:49)
```

```
BC <- 1966:2003
```

```
for (y in 1:38){
```

```

for (x in 1:y){
  # Age-cohort difference
  CALLxD[x, (38 - y + x)] <- CALLxDecompBC[x, y]
  # cumulative age-cohort difference
  CALLxDS[x, (38 - y + x)] <- sum(CALLxDecompBC[(1:x), y])
}
}

options(scipen = 10)

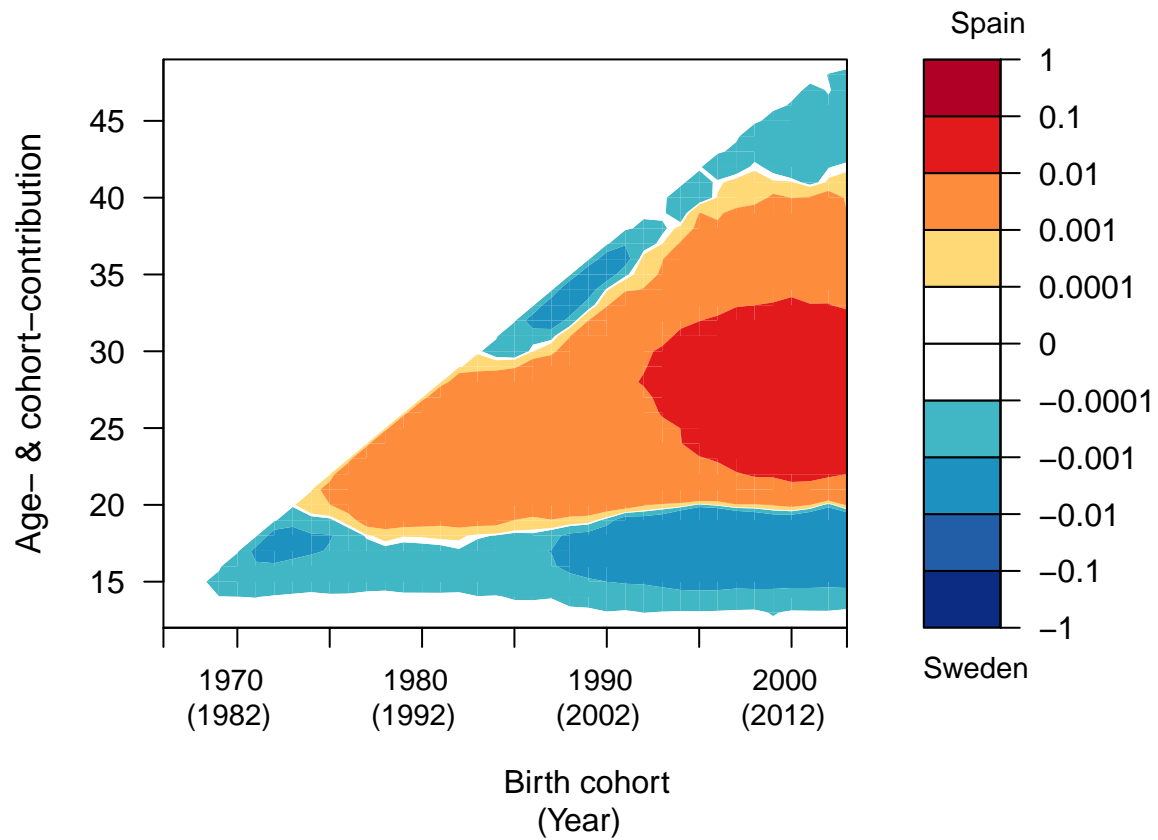
par(cex.axis = 1)
par(oma = c(1, 0, 0, 0))
filled.contour(BC, Age, t(CALLxD),
  levels = levels,
  col = WildColors,
  key.axes = customAxis(),
  ylab = "Age- & cohort-contribution",
  xlab = "",
  cex.lab = 1.1,
  plot.axes = {
    axis(1,
      at = c(1966, seq(1970, 2000, by = 5), 2003),
      labels = c("", "1970\n(1982)",
        "1975\n(1987)", "1980\n(1992)",
        "1985\n(1997)", "1990\n(2002)",
        "1995\n(2007)", "2000\n(2012)",
        "2003\n(2015)"),
      hadj = 0.6, padj = 0.5,
      cex.axis = 0.9)
    axis(2, at = seq(15, 50, by = 5),
      labels = seq(15, 50, by = 5))})

mtext("Birth cohort\n(Year)", 1, line = 4.5, adj = 0.4)

mtext(Names2[2], 3, 0.5, adj = 0.9, cex = 0.9)

mtext(Names2[1], 1, 0.5, adj = 0.9, cex = 0.9)

```



We can also see the cumulative version of this figure. Which is the exact same figure as the one that appeared in the manuscript.

```
options(scipen = 10)

par(cex.axis = 1)
par(oma = c(1, 0, 0, 0)) #bottom, right, top, left
filled.contour(BC, Age, t(CALLxDS),
               levels = levels,
               col = WildColors,
               key.axes = customAxis(),
               ylab = "Cumulative age- & cohort-contribution",
               xlab = "", cex.lab = 1.1,
```

```

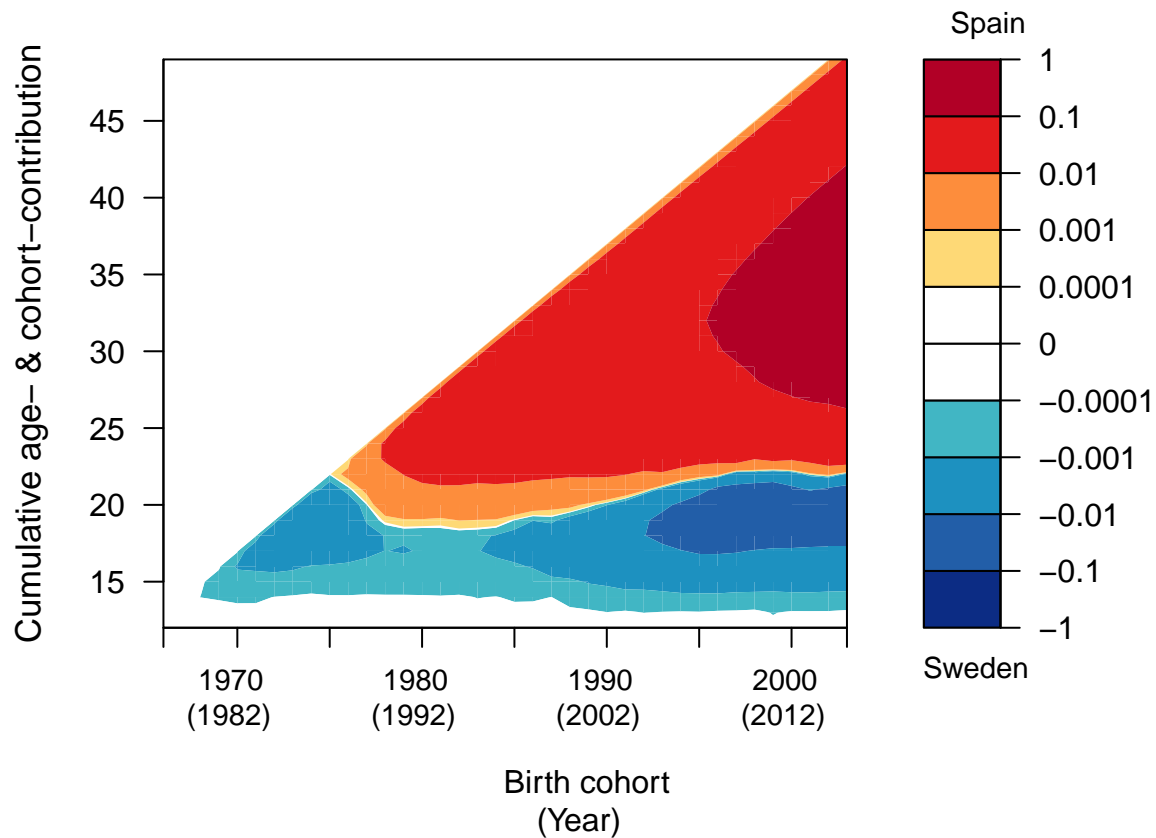
plot.axes = {
  axis(1, at = c(1966,
                 seq(1970, 2000, by = 5),
                 2003),
       labels = c("", "1970\n(1982)",
                  "1975\n(1987)", "1980\n(1992)",
                  "1985\n(1997)", "1990\n(2002)",
                  "1995\n(2007)", "2000\n(2012)",
                  "2003\n(2015)"),
       hadj = 0.6, padj = 0.5, cex.axis = 0.9)
  axis(2, at = seq(15, 50, by = 5),
       labels = seq(15, 50, by = 5)))

mtext("Birth cohort\n(Year)", 1, line = 4.5, adj = 0.4)

mtext(Names2[2], 3, 0.5, adj = 0.9, cex = 0.9)

mtext(Names2[1], 1, 0.5, adj = 0.9, cex = 0.9)

```



## Net-Reproduction Rate ( $NRR$ )

### Method Explained

Until now we have seen fertility measures that use exclusively fertility information: such as  $CCF$  (or its period equivalent  $TFR$ ), and measures that use the life table methodology such as the “expected years without children”, or its equivalent that incorporates cohort history the “cross-sectional average length of life childless” ( $CALC$ ). However, there are other fertility measures that incorporate both, the fertility information and the mortality information of women from birth to the upper limit of their reproductive years (Age 50 or 55). One example of such measure and which we will use here is the “Net-Reproduction Rate” ( $NRR$ ).

The  $NRR$  is defined as the average number of daughters that would be born over a woman's lifetime according to the age-specific rates of fertility and mortality in a given period. The measure focuses on the the number of female births that would 'replace' the women (including those who do not survive). Mathematically, this measure is expressed as:

$$NRR(t) = \int_{\alpha}^{\beta} \ell(a, t) m(a, t) da,$$

where  $m(a, t)$  is the maternity rate for female births at age  $a$  and  $\ell(a, t)$  is the probability of a woman surviving from age zero to age  $a$ , both at time  $t$ . The age-specific maternity rate is equal to  $m(a, t) = s(a, t)f(a, t)$ , where the proportion of female births to women at age  $a$  is denoted as  $s(a, t)$ , and  $f(a, t)$  is the age-specific fertility rates. The data for  $s(a, t)$  are not usually available, so we assume that it is fixed at  $\frac{1}{2.05}$ . The notation  $\alpha$  and  $\beta$  refers to the lower and upper bounds of the fertility ages.

The decomposition of the  $NRR$  was developed by [Shen, Lazzari, and Canudas-Romo \(2023\)](#) and it is written as follows:

$$N\dot{R}R(t) = \int_{\alpha}^{\beta} \dot{m}(a, t)\ell(a, t) + m(a, t)\dot{\ell}(a, t) da. \quad (12)$$

The first term within the integral accounts for changes in fertility, while the second term within the integral refers to change in mortality. Combining both components, fertility and mortality changes, helps explain the  $NRR$  changes. The components can be further decomposed. The fertility component can be disaggregated as

$$\dot{m}(a, t) = [\dot{s}(a, t) * f(a, t) + s(a, t) * \dot{f}(a, t)].$$

The mortality component can be further expressed using the age-specific probabilities of surviving  $p_x$  used in the  $CAL$  calculations, as:

$$\int_{\alpha}^{\beta} m(a, t)\dot{\ell}(a, t) da = \sum_{x=0}^{\beta-1} \frac{\dot{p}_x}{p_x} \int_{x+1}^{\beta} m(a, t)\ell(a, t) da$$

## Examples

The data for France from both HMD and HFD will be used in this example. The codes are adapted from the paper by Shen, Lazzari, and Canudas-Romo (2023).

```
library(tidyverse)

# data cleaning ----

Birth <- read.table("data/FertDecomp/FRATNPbirthsRR.txt",
                    header = TRUE, fill = TRUE, skip = 2)

Birth$Age[Birth$Age=="12-"] <- "12"
Birth$Age[Birth$Age=="55+"] <- "55"

Birth$Age <- as.numeric(Birth$Age)

Pop <- read.table("data/FertDecomp/FRATNPexposRR.txt",
                  header = TRUE, fill = TRUE, skip = 2)

Birth <- left_join(Birth, Pop)
```

Joining with `by = join\_by(Year, Age)`

```
Tot.birth = read.table("data/FertDecomp/Births.txt",
                       header = TRUE, fill = TRUE, skip = 1)

Tot.birth = Tot.birth %>% mutate(F.per = Female/Total)

lx = read.table("data/FertDecomp/fltper_1x1.txt",
                header = TRUE, fill = TRUE, skip = 1)[,c(1,2,6)]
lx$Age = gsub("[+]", "", lx$Age)
lx$Age = as.numeric(lx$Age)

Birth.C = Birth

Birth.C = inner_join(Birth.C, Tot.birth[,c(1,5)])
```

Joining with `by = join\_by(Year)`



```
Birth.C = Birth.C %>% mutate(Female = Total * F.per)

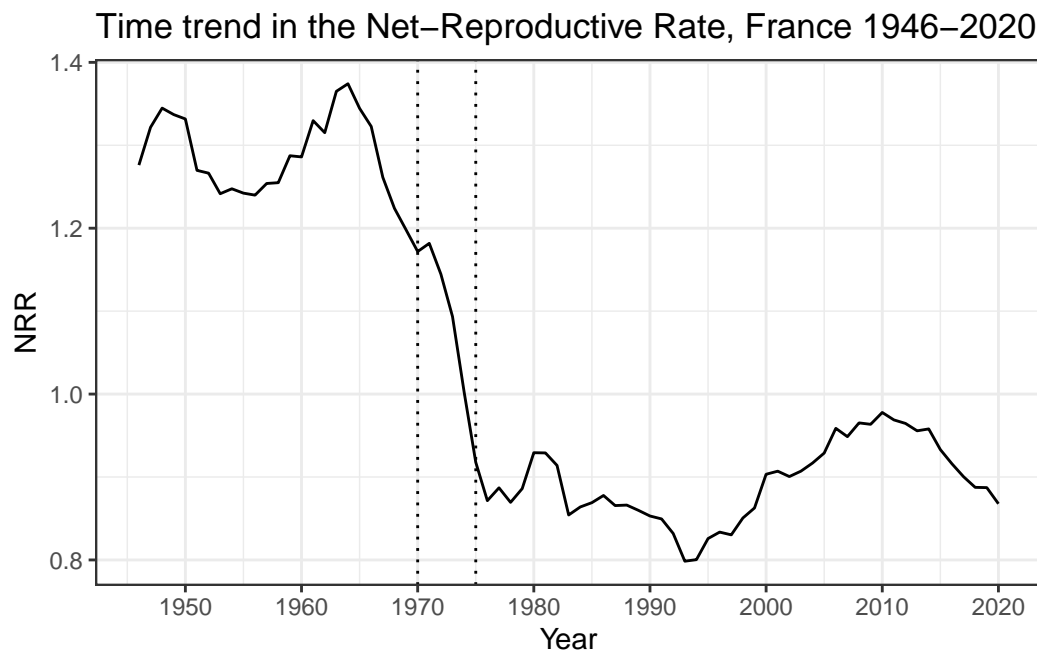
Birth.C = Birth.C %>% mutate(ASFR = Total/Exposure)
Birth.C = inner_join(Birth.C,lx)
```

Joining with `by = join\_by(Year, Age)`

```
Birth.C = Birth.C %>% mutate(lx = lx/100000) %>%
  mutate(ASNRR = ASFR * lx * F.per)

NRR <- Birth.C %>%group_by(Year) %>% summarise(NRR=sum(ASNRR))

ggplot(NRR,aes(x=Year,y=NRR))+
  geom_line()+
  scale_x_continuous(n.breaks = 10)+
  geom_vline(xintercept = c(1970,1975),linetype=3)+
  theme_bw()+
  labs(x="Year",y="NRR",title = "Time trend in the Net-Reproductive Rate, France 1946–2020")
```



During the period 1970-1975 the NRR crossed the value of one, considered the level of

replacement for a population. To disentangle this period of change we apply the method of the *NRR* decomposition.

```
# main decomposition ----

Year1 <- 1970

Year2 <- 1975

r_nrr <-
  log(NRR$NRR[which(NRR$Year==Year2)]/
      NRR$NRR[which(NRR$Year==Year1)])/(Year2-Year1)

d_nrr <-
  sqrt(NRR$NRR[which(NRR$Year==Year2)]*
      NRR$NRR[which(NRR$Year==Year1)])*r_nrr

r_tfr <-
  log(Birth.C$ASFR[which(Birth.C$Year==Year2)]/
      Birth.C$ASFR[which(Birth.C$Year==Year1)])/(Year2-Year1)

r_tfr[is.na(r_tfr)] <- 0
r_tfr[is.nan(r_tfr)] <- 0
r_tfr[is.infinite(r_tfr)] <- 0

r_lx <-
  log(Birth.C$lx[which(Birth.C$Year==Year2)]/
      Birth.C$lx[which(Birth.C$Year==Year1)])/(Year2-Year1)

r_per <-
  log(Birth.C$F.per[which(Birth.C$Year==Year2)]/
      Birth.C$F.per[which(Birth.C$Year==Year1)])/(Year2-Year1)

m_tfr <-
  sqrt(Birth.C$ASFR[which(Birth.C$Year==Year1)]*
      Birth.C$ASFR[which(Birth.C$Year==Year2)])

m_lx <- sqrt(Birth.C$lx[which(Birth.C$Year==Year1)]*
             Birth.C$lx[which(Birth.C$Year==Year2)])
```

```

m_per <- sqrt(Birth.C$F.per[which(Birth.C$Year==Year1)]*
              Birth.C$F.per[which(Birth.C$Year==Year2)])

d_tfr <- sum(m_tfr*m_lx*m_per*r_tfr)
d_lx <- sum(m_tfr*m_lx*m_per*r_lx)
d_per <- d_nrr-d_tfr-d_lx

table <-
  matrix(round(
    c(NRR[NRR$Year==Year1,]$NRR,
      NRR[NRR$Year==Year2,]$NRR,
      d_nrr,d_lx,d_tfr+d_per),3),ncol = 1)

row.names(table) <-
  c(paste0("NRR in ",Year1),
    paste0("NRR in ",Year2),
    "Annualized change of NRR",
    "Mortality Component",
    "Fertility Component")

colnames(table) <- "France"

kable(table,caption = "NRR decomposition")

```

Table 25: NRR decomposition

	France
NRR in 1970	1.172
NRR in 1975	0.918
Annualized change of NRR	-0.051
Mortality Component	0.001
Fertility Component	-0.052

## Assignment 4

Select one of the suggestions below and submit in ONE page: one Figure (or Table) and a brief paragraph describing the results that you find.

- Apply the measures in the examples of this section to a different population from HMD or HFD.
- The decomposition of the change in  $NRR$  for France showed that the mortality component was small in comparison to the fertility component. However, mortality component might be more significant in developing countries. With the data from “World Population Prospect, 2019”, select a country and compare two time points. Describe what has happened to the mortality component and how it contributes to the changes in  $NRR$ .

## Day 5. Decomposition for Alternative Measures

### Life years lost (LYL)

#### Methods Explained

Life expectancy is the average number of years a group of people is expected to live subjected to a set of age-specific death rates. To calculate life expectancy the mortality of a real cohort, or a synthetic cohort, is followed until the last age attained in the population. The radix of the life table, or initial cohort is usually set to  $\ell(0, t) = 1$ . So over age when individuals died the number surviving reduces and the accumulated mortality increases. Mathematically, we can write that for each age  $x$  the following holds:

$$\ell(x, t) + \lambda(x, t) = 1,$$

where  $\lambda(x, t)$  is the cumulative deaths  $\lambda(x, t) = \int_0^x f(a, t) da$ , and as before  $f(a, t)$  is the life tables death distribution. Integrating both sides from zero to age  $y$  gives us the equation:

$${}_xe_0(t) + {}_xLYL_0(t) = x.$$

The notation  ${}_xe_0(t)$  is the life expectancy between ages zero and  $x$ , while the notation  ${}_xLYL_0(t)$  is referred by [Andersen, Canudas-Romo, and Keiding \(2013\)](#) as “life years lost”. This equation shows that life years lost are the complement of the life expectancy up to a certain age  $x$  (which can be between 0 to  $\omega$ ). Life years lost  $LYL$  can be further partitioned by causes  $i$ , as

$$\begin{aligned} {}_xe_0(t) + \sum_i {}_xLYL_0^i(t) &= x \\ {}_xe_0(t) + \sum_i \int_0^x f_i(a, t) da &= x \end{aligned}$$

The notation  $f_i(a, t)$  represents the life table distribution of deaths from specific cause  $i$ . To apply this method first we will look at the functions that will be used and then an example using French data, and where the user can select a year (1950 to 2014), sex (females or males), and the age of the interval under study (0 to 100). The visualization of the death distribution is particularly useful to assess the levels of mortality in a population and the excess mortality (or lack of it) by specific causes of death.

```

# This function is to help calculating the
# cause specific factor R_i with counts of
# death by each cause

RxiMatrix<-function(PP,Cum){
  #PP<-B
  NumC<-dim(PP)[2]
  NumR<-dim(PP)[1]

  G<-PP
  FD<-colSums(PP)

  if (Cum==1){
    G<-t(apply(G,1,cumsum))
    FD<-cumsum(FD)
  }

  FRx3<-t(matrix(rep(FD,(NumR)),NumC))
  FRx<-G/FRx3
  FRx[is.infinite(FRx)]<-0
  return(FRx)
}

# A set of life table functions

Calculate_a0 <- function(m0,sex) {
  #Andreev-Kingkade formulas for computing a0 given m0
  # HMD full protocol Table 1 pg37
  #Males
  if(sex=="m"){
    if (m0<0.02300) {a0<-0.14929-(1.99545*m0)}
    if ((0.0230<= m0)&(m0<0.08307)) {a0<-0.02832+(3.26021*m0)}
    if (0.08307<= m0) {a0<-0.29915}
  }
  if (sex=="f"){
    #Females
    if (m0<0.01724) {a0<-0.14903-(2.05527*m0)}
    if ((0.01724 <= m0)&(m0< 0.06891)) {a0<-0.04667+(3.88089*m0)}
    if (0.06891<= m0) {a0<-0.31411}
  }
}

```

```

    return(a0) }

lifetable.mx<-function(mx,sex){

  N<-length(mx)
  AgeI<-rep(1,N)
  a0<-0.5
  if (AGEF==0){a0<-Calculate_a0(mx[1],sex)}
  ax<-c(a0,rep(0.5,(N-1)))
  if (mx[N]>0){ax[N]<-1/mx[N]}
  qx<-mx/(1+(1-ax)*mx)
  qx[N]<-1

  px<-1-qx

  lx<-100000

  for(y in 1:(N-1)){
    lx[y+1]<-lx[y]*px[y]
  }

  dx<-lx*qx
  dx[N]<-lx[N]

  Lx<-lx+(ax-AgeI)*dx
  Lx[N]<-lx[N]*ax[N]

  Tx<-c()
  for(y in 1:N){
    Tx[y]<-sum(Lx[y:N])
  }

  ex<-Tx/lx
  Age<-AGEF:(AGEF+N-1)
  AgeI<-rep(1,N)
  ALL<-cbind(Age, AgeI, ax, mx, qx, lx, dx, Lx, Tx, ex)
  return(ALL)
}

```

```

# Calculating life years lost based on the
# life table we have just calculated

LostYears2<-function(FLT,B){

  N<-dim(FLT)[1]
  AgeI<-rep(1,N)
  lx<-as.numeric(FLT[,6])/100000
  dx<-c(lx[-N]-lx[-1],lx[N])
  ax<-as.numeric(FLT[,5])
  Lx<-AgeI*lx+(ax-AgeI)*dx
  Lx[N]<-lx[N]*ax[N]

  Rxi<-RxiMatrix(B,0)
  fxi<-RxiMatrix(B,1)

  Nrow<-dim(Rxi)[1]
  Ncol<-dim(Rxi)[2]

  ## we use the life table functions to separate the
  ## person years and person lost

  LYL<-AgeI-Lx

  # age-specific LYLs
  LYLi2<-c()
  LYLi3<-matrix(0,Nrow,Ncol)
  for (y in 1:(Ncol-1)){
    L<-matrix(0,Nrow,Ncol)
    L[,y]<-(AgeI[y]-ax[y])*dx[y]*Rxi[,y]
    L[, (y+1):Ncol]<-
      matrix(rep(AgeI[(y+1):Ncol],each=Nrow),Nrow)*
      matrix(rep(dx[y]*Rxi[,y],length((y+1):Ncol)),Nrow)
    LYLi2<- cbind(LYLi2,rowSums(L))
    LYLi3<- (LYLi3+L)
  }
  y<-Ncol
  LYLi2<- cbind(LYLi2,(AgeI[y]-ax[y])*dx[y]*Rxi[,y])
  LYLi3[,y]<-LYLi3[,y]+LYLi2[,y]

```



```

    return(LYLi2)
}

# This function returns the life years lost
# visualization in the terms of the complement of
# the survival curve in a life table.

NicePlot2<-function(FLT,B,S){

  causesN0<-Causes
  causesN<-causesN0
  NcausesT<-length(Causes)
  Ncauses<-length(S)
  Dif<-length(Causes)-Ncauses

  #Selected causes
  if(Dif>0){causesN<- c(causesN0[S],"Other causes")}
  if(Dif>0){Ncauses<-(length(S)+1)}

  # COL<-matlab.like2(NcausesT)
  COL<-primary.colors(NcausesT)

  N<-dim(FLT)[1]
  AgeI<-rep(1,N)
  lx<-as.numeric(FLT[,6])/100000
  dx<-c(lx[-N]-lx[-1],lx[N])

  FRx<-RxiMatrix(B,0)
  Fdxi<-FRx*matrix(rep(dx,each=NcausesT),NcausesT)
  dxi0<-apply(Fdxi,1,cumsum)

  ## now we make the plot of the survival
  ## function and the causes of death
  ## contributing to the lost years

  ## first just the lines
  dxi <- dxi0
  if(Dif>0){ dxi <- cbind(dxi0[,S],rowSums(dxi0[, -S]))}

```

```

dxi.1 <- rep(0,dim(dxi)[2])
dxi.f <- rbind(dxi.1, dxi[-N,])

test.p0 <- cbind(lx, dxi.f)
test.p1 <- cbind(test.p0[1:length(Age),], Age )
test.p1[is.na(test.p1)]<-0
test.p2 <- as.data.frame(test.p1)
names(test.p2)

test.p <- reshape2::melt(test.p2, id.vars = c("lx", "Age"))

# Alt colors
ggplot(test.p, aes(x = Age, y = value,
                  group = variable, fill = variable)) +
  geom_area(size = 0.3, color = "black",
           position = position_stack()) +
  scale_y_reverse(name = "Probability of surviving and life years lost",
                 breaks = c(0, .2, .4, .6, .8, 1),
                 labels = c(1, .8, .6, .4, .2, 0),
                 limits = c(1,0)) +
  scale_x_continuous(name = "Age",
                    limits = c(AGEF,AGEL)) +
  scale_fill_manual(name = "", values = COL,
                  labels = causesN) +
  theme_bw(base_size = 10) +
  theme(legend.position = "bottom") +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
}

```

## Examples

The following code produces a visualization of the life years lost, as the area above the life table survival function. Furthermore, the life years lost are separated by colors denoting the difference causes of death.

```

library(colorRamps)
library(reshape2)

```

```

#sex
SEX<-c("Male","Female")
Sex<-2

#Ages
AGEF<-0
AGEL<-90 #Age Limit

# population, we will use France
Names<-c("FRATNP")
Names2<-c("France")

#year
Year<-1950

#country codes
CO<-c(4080)

Causes<-c("Infectious","Neoplasms","CVD",
          "Not classified","Mental","Nervous",
          "Diabetes","Digestive",
          "Genitourinary","Congenital","Respiratory",
          "External")
## causes of death + 1 (since the first place is for the total deaths)
## it is done one by one

CONTIf<-c()
CONTIm<-c()

Af<-read.table("data/LYL/fltper_1x1.txt",
               header=TRUE,fill=TRUE,skip=1)
Am<-read.table("data/LYL/mltper_1x1.txt",
               header=TRUE,fill=TRUE,skip=1)
Af<-Af[Af$Year==Year,]
Am<-Am[Am$Year==Year,]

for (CC in 2:13){

B1C<-t(read.csv(paste("data/LYL/", "Country",CO[1], "Cause",

```

```

      CC,"f",".txt",sep=""),
      header=TRUE,fill=TRUE,skip=0)[-1])
B1Cm<-t(read.csv(paste("data/LYL/", "Country", CO[1], "Cause",
      CC,"m",".txt",sep=""),
      header=TRUE,fill=TRUE,skip=0)[-1])

CONTIm<-rbind(CONTIm,B1Cm[B1Cm[,1]==Year,])
CONTIf<-rbind(CONTIf,B1C[B1C[,1]==Year,])

}

CD <- c()
A1 <- c()

if (Sex==1){(A1<-Am)}
if (Sex==1){(sex<-c("m"))}
if (Sex==1){(CD<-CONTIm)}
if (Sex==2){(A1<-Af)}
if (Sex==2){(sex<-c("f"))}
if (Sex==2){(CD<-CONTIf)}

LT<-A1[c(AGEF:AGEL)+1,]

WW<-CD[,c(AGEF:AGEL)+2]

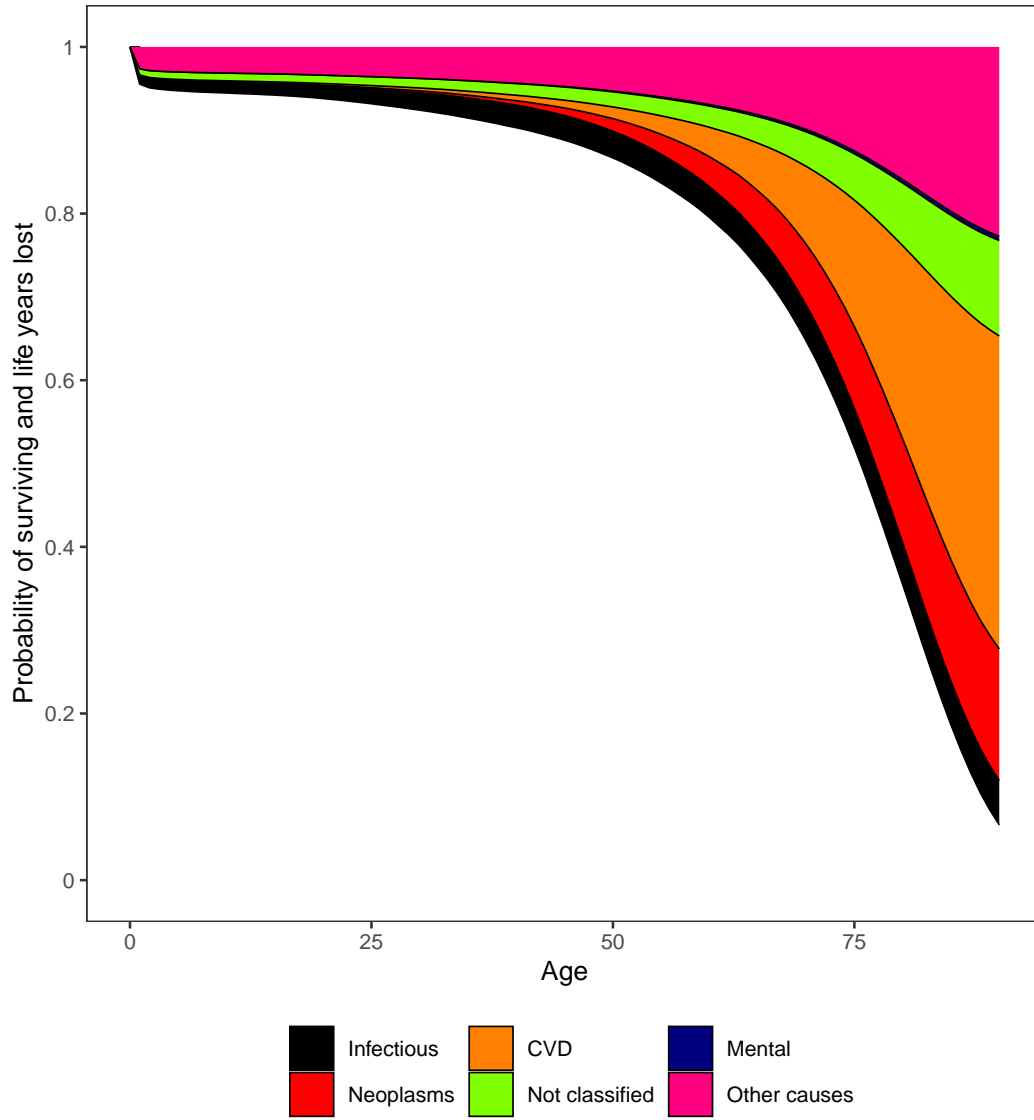
Age <- c(AGEF:AGEL)

S<-c(1:5)

NicePlot2(LT,WW,S)

```

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
 i Please use `linewidth` instead.



## Variable-r

### Methods Explained

The variable-r method comes from generalizing the Lotka stable population model, changing from one single population growth rate to age-specific growth rates (Arthur and Vaupel 1984; S. H. Preston and Coale 1982). The population growth at the national level  $\bar{r}(t)$  is the weighted average of the age-specific growth rate  $r_x(t)$  multiplied by the

population structure ( $c_x(t)$ ) at age  $x$  and time  $t$ , which writes  $\bar{r}(t) = \int_0^\omega r_x(t)c_x(t)dx$  (Horiuchi and Preston 1988; Canudas-Romo, Shen, and Payne 2022). From each of the age-specific growth rates we can further disentangle the cohort history that involves the birth, survival, and migration history of the population (Horiuchi and Preston 1988; S. H. Preston and Stokes 2012; Canudas-Romo, Shen, and Payne 2022).

In this model any counts of individuals at a given age  $x$  and time  $t$ , can be traced back to its original cohort by taking into account their births, mortality and net-migration, as

$$P(x, t) = B(t - x)\ell^c(x, t - x)n(x, t - x),$$

where the population  $P(x, t)$  is the product of new-born babies  $B(t - x)$  in year  $t - x$ , the cohort survival probability  $\ell^c(x, t - x)$  between birth age  $x$  which is attained in year  $t$ , and a factor of net-migration  $n(x, t - x)$  also between birth and age  $x$ . In reality, since we do not have good enough data in most settings for cohort net-migration the term is calculated inversely from the other components as

$$n(x, t - x) = \frac{P(x, t)}{B(t - x)\ell^c(x, t - x)}.$$

This use of the variable-r method can be applied to study changes over time, where then age-specific growth rates play an important role. But it can also be applied to study the ratio between populations at the same age and time. For instance, the sex ratio at each age, defined as the ratio of males over females  $s(x, t) = \frac{P_M(x, t)}{P_F(x, t)}$  can be represented as:

$$s(x, t) = R_B(t - x)R_{\ell^c}(x, t)R_n(x, t),$$

where the sex ratio at each age is the product between the sex ratio at birth or  $R_B(t - x) = \frac{B_M(t - x)}{B_F(t - x)}$ , the survival ratio or  $R_{\ell^c}(t - x) = \frac{\ell_M^c(t - x)}{\ell_F^c(t - x)}$  and the net-migration ratio or  $R_n(t - x) = \frac{n_M(t - x)}{n_F(t - x)}$ . The notation  $M$  stand for male and  $F$  stands for female.

## Examples

We will illustrate this with Swedish data from HMD in 2020. We are going to use a new package called “data.table”. This package uses syntax similar to base r but significantly faster, especially dealing with huge amount of data. The annotated codes are shown below.

```

Cnty <- "SWE"

Year <- 2020

Age <- 100

library(data.table)

decomp_data <- function(Cnty,Year,Age){

  country <- Cnty
  year <- Year
  age <- Age

  # First step is read data. We will need birth, mortality,
  # and Population data from HMD. We will be using lexis deaths
  # since we want to get a more accurate cohort mortality estimate.

  # Population
  Pop <- fread(paste0("data/Variabler/",
                      country,".Population.txt"),
              header = T,skip=1)
  colnames(Pop) <- c("Year","Age","Female","Male","Total")
  Pop$Age[Pop$Age=="110+"]<-110
  Pop$Age <- as.numeric(Pop$Age)
  Pop$Year = gsub("[-]", "", Pop$Year)
  Pop$Year <- as.numeric(Pop$Year)

  # Birth
  Birth <- fread(paste0("data/Variabler/",
                        country,".Births.txt"),
                header = T,skip=1)
  Pop$Year <- as.numeric(Pop$Year)

  # Death
  Death <- fread(paste0("data/Variabler/",
                        country,".Deaths_lexis.txt"),
                header = T,skip=1)
  Death$Cohort[Death$Cohort=="."]<-0
  Death$Cohort <- as.numeric(Death$Cohort)

```

```

Death$Age[Death$Age=="110+"]<-110
Death$Age <- as.numeric(Death$Age)

# specify the vectors needed.

sex_ratio <- rep(0,age+1)

# For female and male population along
P_male <- rep(0,age+1)
P_female <- rep(0,age+1)
B_male <- rep(0,age+1)
B_female <- rep(0,age+1)
S_male <- rep(0,age+1)
S_female <- rep(0,age+1)
M_male <- rep(0,age+1)
M_female <- rep(0,age+1)

# for the ratio
rB <- rep(0,age+1)
rS <- rep(0,age+1)
rM <- rep(0,age+1)

# We are calculating the cohort survival per
# HMD protocol which is the parallelogram
# but first age 0 there is only a triangle.

P_male[1] <- Pop[Year==paste0(year)&Age==0,Male]
P_female[1] <- Pop[Year==paste0(year)&Age==0,Female]

sex_ratio[1] <- P_male[1]/P_female[1]

#### Birth

B_male[1] <- Birth[Year==year-1,Male]
B_female[1] <- Birth[Year==year-1,Female]

rB[1] <- B_male[1]/B_female[1]

#### Survival

```



```

# Survival is calculated as the probability
# of dying for people exposed.

#### male survival
qx_male <- Death[Cohort==year-1&
                  Age==0&
                  Year==year-1,][,c(1:3,5)]
qx_male <- qx_male[,sum(Male),by=.(Age,Cohort)]
# This is essentially the group_by and summarise function
# combined
names(qx_male)[3] <- "D"

qx_male$P <-
  Pop[Year==year&Age==0,Male]+qx_male$D

qx_male$Mx <- qx_male$D/qx_male$P

S_male[1] <- (1-qx_male$Mx)

#### female survival
qx_female <- Death[Cohort==year-1&
                   Age==0&
                   Year==year-1,][,c(1:3,4)]
qx_female <- qx_female[,sum(Female),by=.(Age,Cohort)]
names(qx_female)[3] <- "D"

qx_female$P <-
  Pop[Year==year&Age==0,Female]+qx_female$D

qx_female$Mx <- qx_female$D/qx_female$P

S_female[1] <- (1-qx_female$Mx)

#### Survival ratio
rS[1] <- S_male[1]/S_female[1]

#### Net-migration
M_male[1] <- P_male[1]/(B_male[1]*S_male[1])

```

```

M_female[1] <- P_female[1]/(B_female[1]*S_female[1])

rM[1] <- M_male[1]/M_female[1]

for (a in 1:age) {

  ##### Sex ratio

  P_male[a+1] <- Pop[Year==paste0(year)&Age==a, Male]
  P_female[a+1] <- Pop[Year==paste0(year)&Age==a, Female]

  sex_ratio[a+1] <- P_male[a+1]/P_female[a+1]

  ##### Birth

  B_male[a+1] <- Birth[Year==year-a-1, Male]
  B_female[a+1] <- Birth[Year==year-a-1, Female]

  rB[a+1] <- B_male[a+1]/B_female[a+1]

  ##### Survival

  ### male survival
  qx_male <- Death[Cohort==year-a-1&Age<=a,][,c(1:3,5)]
  qx_male <- qx_male[,sum(Male),by=.(Age,Cohort)]
  names(qx_male)[3] <- "D"

  for (c in (a-1):0) {
    qx_male$P[which(qx_male$Age==a-c-1)] <-
      Pop[Year==year-c-1&Age==a-c-1, Male]+
      Death[Year==year-c-2&Age==a-c-1, Male][1]
  }

  # The last age group is also a triangle so
  # we calculate them separately.

  last_qx_male <-
    Death[Cohort==year-a-1&
      Age==a&

```

```

      Year==year-1,][,c(1:3,5)]

names(last_qx_male)[4] <- "D"

last_qx_male$P <-
  Pop[Year==year&Age==a, Male]+last_qx_male$D

last_qx_male$Mx <- last_qx_male$D/last_qx_male$P

qx_male$Mx <- qx_male$D/(qx_male$P)

S_male[a+1] <-
  tail(cumprod(1-qx_male$Mx),1)*(1-last_qx_male$Mx)

### female survival
qx_female <- Death[Cohort==year-a-1&Age<a,][,c(1:3,4)]
qx_female <- qx_female[,sum(Female),by=(Age,Cohort)]
names(qx_female)[3] <- "D"

for (c in (a-1):0) {
  qx_female$P[which(qx_female$Age==a-c-1)] <-
    Pop[Year==year-c-1&Age==a-c-1, Female]+
    Death[Year==year-c-2&Age==a-c-1, Female][1]
}

last_qx_female <-
  Death[Cohort==year-a-1&
    Age==a&
    Year==year-1,][,c(1:3,4)]

names(last_qx_female)[4] <- "D"

last_qx_female$P <-
  Pop[Year==year&Age==a, Female]+last_qx_female$D

last_qx_female$Mx <- last_qx_female$D/last_qx_female$P

qx_female$Mx <- qx_female$D/(qx_female$P)

```

```

S_female[a+1] <-
  tail(cumprod(1-qx_female$Mx),1)*(1-last_qx_female$Mx)

### Survival ratio
rS[a+1] <- S_male[a+1]/S_female[a+1]

#### Net-migration
M_male[a+1] <- P_male[a+1]/
  (B_male[a+1]*S_male[a+1])

M_female[a+1] <- P_female[a+1]/
  (B_female[a+1]*S_female[a+1])

# rM[a+1] <- sex_ratio[a+1]/
#   (rS[a+1]*rB[a+1])

rM[a+1] <- M_male[a+1]/
  M_female[a+1]
}

table <- data.table(
  Age = 0:age,
  sex_ratio = sex_ratio,
  P_male = P_male, P_female = P_female,
  rB = rB, rS = rS, rM = rM,
  B_male = B_male, B_female = B_female,
  S_male = S_male, S_female = S_female,
  M_male = M_male, M_female = M_female
)

table <- table[,`:=`(Cx_f = P_female/sum(P_female),
  Cx_m = P_male/sum(P_male))]

return(table)
}

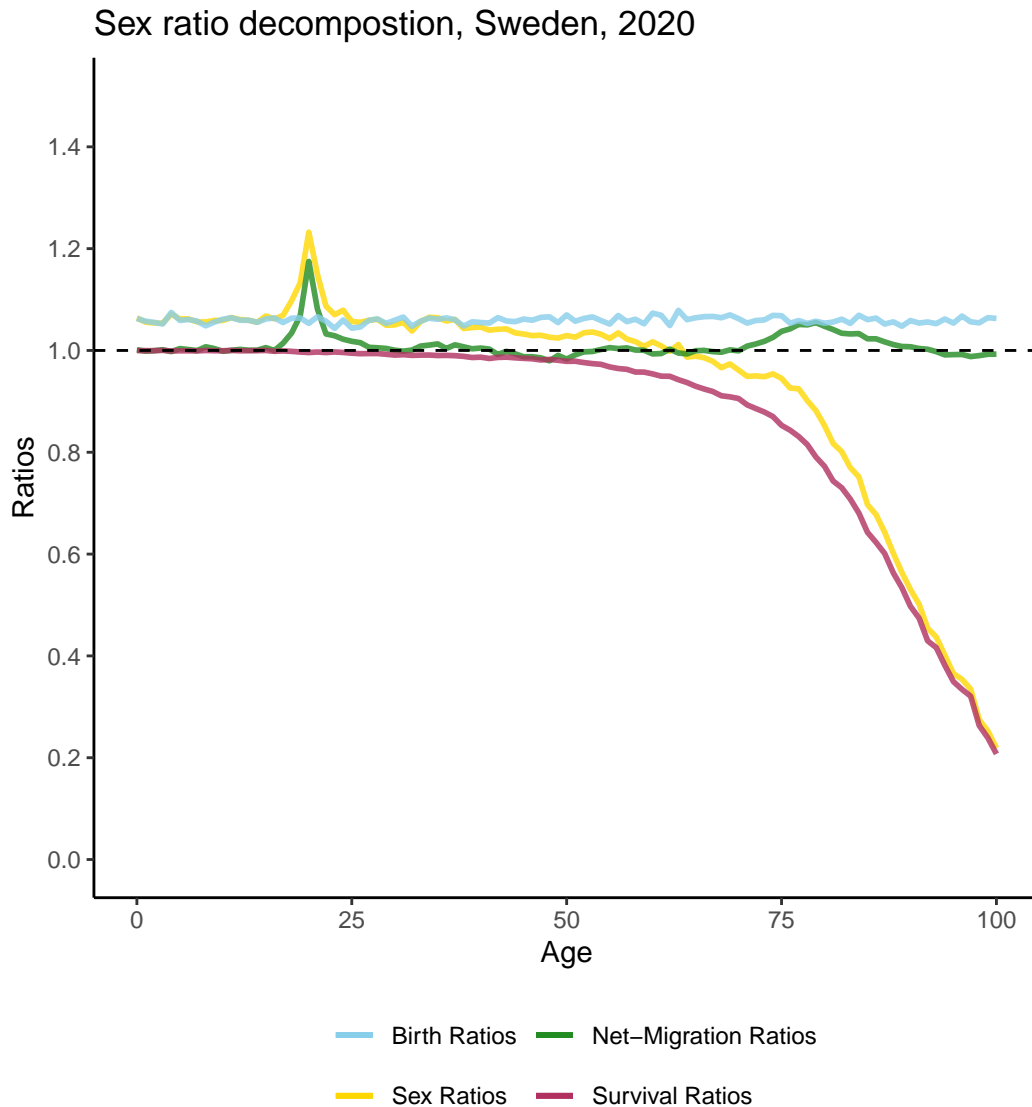
Data <- decomp_data(Cnty,Year,Age)

```

```

ggplot(Data,mapping=aes(x=Age))+
  geom_line(mapping=aes(y=sex_ratio,color="Sex Ratios"),
            linewidth=1,alpha=0.8)+
  geom_line(mapping=aes(y=rM,color="Net-Migration Ratios"),
            linewidth=1,alpha=0.8)+
  geom_line(mapping=aes(y=rS,color="Survival Ratios"),
            linewidth=1,alpha=0.8)+
  geom_line(mapping=aes(y=rB,color="Birth Ratios"),
            linewidth=1,alpha=0.8)+
  geom_hline(yintercept = 1,linetype=2)+
  scale_color_manual(values = c("skyblue","forestgreen",
                                "gold","maroon"))+
  scale_x_continuous(n.breaks=5,limits = c(0,100))+
  scale_y_continuous(n.breaks=10,limits = c(0,1.5))+
  guides(color=guide_legend(title="",
                             nrow = 2,byrow = T))+
  theme_classic()+
  theme(legend.position = "bottom")+
  labs(title = "Sex ratio decomposition, Sweden, 2020",
        x="Age",y="Ratios")

```



## Assignment 5

Select one of the suggestions below and submit in ONE page: one Figure (or Table) and a brief paragraph describing the results that you find.

- Apply the measures in the examples of this section to a different population from HMD or HFD.

# Assignment Answers

## Assignment 1 Answer

### Population Growth Rate ( $\bar{r}$ )

This can be written as:

$$\dot{\bar{r}} = \bar{r} + \sigma^2(r)$$

This equation is a famous one with many applications.

```
Year1 <- 2010

Year2 <- 2020

yeard <- 10

JPN_pop <- read.table("data/standard/JPN.Population.txt",
                      header = T, skip = 2)

JPN_gth1 <- log_func(JPN_pop[JPN_pop$Year==Year1,"Female"],
                     JPN_pop[JPN_pop$Year==(Year1-yeard),"Female"],
                     yeard)

JPN_gth2 <- log_func(JPN_pop[JPN_pop$Year==Year2,"Female"],
                     JPN_pop[JPN_pop$Year==(Year2-yeard),"Female"],
                     yeard)

JPN_pop1 <- (JPN_pop[JPN_pop$Year==Year1,"Female"]*
             JPN_pop[JPN_pop$Year==(Year1-yeard),"Female"])^0.5

JPN_pop2 <- (JPN_pop[JPN_pop$Year==Year2,"Female"]*
             JPN_pop[JPN_pop$Year==(Year2-yeard),"Female"])^0.5

Comp1 <- JPN_pop1/sum(JPN_pop1)

Comp2 <- JPN_pop2/sum(JPN_pop2)

Comp_avg <-
```

```

(JPN_pop2*JPN_pop1)^0.5/
sum((JPN_pop2*JPN_pop1)^0.5)

r_bar1 <- sum(JPN_gth1*Comp1)*100

r_bar2 <- sum(JPN_gth2*Comp2)*100

r_bar_diff <-
  (r_bar2-r_bar1)/(Year2-Year1)

r_bar_direct <-
  sum((JPN_gth2-JPN_gth1)/(Year2-Year1)*Comp_avg)*100

gth_avg <- log_func(JPN_pop2,JPN_pop1,Year2-Year1)

r_bar_comp <-
  (sum(gth_avg^2*Comp_avg)-
   sum((gth_avg*Comp_avg)^2))*100

r_bar_decomp <- r_bar_direct+r_bar_comp

table <- matrix(round(c(r_bar1,r_bar2,r_bar_diff,
                       r_bar_direct,r_bar_comp,
                       r_bar_decomp),2))

row.names(table) <- c("growth rate 2000-2010",
                     "growth rate 2010-2020",
                     "difference in growth rates",
                     "direct component",
                     "composition component",
                     "total estimated difference")

colnames(table) <- "Japan"

kable(table,caption = "Growth rate decompostion (VCR)")

```



Table 26: Growth rate decomposition (VCR)

	Japan
growth rate 2000-2010	0.12
growth rate 2010-2020	-0.19
difference in growth rates	-0.03
direct component	-0.06
composition component	0.03
total estimated difference	-0.03

Note: Results are multiplied by 100

## Assignment 2 Answer

### Crude Birth Rate (CBR)

Crude Birth rate and its decomposition using the VCR method can be written as:

$$\dot{\bar{b}} = \bar{b} + Cov(f_x, r_f)$$

```
Year1 <- 2010

Year2 <- 2020

KOR_pop <- read.table("data/standard/KOR.Population.txt",
                      header = T, skip = 2)

KOR_fpop1 <-
  (KOR_pop[KOR_pop$Year==Year1, "Female"] *
   KOR_pop[KOR_pop$Year==Year1+1, "Female"])^0.5

KOR_fpop2 <-
  (KOR_pop[KOR_pop$Year==Year2, "Female"] *
   KOR_pop[KOR_pop$Year==Year2+1, "Female"])^0.5

KOR_totpop2 <-
  (KOR_pop[KOR_pop$Year==Year2, "Total"] *
   KOR_pop[KOR_pop$Year==Year2+1, "Total"])^0.5

KOR_totpop1 <-
  (KOR_pop[KOR_pop$Year==Year1, "Total"] *
   KOR_pop[KOR_pop$Year==Year1+1, "Total"])^0.5

KOR_fx <- read.table("data/standard/KORasfrRR.txt",
                     header = T, skip = 2)

KOR_fx1 <- c(rep(0, 12),
             KOR_fx[KOR_fx$Year==Year1, "ASFR"],
             rep(0, 55))

KOR_fx2 <- c(rep(0, 12),
             KOR_fx[KOR_fx$Year==Year2, "ASFR"],
```

```

rep(0,55))

KOR_comp1 <- (KOR_fpop1)/sum(KOR_totpop1)

KOR_comp2 <- (KOR_fpop2)/sum(KOR_totpop2)

KOR_comp <-
  (KOR_fpop1*KOR_fpop2)^0.5/
  sum((KOR_totpop1*KOR_totpop2)^0.5)

CBR_Y1 <- sum(KOR_fx1*KOR_comp1)*1000

CBR_Y2 <- sum(KOR_fx2*KOR_comp2)*1000

CBR_diff <-
  log(CBR_Y2/CBR_Y1)/(Year2-Year1)*
  (CBR_Y1*CBR_Y2)^0.5

CBR_direct <-
  sum(log_func(KOR_fx2,KOR_fx1,Year2-Year1)*
      (KOR_fx2*KOR_fx1)^0.5*KOR_comp,na.rm = T)*1000

CBR_comp <-
  (sum((KOR_fx2*KOR_fx1)^0.5*
      log_func(KOR_totpop2,KOR_totpop1,
                Year2-Year1)*KOR_comp)-
   sum((KOR_fx2*KOR_fx1)^0.5*KOR_comp)*
   sum(log_func(KOR_totpop2,KOR_totpop1,
                Year2-Year1)*KOR_comp))*1000

CBR_decomp <- (CBR_direct+CBR_comp)

table <-
matrix(round(c(CBR_Y1,CBR_Y2,
              CBR_diff,CBR_direct,
              CBR_comp,CBR_decomp),2),
       ncol=1)

colnames(table) <- "Korea"

```

```

row.names(table) <- c(paste0("CBR in ",Year1),
                      paste0("CBR in ",Year2),
                      "CBR annualized difference",
                      "Direct component",
                      "Compositional component",
                      "Total estimated difference")

kable(table,caption = "CBR decomposition (VCR)")

```

Table 27: CBR decomposition (VCR)

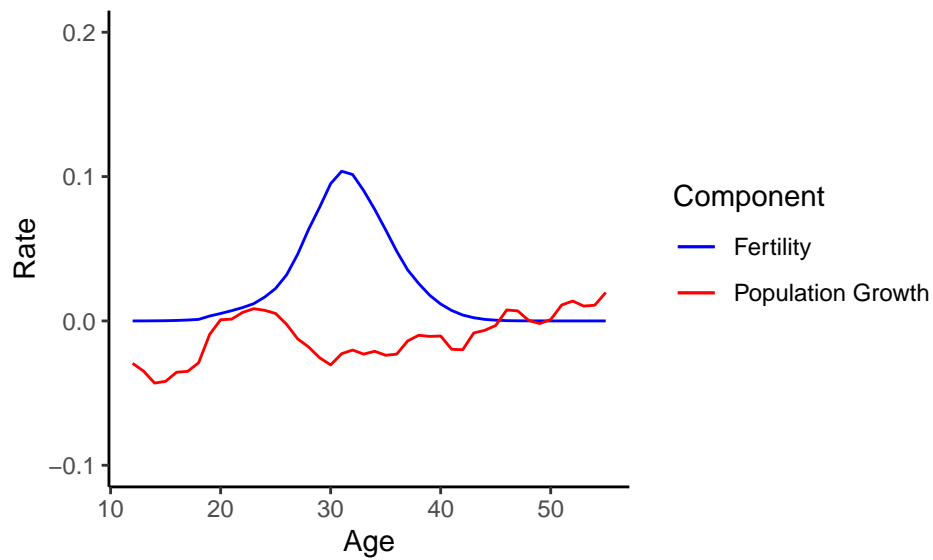
	Korea
CBR in 2010	9.41
CBR in 2020	5.31
CBR annualized difference	-0.40
Direct component	-0.25
Compositional component	-0.13
Total estimated difference	-0.38

Note: Results are multiplied by 1000

```

ggplot()+
  geom_line(aes(x=0:110,
                y=(KOR_fx2*KOR_fx1)^0.5,
                color = "Fertility"))+
  geom_line(aes(x=0:110,
                y=log_func(KOR_fpop2,KOR_fpop1,Year2-Year1),
                color = "Population Growth"))+
  scale_x_continuous(limits = c(12,55))+
  scale_y_continuous(limits = c(-0.1,0.2))+
  scale_color_manual(values = c("blue","red"))+
  labs(x="Age",y="Rate",color="Component")+
  theme_classic()

```



## Assignment 3 Answer

### Part 1

```
USA_Mx <- read.table("data/MortDecomp/USA.Mx_1x1.txt",
                     header = T, skip = 2, fill = T)

Mx1 <- USA_Mx[USA_Mx$Year==2000,3]

Mx2 <- USA_Mx[USA_Mx$Year==2010,3]

Mx3 <- USA_Mx[USA_Mx$Year==2018,3]

ex_diff1 <-
  life.table(Mx2,"f")$ex[1] -
  life.table(Mx1,"f")$ex[1]

ex_diff2 <-
  life.table(Mx3,"f")$ex[1] -
  life.table(Mx2,"f")$ex[1]

ex_decomp1 <-
```

```

arriaga(as.numeric(Mx1),
        as.numeric(Mx2),
        sex = "f",
        breakdown = F)

ex_decomp2 <-
arriaga(as.numeric(Mx2),
        as.numeric(Mx3),
        sex = "f",
        breakdown = F)

cause_count <- read.csv("data/MortDecomp/USA_Cause16_counts.csv")

cause_count1 <- cause_count[cause_count$Year==2000&
                           cause_count$Sex==2,]

cause_count1 <- matrix(c(cause_count1$Count),
                      ncol = length(unique(cause_count1$Cause)),
                      nrow = length(unique(cause_count1$Age)))

row.names(cause_count1) <- 0:110

colnames(cause_count1) <- 1:16

cause_total1 <- rowSums(cause_count1)

cause_prop1 <- apply(cause_count1, 2, function(x){x/cause_total1})

cause_count2 <- cause_count[cause_count$Year==2010&
                           cause_count$Sex==2,]

cause_count2 <- matrix(c(cause_count2$Count),
                      ncol = length(unique(cause_count2$Cause)),
                      nrow = length(unique(cause_count2$Age)))

row.names(cause_count2) <- 0:110

colnames(cause_count2) <- 1:16

```

```

cause_total2 <- rowSums(cause_count2)

cause_prop2 <- apply(cause_count2, 2, function(x){x/cause_total2})

cause_count3 <- cause_count[cause_count$Year==2010&
                             cause_count$Sex==2,]

cause_count3 <- matrix(c(cause_count3$Count),
                       ncol = length(unique(cause_count3$Cause)),
                       nrow = length(unique(cause_count3$Age)))

row.names(cause_count3) <- 0:110

colnames(cause_count3) <- 1:16

cause_total3 <- rowSums(cause_count3)

cause_prop3 <- apply(cause_count3, 2, function(x){x/cause_total3})

cause_fac1_1 <-
  (cause_prop1*Mx1)/
  ifelse((Mx2-Mx1)==0,1,Mx2-Mx1)*ex_decomp1

cause_fac2_1 <-
  (cause_prop2*Mx2)/
  ifelse((Mx2-Mx1)==0,1,Mx2-Mx1)*ex_decomp1

cause_fac2_2 <-
  (cause_prop2*Mx2)/
  ifelse((Mx3-Mx2)==0,1,Mx3-Mx2)*ex_decomp2

cause_fac3_2 <-
  (cause_prop3*Mx3)/
  ifelse((Mx3-Mx2)==0,1,Mx3-Mx2)*ex_decomp2

cause_mat1 <- cause_fac2_1 - cause_fac1_1

cause_mat2 <- cause_fac3_2 - cause_fac2_2

```

```

table <- matrix(c(
  round(c(life.table(Mx1,"f")$ex[1],
    life.table(Mx2,"f")$ex[1]),2),
  round(c(
    sum(ex_decomp1)/10,
    sum(cause_mat1[,6:7])/10,
    sum(cause_mat1[,2])/10,
    sum(cause_mat1[,c(1,3:5,8:15)])/10,
    sum(cause_mat1[,16])/10,
    sum(cause_mat1)/10),3)))

row.names(table) <- c(
  "Life expectancy at birth for USA, 2000",
  "Life expectancy at birth for USA, 2010",
  "Life expectancy annualized difference between 2000 and 2010",
  "Cardiovascular disease component",
  "Neoplasms component",
  "Other caueses component",
  "External causes component",
  "Estimated total difference from decompositon")

colnames(table) <- "USA"

kable(table, caption = "Arriaga Decomposition by age and cause")

```

Table 28: Arriaga Decomposition by age and cause

	USA
Life expectancy at birth for USA, 2000	79.430
Life expectancy at birth for USA, 2010	81.060
Life expectancy annualized difference between 2000 and 2010	0.164
Cardiovascular disease component	0.080
Neoplasms component	0.045
Other caueses component	0.047
External causes component	-0.007
Estimated total difference from decompositon	0.164



```

table <- matrix(c(
  round(c(life.table(Mx2,"f")$ex[1],
    life.table(Mx3,"f")$ex[1]),2),
  round(c(
    sum(ex_decomp2)/8,
    sum(cause_mat2[,6:7])/8,
    sum(cause_mat2[,2])/8,
    sum(cause_mat2[,c(1,3:5,8:15)])/8,
    sum(cause_mat2[,16])/8,
    sum(cause_mat2)/8,3)))

row.names(table) <- c(
  "Life expectancy at birth for USA, 2010",
  "Life expectancy at birth for USA, 2018",
  "Life expectancy annualized difference between 2010 and 2018",
  "Cardiovascular disease component",
  "Neoplasms component",
  "Other caueses component",
  "External causes component",
  "Estimated total difference from decompositon")

colnames(table) <- "USA"

kable(table, caption = "Arriaga Decomposition by age and cause")

```

Table 29: Arriaga Decomposition by age and cause

	USA
Life expectancy at birth for USA, 2010	81.060
Life expectancy at birth for USA, 2018	81.350
Life expectancy annualized difference between 2010 and 2018	0.035
Cardiovascular disease component	0.013
Neoplasms component	0.009
Other caueses component	0.021
External causes component	-0.006
Estimated total difference from decompositon	0.035

## Part 2

Applying the Canudas-Romo (2003) method to a later date, say 2009 to 2019.

```
# Remember to load the two essential functions,
# the life table function and the function to
# calculate e-dagger or life disparity.

Y1 <- 2009
Y2 <- 2019

tp <- Y2-Y1

LT1 <- read.table('data/MortDecomp/SWE.bltper_1x1.txt',
                  header=T,skip=2)
LT1 <- LT1[LT1$Year==Y1,]

LT2 <- read.table('data/MortDecomp/SWE.bltper_1x1.txt',
                  header=T,skip=2)
LT2 <- LT2[LT2$Year==Y2,]

gap <-
  (life.table(LT2$mx,sex="m")$ex[1]-
   life.table(LT1$mx,sex="m")$ex[1])+
  (life.table(LT2$mx,sex="f")$ex[1]-
   life.table(LT1$mx,sex="f")$ex[1])

gap <- gap/(tp*2)

ex <- (LT1$ex+LT2$ex)/2

edag <- (ineq_edag(0:110,LT1$dx,LT1$lx,LT1$ex,LT1$ax)*
        ineq_edag(0:110,LT2$dx,LT2$lx,LT2$ex,LT2$ax))^0.5

rho <- -log(LT2$mx/LT1$mx)/tp

fx <- (LT1$dx/100000*LT2$dx/100000)^0.5

main <- sum(rho*fx)*edag[1]
```

```

error <- sum((rho-sum(rho*fx))*(ex-edag[1])*fx)

gap2 <- main+error

table <-
  matrix(round(c(LT1$ex[1],LT2$ex[1],gap,
               sum(rho*fx),edag[1],
               main,error,gap2),3),ncol=1)

row.names(table) <- c(paste0("Life expectancy at ",Y1),
  paste0("Life expectancy at ",Y2),
  paste0("Annualized change between ",
        Y1,"-",Y2),
  "Average improvements in mortality",
  "Life disparity at birth",
  "Direct component",
  "Covarinace component",
  "Total estimated difference")

colnames(table) <- "SWE"

kable(table,caption = "Decomposition of life expectancy (VCR)")

```

Table 30: Decomposition of life expectancy (VCR)

	SWE
Life expectancy at 2009	81.370
Life expectancy at 2019	83.050
Annualized change between 2009-2019	0.169
Average improvements in mortality	0.015
Life disparity at birth	9.402
Direct component	0.144
Covarinace component	0.024
Total estimated difference	0.169

## Assignments 4 Answer

```
library(tidyverse)
library(readxl)
```

Warning: package 'readxl' was built under R version 4.2.3

```
c="Nigeria"

p = c("1980-1985", "1985-1990", "1990-1995",
      "1995-2000", "2000-2005", "2005-2010",
      "2010-2015", "2015-2020")

yeard = 5

# data cleaning ----
SRB <-
  read_excel("data/WPP/WPP2019_FERT_F02_SEX_RATIO_AT_BIRTH.xlsx",
             sheet = "ESTIMATES", range = "C17:U272") %>%
  filter(Type == "Country/Area") %>% select(c(1,6:19)) %>%
  pivot_longer(2:15, names_to = "Period")

Fertility <-
  read_excel("data/WPP/WPP2019_FERT_F07_AGE_SPECIFIC_FERTILITY.xlsx",
             sheet = "ESTIMATES", range = "C17:O3587") %>%
  filter(Type == "Country/Area") %>% select(c(1,6:13)) %>%
  pivot_longer(3:9, names_to = "Age")

Fertility$Age <- substr(Fertility$Age, 1, 2)

Fertility <- left_join(Fertility, SRB,
                      by = c("Region, subregion, country or area *",
                             "Period"))

Mortality <-
  read_excel("data/WPP/WPP2019_MORT_F15_3_LIFE_TABLE_SURVIVORS_FEMALE.xlsx",
             sheet = "ESTIMATES", range = "C17:AD3587") %>%
  filter(Type == "Country/Area") %>% select(c(1,6:28)) %>%
  pivot_longer(3:24, names_to = "Age")
```

```

Mortality$value = as.numeric(Mortality$value)/100000

# NRR
B <- left_join(Fertility,Mortality,
               by = c("Region, subregion, country or area *",
                      "Period", "Age"))

B = B %>% filter(`Region, subregion, country or area *`==c,Period %in% p)

colnames(B) <- c("Country","Year","Age","ASFR","SRB","lx")

B = B %>% mutate(ASFR= as.numeric(ASFR)*5/1000) %>% mutate(F.per=(1-as.numeric(SRB)/

B$Year <- substr(B$Year,1,4)

B$Year <- as.numeric(B$Year)

Sums = B %>% group_by(Year) %>% summarise(NRR = sum(ASNRR))

Year1 <- 2000

Year2 <- 2010

r_nrr <-
  log(Sums$NRR[which(Sums$Year==Year2)]/
      Sums$NRR[which(Sums$Year==Year1)])/(Year2-Year1)

d_nrr <-
  sqrt(Sums$NRR[which(Sums$Year==Year2)]*
      Sums$NRR[which(Sums$Year==Year1)])*r_nrr

r_tfr <-
  log(B$ASFR[which(B$Year==Year2)]/
      B$ASFR[which(B$Year==Year1)])/
  (Year2-Year1)

r_tfr[is.na(r_tfr)] <- 0
r_tfr[is.nan(r_tfr)] <- 0
r_tfr[is.infinite(r_tfr)] <- 0

```

```

r_lx <-
  log(B$lx[which(B$Year==Year2)]/
      B$lx[which(B$Year==Year1)])/(Year2-Year1)

r_per <-
  log(B$F.per[which(B$Year==Year2)]/
      B$F.per[which(B$Year==Year1)])/(Year2-Year1)

m_tfr <-
  sqrt(B$ASFR[which(B$Year==Year1)]*
      B$ASFR[which(B$Year==Year2)])

m_lx <- sqrt(B$lx[which(B$Year==Year1)]*
            B$lx[which(B$Year==Year2)])

m_per <- sqrt(B$F.per[which(B$Year==Year1)]*
            B$F.per[which(B$Year==Year2)])

d_tfr <- sum(m_tfr*m_lx*m_per*r_tfr)
d_lx <- sum(m_tfr*m_lx*m_per*r_lx)
d_per <- d_nrr-d_tfr-d_lx

table <-
  matrix(round(
    c(Sums$NRR[Sums$Year==Year1],
      Sums$NRR[Sums$Year==Year2],
      d_nrr,d_lx,d_tfr+d_per),3),ncol = 1)

row.names(table) <-
  c(paste0("NRR in ",Year1),
    paste0("NRR in ",Year2),
    "Annualized change of NRR",
    "Mortality Component",
    "Fertility Component")

colnames(table) <- c

kable(table,caption = "NRR decomposition")

```

Table 31: NRR decomposition

	Nigeria
NRR in 2000	2.088
NRR in 2010	2.152
Annualized change of NRR	0.006
Mortality Component	0.018
Fertility Component	-0.011

## References

- Andersen, Per Kragh, Vladimir Canudas-Romo, and Niels Keiding. 2013. "Cause-Specific Measures of Life Years Lost." Journal Article. *Demographic Research* 29: 1127–52. <https://doi.org/10.4054/DemRes.2013.29.41>.
- Arriaga, Eduardo E. 1984. "Measuring and Explaining the Change in Life Expectancies." *Demography* 21 (1): 83–96. <https://doi.org/10.2307/2061029>.
- Arthur, W. Brian, and James W. Vaupel. 1984. "Some General Relationships in Population Dynamics." Journal Article. *Population Index* 50 (2): 214–26. <https://doi.org/10.2307/2736755>.
- Brouard, Nicholas. 1986. "Structure and Dynamics of Populations: The 'Years-to-Live' Pyramid, National Aspects and Regional Examples." Journal Article. *Espace Populations Sociétés*. <https://doi.org/10.3406/espos.1986.1120>.
- Canudas-Romo, Vladimir, and Michel Guillot. 2015. "Truncated Cross-Sectional Average Length of Life: A Measure for Comparing the Mortality History of Cohorts." Journal Article. *Population Studies* 69 (2): 147–59. <https://doi.org/10.1080/00324728.2015.1019955>.
- Canudas-Romo, Vladimir, Tianyu Shen, and Collin F. Payne. 2022. "The Components of Change in Population Growth Rates." Journal Article. *Demography*. <https://doi.org/10.1215/00703370-9765067>.
- Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." In, 1–102. Elsevier. [https://doi.org/10.1016/s0169-7218\(11\)00407-2](https://doi.org/10.1016/s0169-7218(11)00407-2).
- Guillot, Michel. 2003. "The Cross-Sectional Average Length of Life (CAL): A Cross-Sectional Mortality Measure That Reflects the Experience of Cohorts." Journal Article. *Population Studies* 57 (1): 41–54. <https://doi.org/10.1080/0032472032000061712>.
- Gupta, P. Das. 1993. *Standardization and Decomposition of Rates: A User's Manual (No. 186)*. Book. US Department of Commerce, Economics; Statistics Administration, Bureau of the Census.
- HFD. 2023. "Human Fertility Database (HFD)." Dataset. Max Planck Institute for Demographic Research (Germany); Vienna Institute of Demography (Austria). [Available at www.humanfertility.org](http://www.humanfertility.org).
- HMD. 2023. "HMD (Human Mortality Database)." Dataset. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA); French Institute for Demographic Studies (France). [Available at: http://www.mortality.org](http://www.mortality.org).
- Horiuchi, Shiro. 1991. "Assessing the Effects of Mortality Reduction on Population Ageing." *Population Bulletin of the United Nations Nos.31-32*, 31–51.
- Horiuchi, Shiro, and Samuel H. Preston. 1988. "Age-Specific Growth Rates: The Legacy of Past Population Dynamics." Journal Article. *Demography* 25 (3): 429–41. <https://doi.org/10.2307/2061542>.
- Kim, Young J., and Donna M. Strobino. 1984. "Decomposition of the Difference Be-



- tween Two Rates with Hierarchical Factors.” *Demography* 21 (3): 361–72. <https://doi.org/10.2307/2061165>.
- Kitagawa, Evelyn M. 1955. “Components of a Difference Between Two Rates.” *Journal of the American Statistical Association* 50 (272): 1168. <https://doi.org/10.2307/2281213>.
- Lazzari, Ester, Ryohei Mogi, and Vladimir Canudas-Romo. 2021. “Educational Composition and Parity Contribution to Completed Cohort Fertility Change in Low-Fertility Settings.” Journal Article. *Population Studies* 75 (2): 153–67. <https://doi.org/10.1080/00324728.2021.1895291>.
- Mogi, Ryohei, Jessica Nisén, and Vladimir Canudas-Romo. 2021. “Cross-Sectional Average Length of Life Childless.” Journal Article. *Demography* 58 (1): 321–44. <https://doi.org/10.1215/00703370-8937427>.
- Newton, Isaac. 1704. *Opticks: Or, a Treatise of the Reflections, Refractions, Inflexions and Colours of Light*. Book. London: Printed for Sam. Smith,; Benj. Walford. Retrieved from <https://doi.org/10.5479/sil.302475.39088000644674>.
- Pollard, J. H. 1988. “On the Decomposition of Changes in Expectation of Life and Differentials in Life Expectancy.” *Demography* 25 (2): 265–76. <https://doi.org/10.2307/2061293>.
- Preston, Samuel H. 1976. “Family Sizes of Children and Family Sizes of Women.” *Demography* 13 (1): 105–14. <https://doi.org/10.2307/2060423>.
- Preston, Samuel H., and Ansley J. Coale. 1982. “Age Structure, Growth, Attrition, and Accession: A New Synthesis.” Journal Article. *Population Index* 48 (2): 217–59. <https://doi.org/10.2307/2735961>.
- Preston, Samuel H., and Andrew Stokes. 2012. “Sources of Population Aging in More and Less Developed Countries.” Journal Article. *Population and Development Review* 38 (2): 221–36. <https://doi.org/10.1111/j.1728-4457.2012.00490.x>.
- Preston, Samuel, Patrick Heuveline, and Michael Guillot. 2001. “Demography: Measuring and Modeling Population Processes.” Journal Article. *Malden, MA: Blackwell Publishers*.
- Shen, Tianyu, Ester Lazzari, and Vladimir Canudas-Romo. 2023. “The Contribution of Survival to Changes in the Net Reproduction Rate.” *Population Studies*, March, 1–16. <https://doi.org/10.1080/00324728.2023.2187441>.
- Vaupel, James W., and Vladimir Canudas-Romo. 2002. “Decomposing Demographic Change into Direct Vs. Compositional Components.” *Demographic Research* 7 (July): 1–14. <https://doi.org/10.4054/demres.2002.7.1>.
- Vaupel, James W., and Vladimir Canudas Romo. 2003. “Decomposing Change in Life Expectancy: A Bouquet of Formulas in Honor of Nathan Keyfitz’s 90th Birthday.” *Demography* 40 (2): 201–16. <https://doi.org/10.1353/dem.2003.0018>.