European Doctoral School of Demography (EDSD)
Decomposition Techniques - Final Assignment

Ainhoa-Elena Leger
Gonzalo Daniel Garcia
Liliana Patricia Calderon Bernal
Marilyn-Anne Tremblay
Özge Elif Özer

01/6/2021

We should include some note to say JM to type the username and password.

Challenge 1

Proof Kitagawa decomposition (1995) without interactions

Define the difference between the crude death rates as Δ.

$$\Delta\text{CDR} = \sum_x M_x(t_2)\frac{N_x(t_2)}{N(t_2)} - \sum_x M_x(t_1)\frac{N_x(t_1)}{N(t_1)}$$

I divide each of the terms into two equal parts and add and subtract some additional terms, thereby keeping the difference (Δ) constant.

$$\begin{aligned}\Delta\text{CDR} = &\frac{\sum_x M_x(t_2)\frac{N_x(t_2)}{N(t_2)}}{2} + \frac{\sum_x M_x(t_2)\frac{N_x(t_2)}{N(t_2)}}{2} - \frac{\sum_x M_x(t_1)\frac{N_x(t_1)}{N(t_1)}}{2} - \frac{\sum_x M_x(t_1)\frac{N_x(t_1)}{N(t_1)}}{2}\\ &+ \frac{\sum_x M_x(t_1)\frac{N_x(t_2)}{N(t_2)}}{2} - \frac{\sum_x M_x(t_1)\frac{N_x(t_2)}{N(t_2)}}{2} + \frac{\sum_x M_x(t_2)\frac{N_x(t_1)}{N(t_1)}}{2} - \frac{\sum_x M_x(t_2)\frac{N_x(t_1)}{N(t_1)}}{2}\end{aligned}$$

I now combine the eight terms in Δ into four:

$$\begin{aligned}\Delta\text{CDR} = &\sum_x \frac{N_x(t_2)}{N(t_2)}\left(\frac{M_x(t_2)+M_x(t_1)}{2}\right) - \sum_x \frac{N_x(t_1)}{N(t_1)}\left(\frac{M_x(t_2)+M_x(t_1)}{2}\right)\\ &+ \sum_x M_x(t_2)\left(\frac{\frac{N_x(t_2)}{N(t_2)}+\frac{N_x(t_1)}{N(t_1)}}{2}\right) - \sum_x M_x(t_1)\left(\frac{\frac{N_x(t_2)}{N(t_2)}+\frac{N_x(t_1)}{N(t_1)}}{2}\right).\end{aligned}$$

Finally, we combine the terms into two:

$$\Delta\text{CDR} = \sum_x \left(\frac{M_x(t_2)+M_x(t_1)}{2}\right)\left(\frac{N_x(t_2)}{N(t_2)} - \frac{N_x(t_1)}{N(t_1)}\right) + \sum_x \left(\frac{\frac{N_x(t_2)}{N(t_2)}+\frac{N_x(t_1)}{N(t_1)}}{2}\right)(M_x(t_2) - M_x(t_1)).$$

The first terms is the difference in age composition weighted by the average age-specific mortality, while the second term is the difference in rate schedules weighted by the average age composition. Therefore, Δ is equal to the sum of the contribution of age compositional differences and the contribution of rate schedule differences.

Challenge 2

With data on fertility (e.g. HFD) select 3 countries and analyze the change in their crude fertility rate (CFR) in a recent period (10 years) and decompose these changes following Kitagawa's decomposition and describe your results. Then for the most recent period select the two countries (among the 3) with the highest and lowest CFR and decompose their difference and describe your results.

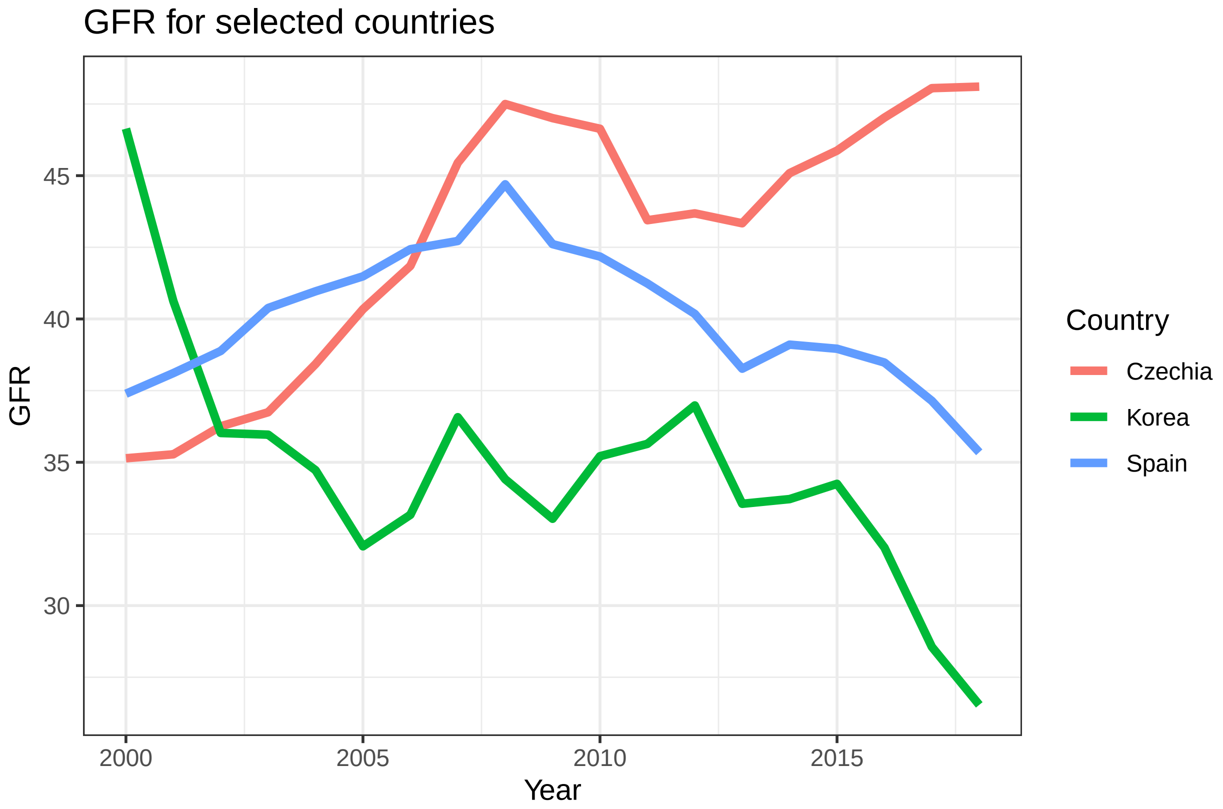
For doing this challenge we decided to take Spain, Korea and Czechia as the three countries to compare. The data was retrieved from the Human Fertility Database (HFD), [@jasilioniene2015methods] using Tim Riffe's package `HMDHFDplus` [@riffe2015reading]. The selected measure is the General Fertility Rate, as defined in page 93 of Preston's book [@preston2000demography]:

$$GFR[0,T] = \frac{Births[0,T]}{Person-years-lived[0,T]}$$

We used this measure instead of the Crude Birth Rate to avoid a bias of different age-structure of populations.

A glimpse on data and some literature

After some data wrangling (code provided), we have the following time series of the GFR for the three countries:



Source: HFD

As we can see, Czechia presents an increasing GFR for the period, contrasting to Korea's GFR which is decreasing and the lowest one for the selected countries. As for Spain, the GFR was increasing until the 2008/9 Financial Crisis, and from that point onward decreases.

Literature suggests that Korea has a decreasing fertility due to socioeconomic factors, especially pertaining to the domestic division of labor and the availability of State help to childcare [@kim2017division; @seo2019low]. The variability of the time series could be related to tempo effects [@yoo2018ultra].

When analyzing Spain, most of literature relates fertility to socioeconomic factors [@barbieri2015rise; @sobotka2011economic]. The GFR curve, with a peak around the Financial Crisis, shows this very clearly.

As for the case of Czechia, the behavior could be related to the country experiencing a Second Demographic Transition after the end of the USSR [@billingsley2010post]. We can also observe that the Financial Crisis impacted the GFR of this country, but the increasing trend remains after the economic recovery [@matysiak2021great].

Time to Decompose

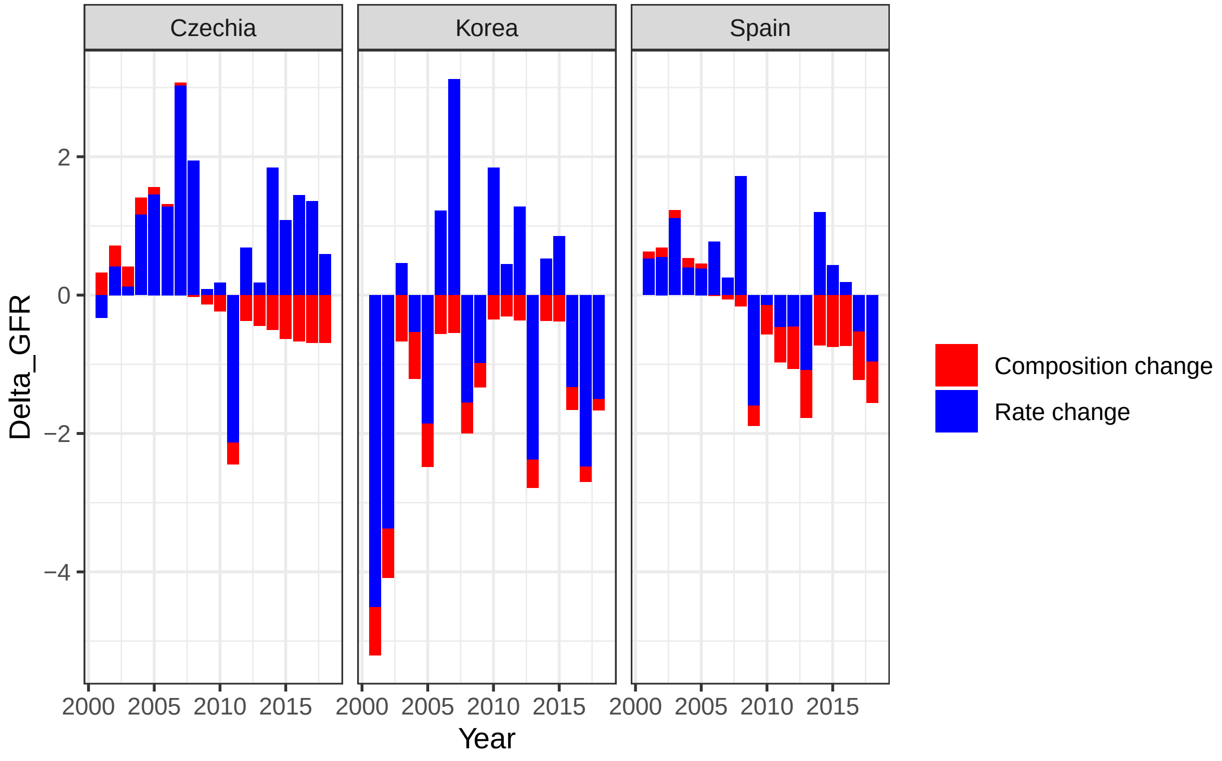
We will be using the Kitagawa decomposition method [@kitagawa1955components] components to unravel changes in composition of the population under analysis from the changes in the rate.

Basically, Kitagawa tells us that a change of any rate `R = A * B` can be decompose in this way [@tonnessen2019declined]:

$$\Delta R = (\Delta A * \overline{B}) + (\Delta B * \overline{A})$$

The following plot shows the decomposition of the GFR for each country for each of the years observed:

Kitagawa decomposition: Czechia, Korea & Spain, 2000−2018



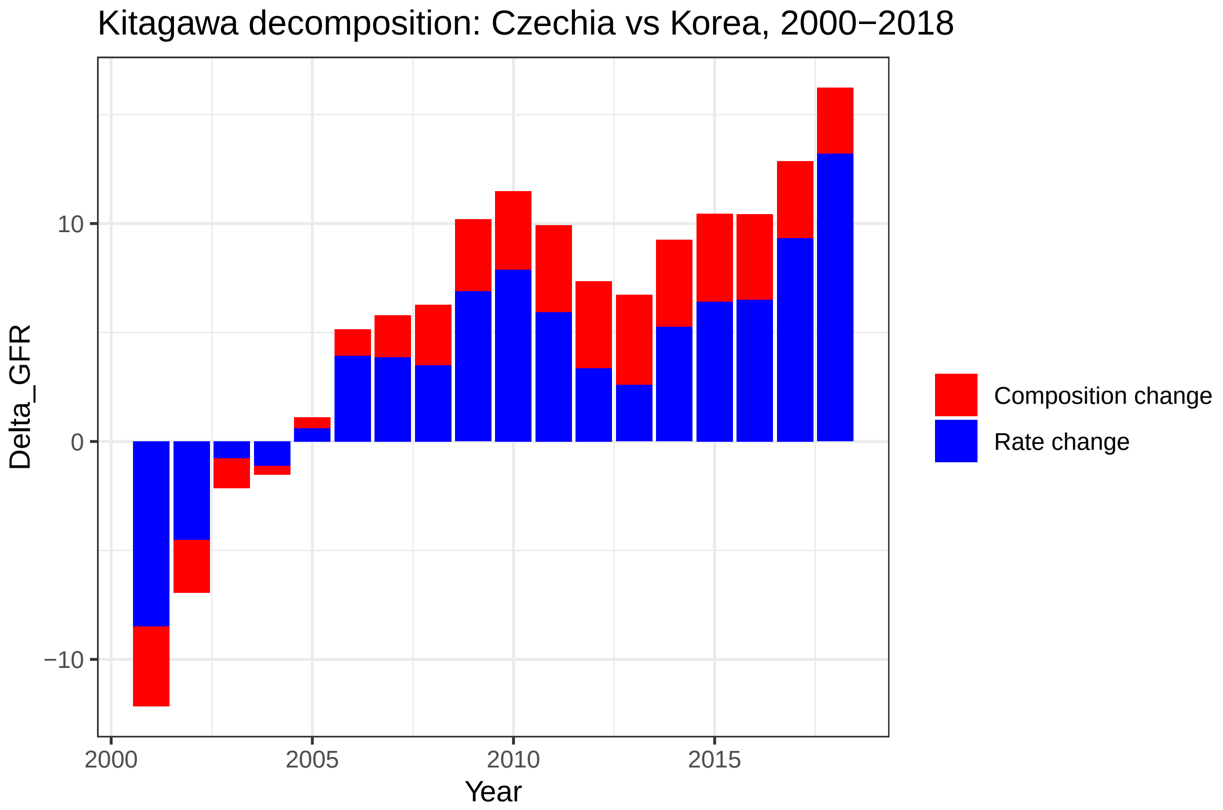
Source: HFD

We can observe that the source of change in the GFR for all countries is mostly due to changes in rate and not in population composition. But it is interesting that for both Czechia and Spain, the amount of change due to changes in composition has been increasing during this 2000-2018 period. This is related to the presence of aging populations, a common problem in European countries.

As for Korea, most part of the change in the GFR is related to changes in rate and not in population composition.

Comparing Countries

To finish this challenge, we can also decompose the difference in the GFR of Czechia and Korea. Since Czechia will be used as the second population and this country has experienced increasing GFR during the period, the decomposition should give positive numbers for most of the observed periods.



Source: HFD

The plot shows that the difference in GFR between the two countries is related more to changes in rates, rather than in population composition. But in those years where the change in rates is small, the composition factor plays a major role in explaining this differences.

Challenge 3

Get (any source) age-specific prevalences and apply the Sullivan method in R. Follow the practical guide if you want to and be aware of the method's limitations (e.g. same mortality schedule).

Check the names of the countries and the the items available in the HMDweb.

We will extract data from Lithuania (LTU) and the Netherlands (NLD), the two European countries with the highest (9.6) and the lowest (3.1) gender gap in life expectancy in 2019 according to Eurostat. https://ec

.europa.eu/eurostat/statistics-explained/index.php?title=File:Life_expectancy_gender_gap_2021-01.jpg

We first extract data on Deaths and Exposures by 5 year age group from the HMD for Lithuania and the Netherlands.

We compute the Age-Specific mortality rates by 5-year age groups from the HMD for each country and sex separately.

To apply the Sullivan method, we will use data on self-reported chronic morbidity (i.e. the presence of long-term (chronic) symptoms, health conditions or diseases)
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Chronic_morbidity.
This dimension of health is captured by one of the three questions of the Minimum European Health Module (MEHM): “Do you have any long-standing illness or health problem?”. THe MEHM is currently implemented in the EU Statistics on Income and Living Conditions (EU-SILC) and the European Health Interview Survey (EHIS)
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Minimum_European_Health_Module_(MEHM) The EU-SILC target population consists of all individuals Individuals aged 16 years old and over living in private households. Hence, people living in collective households and in institutions are generally excluded from the survey.

Data on chronic morbidity from the EU-SILC is expressed as percentages within the population combining various breakdowns: sex, age-group, labour status, educational attainment level, country of birth, country of citizenship, degree of urbanisation and income quintile (group).
https://ec.europa.eu/eurostat/web/health/methodology
Since we are only interested un the total population by sex and age-group, we will use the first data-set with labour status.

We extract the dataset of people having a long-standing illness or health problem, by sex, age and labour status (hlth_silc_04) and get a glimpse on it.

```
## Rows: 190,032
## Columns: 7
## $ unit    <chr> "PC", "PC", "PC", "PC", "PC", "PC", "PC", "PC", "PC", "PC",...
## $ wstatus <chr> "EMP", "EMP", "EMP", "EMP", "EMP", "EMP", "EMP", "EMP", "EM...
## $ age     <chr> "Y16-24", "Y16-24", "Y16-24", "Y16-24", "Y16-24", "Y16-24",...
## $ sex     <chr> "F", "F", "F", "F", "F", "F", "F", "F", "M", "M", "M", "M",...
## $ geo     <chr> "AT", "BG", "DK", "EE", "FI", "HU", "NL", "RO", "AT", "BG",...
## $ time    <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020,...
## $ values  <dbl> 15.7, NA, 21.0, 18.6, 30.4, NA, NA, NA, 18.4, NA, NA, 18.2,...
```

```
## [1] "EMP"      "INAC_OTH" "NEMP"     "NSAL"     "POP"      "RET"      "SAL"
## [8] "UNE"
```

```
##  [1] "Y16-24" "Y16-29" "Y16-44" "Y16-64" "Y25-29" "Y25-34" "Y35-44" "Y45-49"
##  [9] "Y45-54" "Y45-64" "Y55-64" "Y65-74" "Y75-84" "Y_GE16" "Y_GE65" "Y_GE75"
## [17] "Y_GE85"
```

We filter the data to keep only information for Lithuania and the Netherlands in 2009 for each sex separately, as well as the total population (wstatus == POP) and the age values corresponding to a distribution by 10 years age group, since it is the most regular age-grouping across the whole dataset.

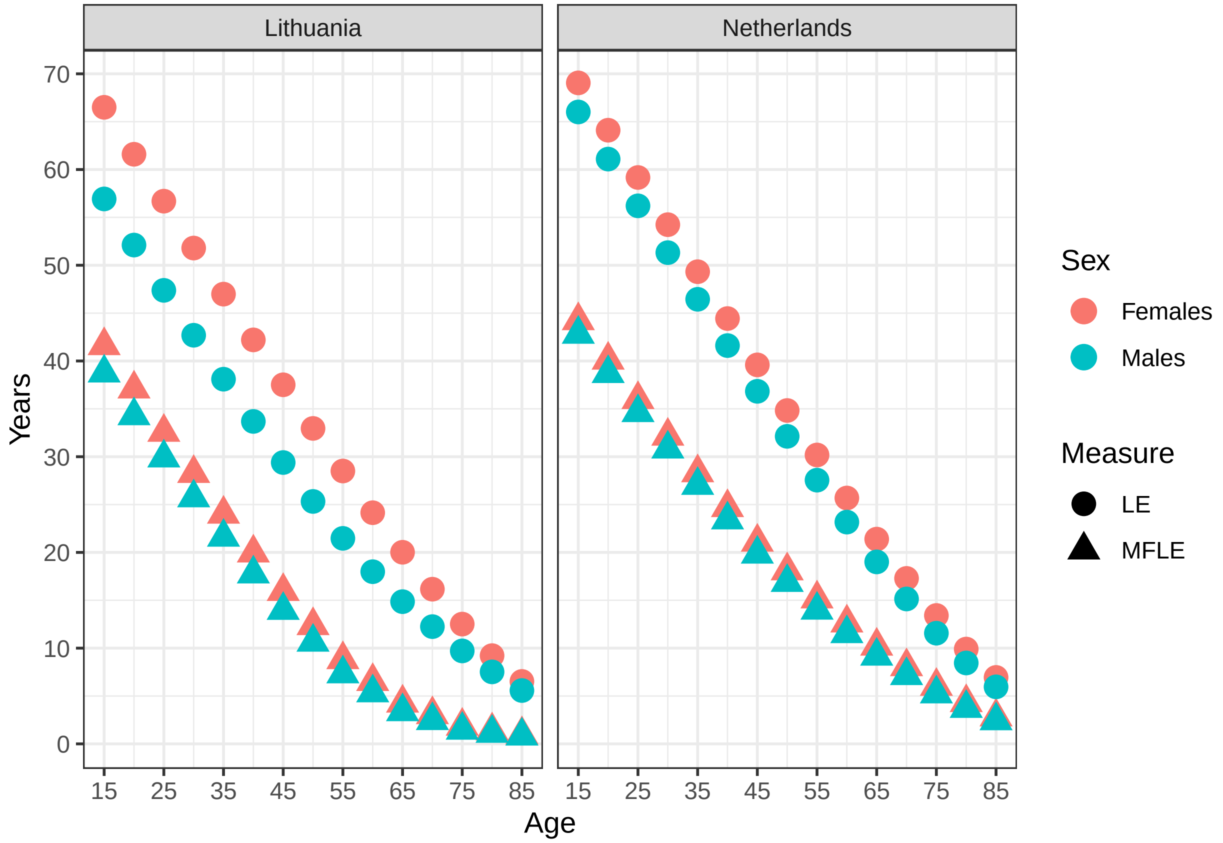
Since the death rates are available by 5-year age group, but chronic morbidity prevalence (pix) by 10-year age group, we need to add a correspondence 10-year age group value to the death rates, allowing us to join both tables. We made the assumption that the prevalence is the same for the two 5-year age groups included in each 10-year each group. We also split each dataset by sex.

To apply the Sullivan Method, we will consider the non-presence of chronic morbidity as an indicator of health status. We create a function allowing to compute general Life Expectancy (LE) and Morbidity Free Life Expectancy (MFLE) above age 15.

We apply the Sullivan function, to each dataset to compute the LE and the MFLE. We merge all datasets in a unique one to plot.

We can now plot all together.

Remaining Life Expectancy (LE) and Morbidity Free Life Expectancy (MFLE) by sex in Lithuania and the Netherlands in 2019.



Challenge 4

Use the linear integral model to decompose the change in the standard deviation of the age-at-death distribution and life expectancy by age and cause of death for 3 countries you might be interested in (over time or between them). Interpret the results of life expectancy alongside standard deviation. Make it interesting. You can use data from HCoD, HMD, WHO, GBD.

In this exercise, we aim to see the changes in the standard deviation of the age-at-death distribution and life expectancy by age and cause of death in Latvia, Russia and Poland from the Soviet Union dissolution to becoming independent states. Since the impact of macrolevel societal changes can take time to be observed in demographic behavior, we decided to look at the change from 1990 to 2010.

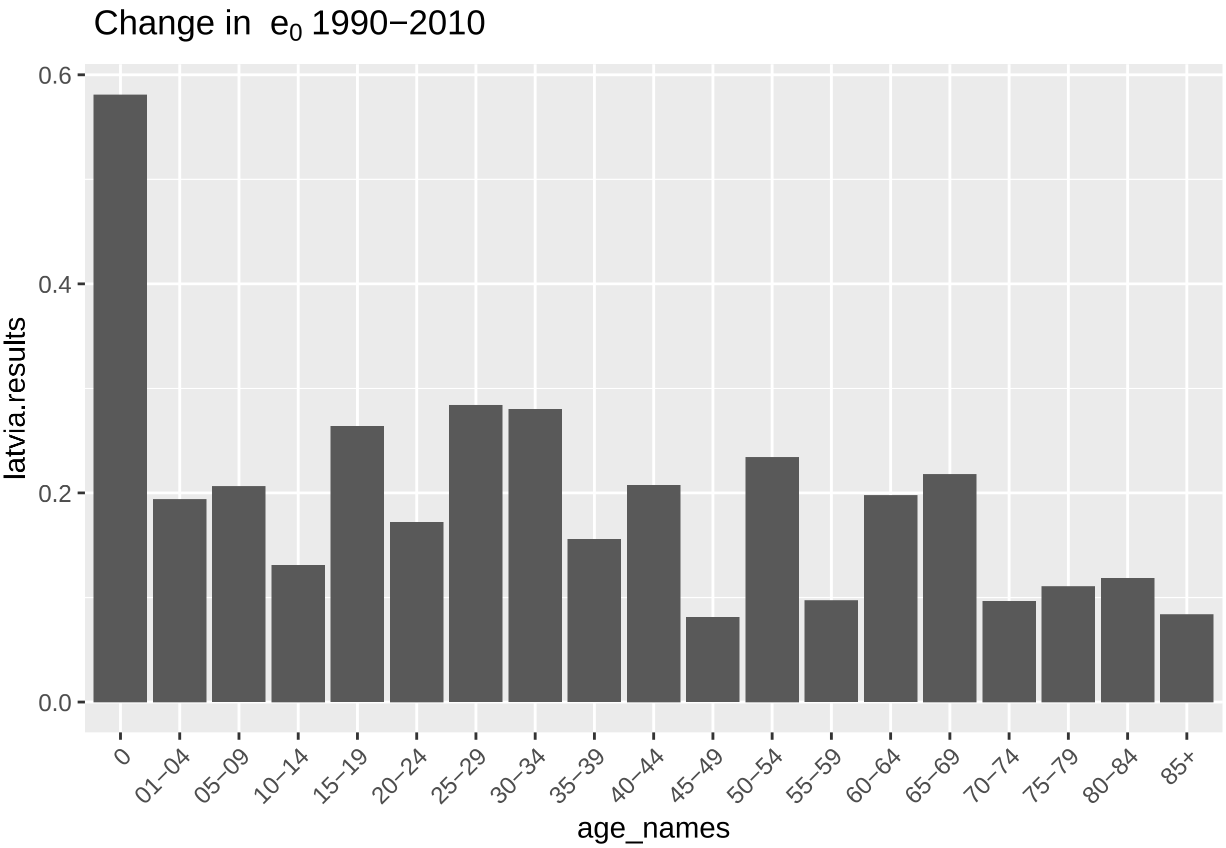
We will use the data from the Human Mortality Database (HMD) and the Cause of Death Database (CDD). As requested, we benefited from the Horiuchi and colleagues' (2008) linear integral decomposition model to decompose the changes in standard deviation and life expectancy. First we will look at it only by age and then with cause of mortality as well.

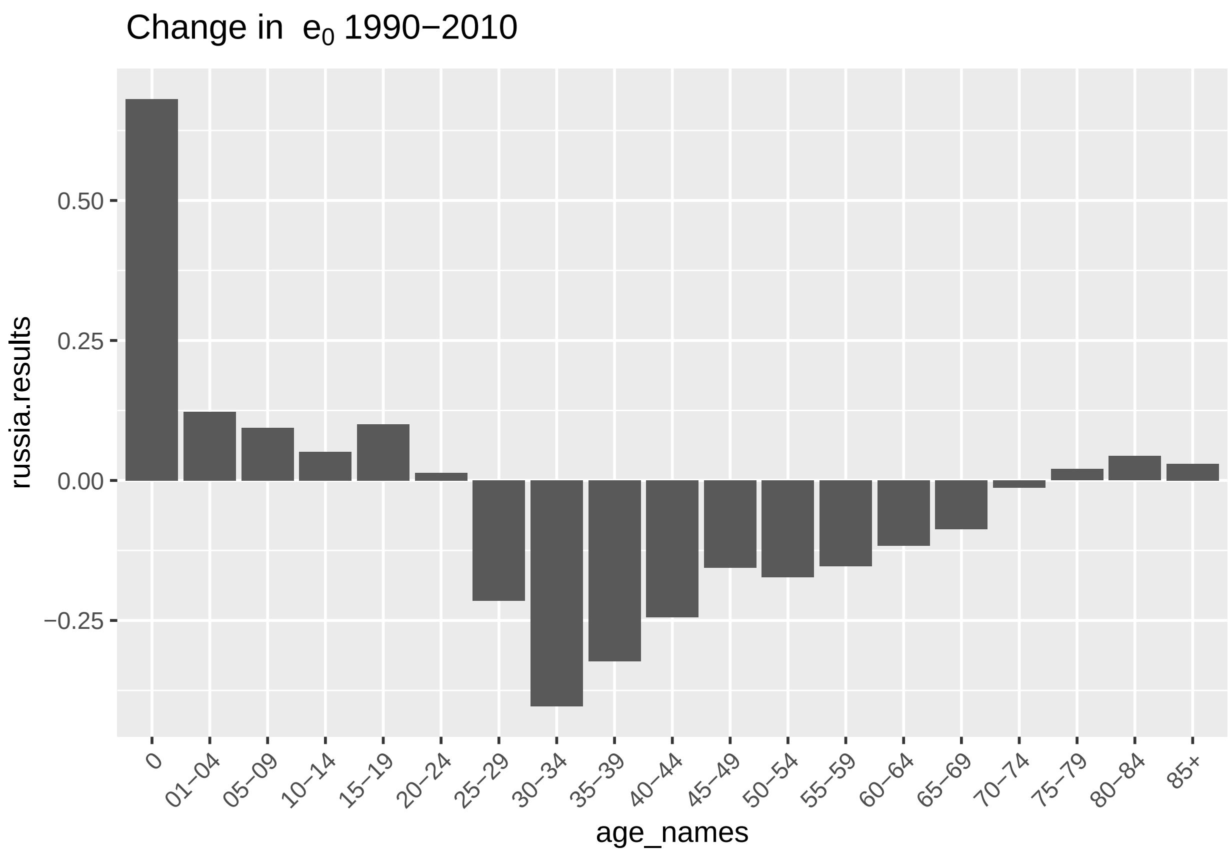
We first needed to arrange the dataset to make it ready for decomposition. The code for this is available on the rmd.

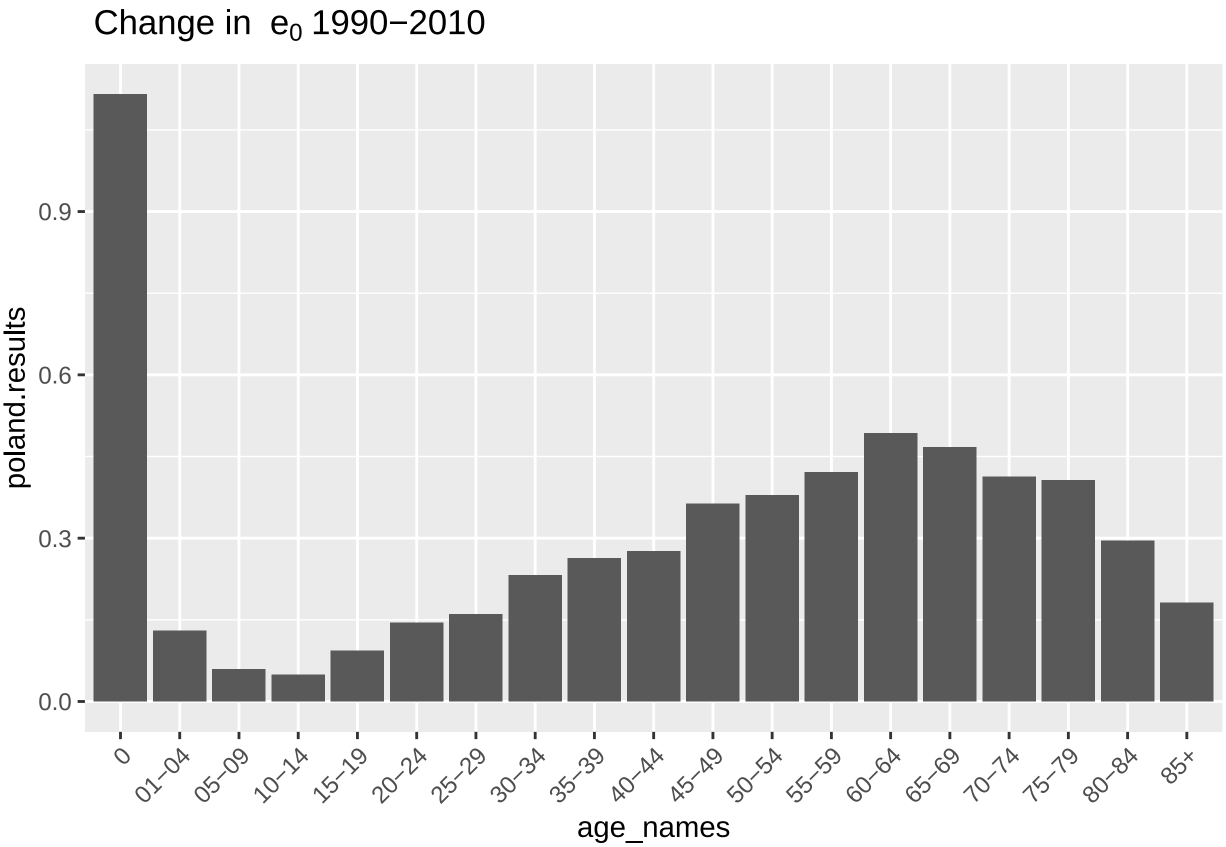
```
## [1]  3.7151230  3.7151224 -0.7273611 -0.7273611  5.9420555  5.9420509
```

```
## [1] -0.00000057451298  0.00000002265514 -0.00000455006681
```

The graphs




Change in e_0 1990−2010
russia.results
0.50
0.25
0.00
−0.25
0
01−04
05−09
10−14
15−19
20−24
25−29
30−34
35−39
40−44
45−49
50−54
55−59
60−64
65−69
70−74
75−79
80−84
85+
age_names



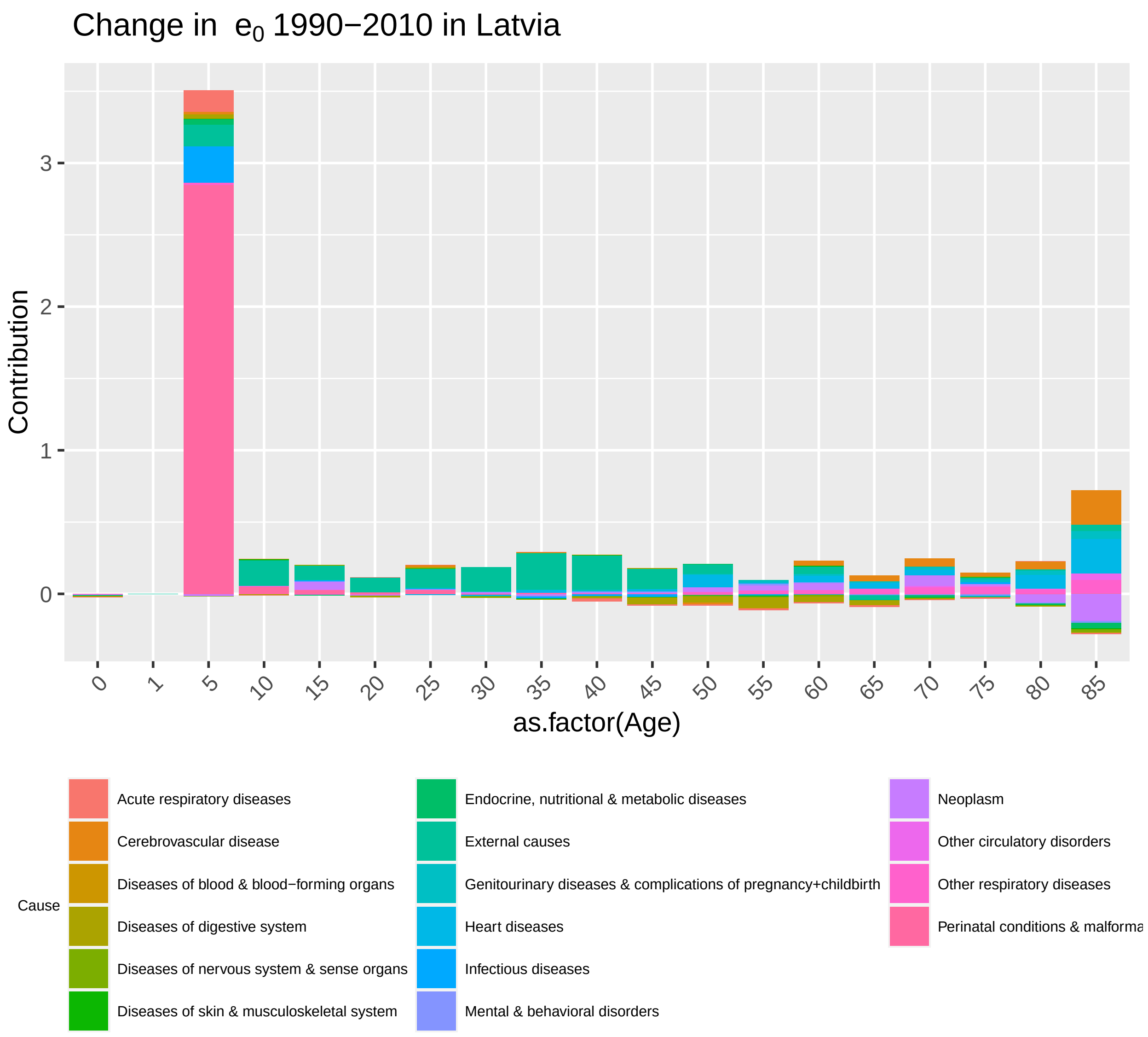
Unlike the graph for Russia, Latvia and Poland shows that there has been overall a positive change in the life expactancy.In all of the countries, the age group that has experienced the highest gain in life expectancy is the 0-5 ages. This is expected considering the increase in availability of communicable disease treatment for the infants and children. Russia shows a substantial decrease in life expectancy for the middle age groups. It should be noted that there are big differences between Russian males and females and we are only observing males here. The increase in alcohol consumtopn and smoking can be a reason for that and additionally the economic and political instability was much prolonged in the case of Russia after the dissolution of Soviet Union, and these can also be considered as causes of changes in the life expectancy.

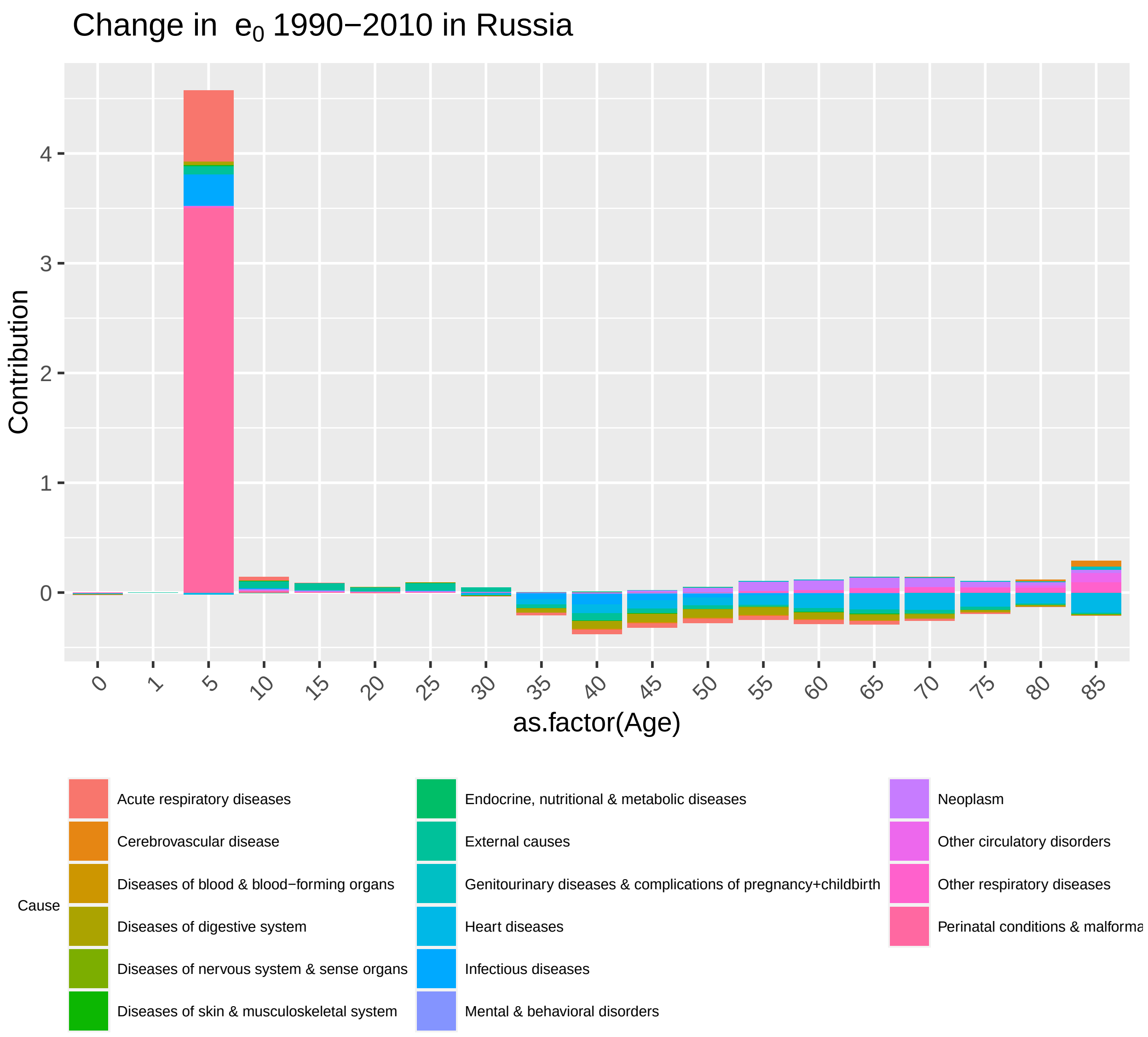
We now extend our findings to cause of death decomposition to see which cause of death plays major role in the life expectancy changes in Latvia, Russia and Poland from 1990 t0 2010.

```
## [1]  6.118467  6.118466  3.198322  3.198322 11.577461 11.577441
```

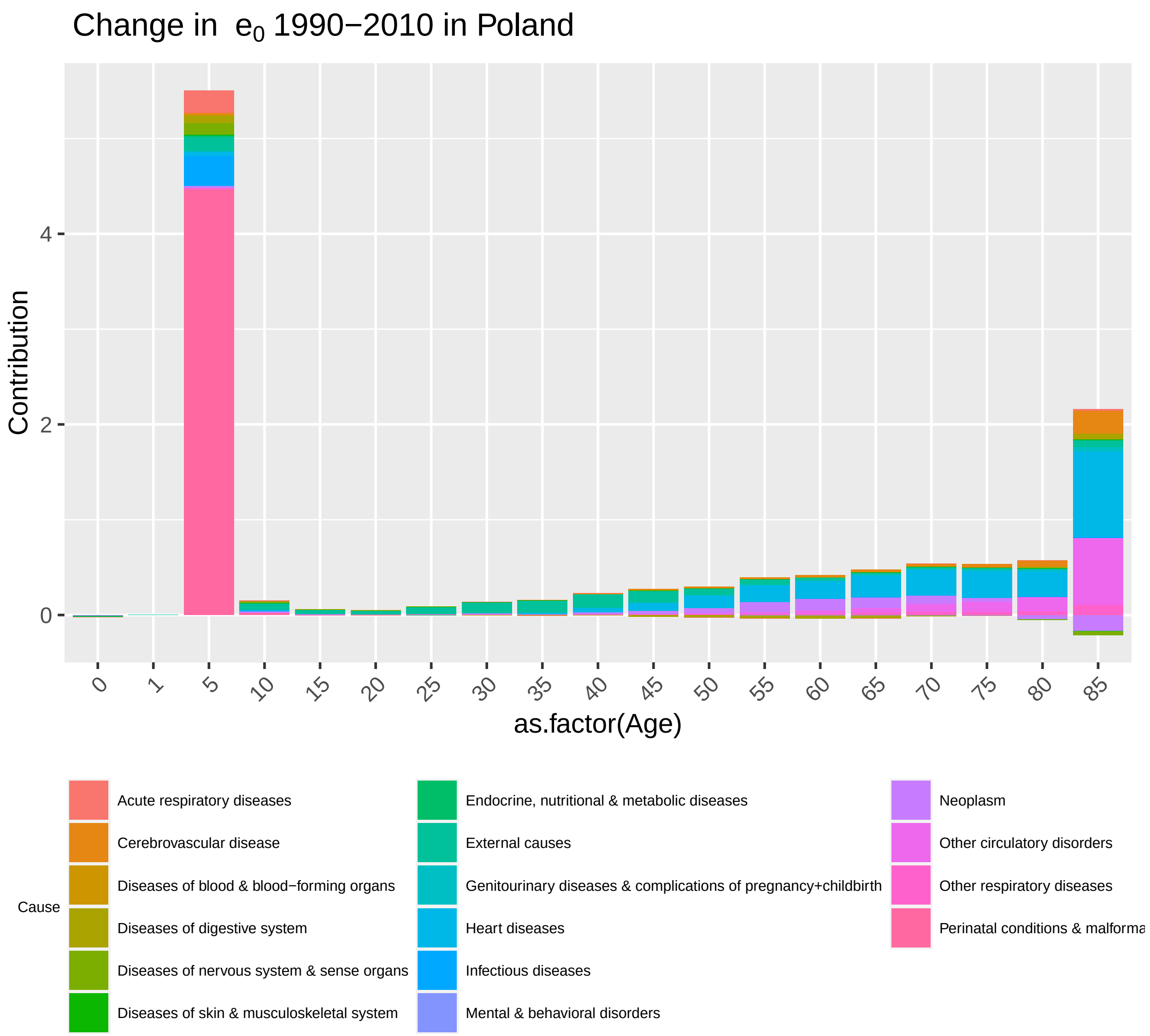
```
## [1] -0.00000081415119 -0.00000006621032 -0.00001934737402
```

The graphs for the life expectancies decomposed by age and cause of death are below.


Change in e_0 1990−2010 in Latvia
Contribution
3
2
1
0
0
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
as.factor(Age)
Cause
Acute respiratory diseases
Cerebrovascular disease
Diseases of blood & blood−forming organs
Diseases of digestive system
Diseases of nervous system & sense organs
Diseases of skin & musculoskeletal system
Endocrine, nutritional & metabolic diseases
External causes
Genitourinary diseases & complications of pregnancy+childbirth
Heart diseases
Infectious diseases
Mental & behavioral disorders
Neoplasm
Other circulatory disorders
Other respiratory diseases
Perinatal conditions & malforma



Change in e_0 1990−2010 in Russia
Contribution
4
3
2
1
0
0
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
as.factor(Age)
Cause
Acute respiratory diseases
Cerebrovascular disease
Diseases of blood & blood−forming organs
Diseases of digestive system
Diseases of nervous system & sense organs
Diseases of skin & musculoskeletal system
Endocrine, nutritional & metabolic diseases
External causes
Genitourinary diseases & complications of pregnancy+childbirth
Heart diseases
Infectious diseases
Mental & behavioral disorders
Neoplasm
Other circulatory disorders
Other respiratory diseases
Perinatal conditions & malforma

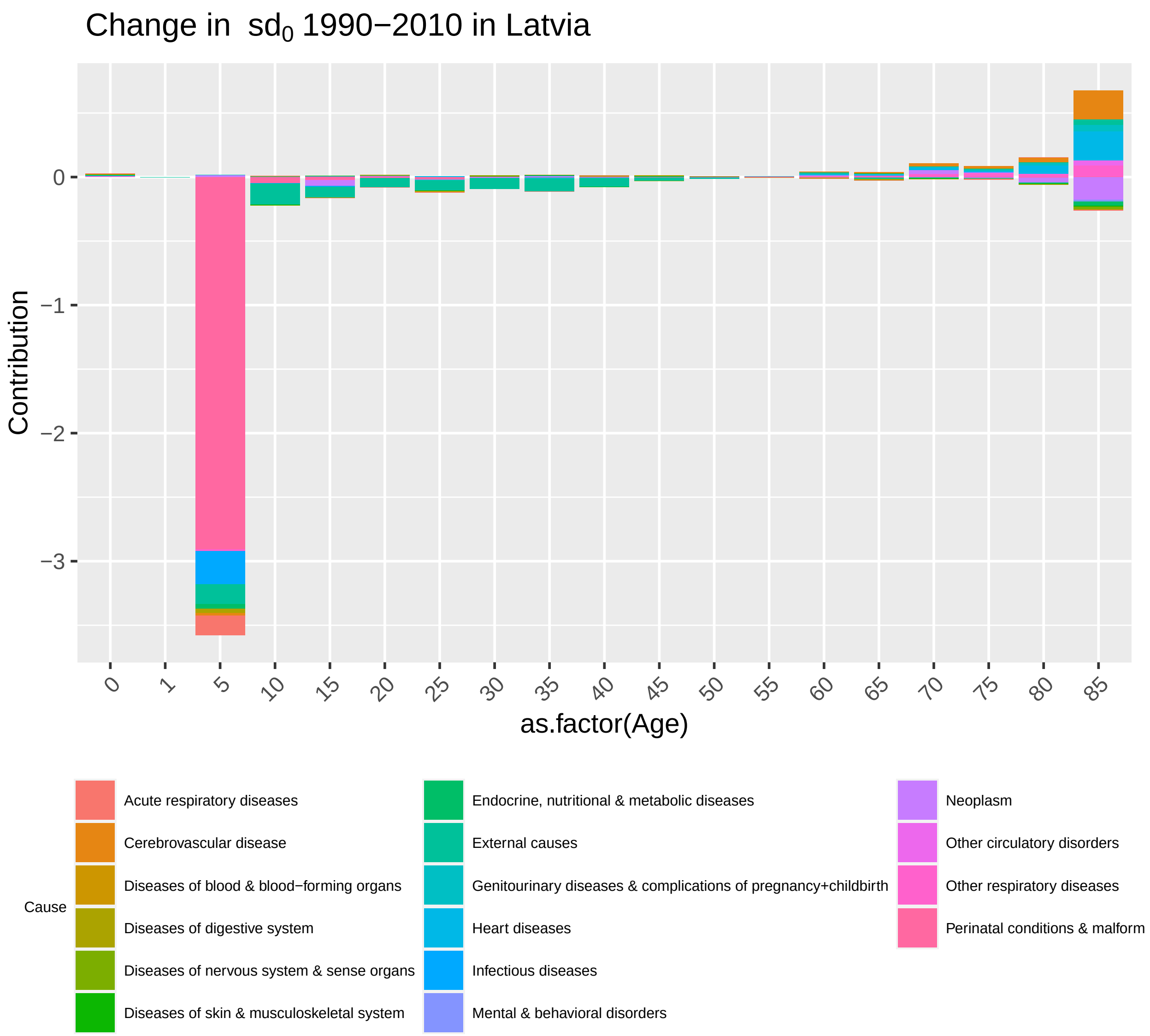


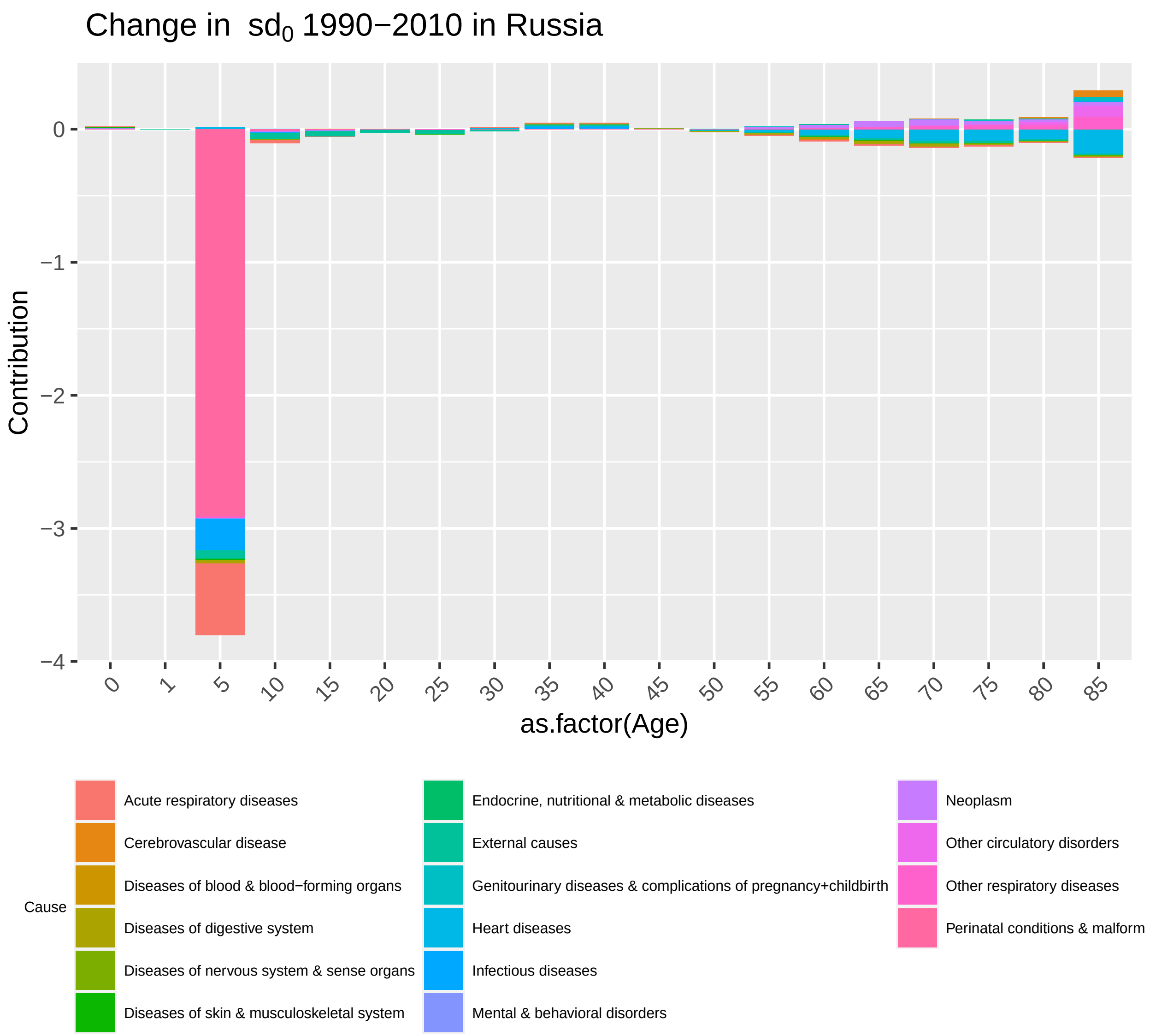
These set of graphs show the change in life expectancy decomposed by age and cause of death. For the all the countries, we can see that as suspected from the first graphs, the increase in the life expectancy at earlier ages is mostly from perinatal conditions and malformations. For the Russian case, the middle age groups had a decrease in life expectancy and the cause of death seemed to show that this is due to a rise in heart diseases. This can be due to change in alcohol and smoking habits as in Soviet Union, Gorbachov had campaigned against alcohol consumption and tobacco usage. For Poland, we observe that below age 55, there has been a decrease in death from external causes and for later ages there is a decrease for heart diseases so these are the main causes of deaths that contributed to the increase life expectancies in these age categories. For Latvia, between age 10 and 55, the main contributors to the change are again in external causes. These can be attributed to the stabilization of political, social and economical conditions for Latvia and Poland.

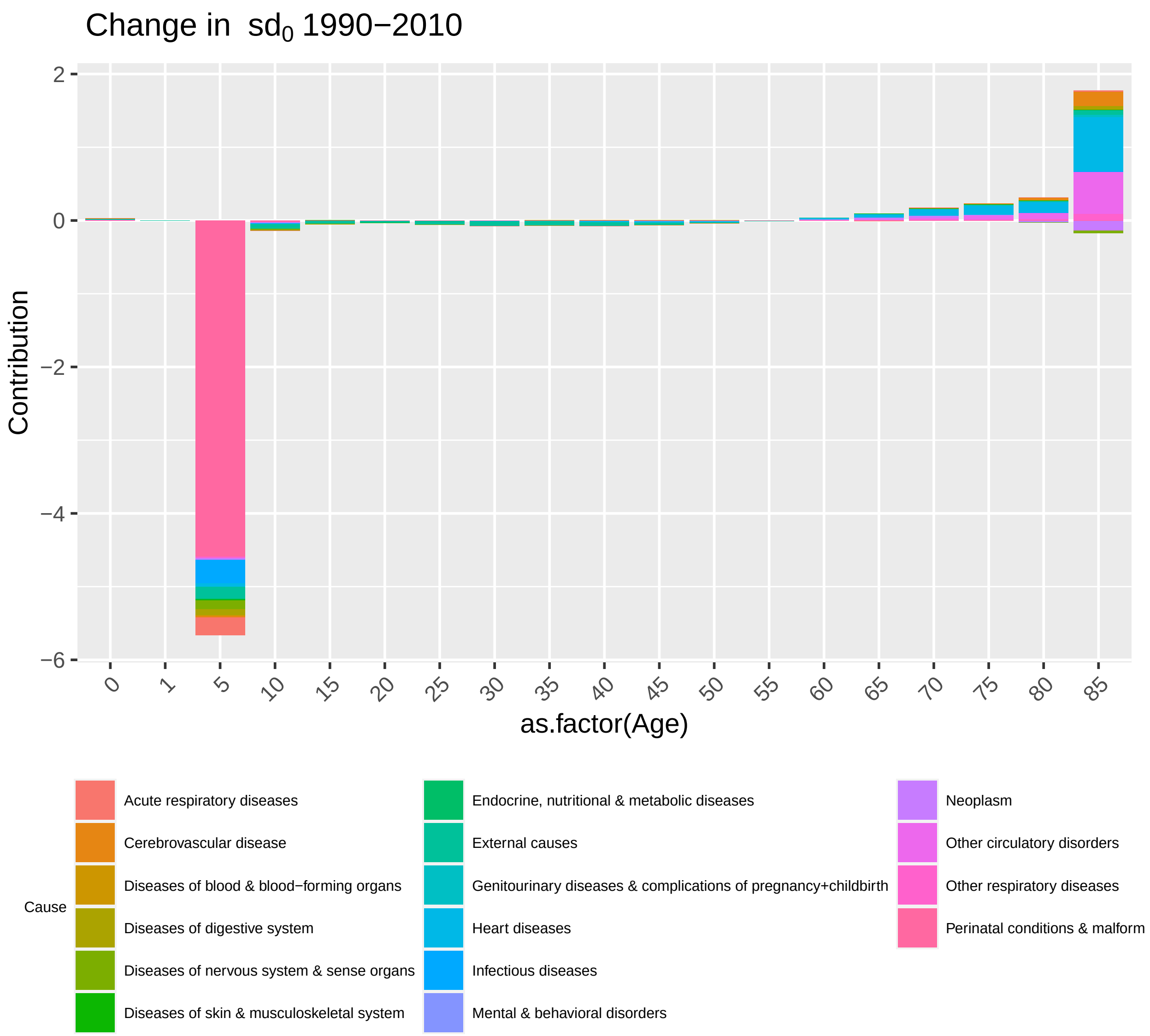
Now we can look at lifespan variation by calculating the standard deviation.

```
## [1] -3.647308 -3.647304 -4.108706 -4.108705 -3.818840 -3.818833
```

```
## [1] 0.000003980185 0.000001251437 0.000007661509
```


Change in sd_0 1990−2010 in Latvia
Contribution
0
−1
−2
−3
0
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
as.factor(Age)
Cause
Acute respiratory diseases
Cerebrovascular disease
Diseases of blood & blood−forming organs
Diseases of digestive system
Diseases of nervous system & sense organs
Diseases of skin & musculoskeletal system
Endocrine, nutritional & metabolic diseases
External causes
Genitourinary diseases & complications of pregnancy+childbirth
Heart diseases
Infectious diseases
Mental & behavioral disorders
Neoplasm
Other circulatory disorders
Other respiratory diseases
Perinatal conditions & malform


Change in sd_0 1990−2010 in Russia
Contribution
0
−1
−2
−3
−4
0
1
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
as.factor(Age)
Cause
Acute respiratory diseases
Cerebrovascular disease
Diseases of blood & blood−forming organs
Diseases of digestive system
Diseases of nervous system & sense organs
Diseases of skin & musculoskeletal system
Endocrine, nutritional & metabolic diseases
External causes
Genitourinary diseases & complications of pregnancy+childbirth
Heart diseases
Infectious diseases
Mental & behavioral disorders
Neoplasm
Other circulatory disorders
Other respiratory diseases
Perinatal conditions & malform



These last plots are for the changes in the standard deviation of the age-at-death distribution decomposed by age and cause of death for Russia, Poland and Latvia. We can observed that for Russian males, the main contributors for the decrease in people ages less than 5 are perinatal conditions and acute respiratory diseases. For the people aged more than 60, this seems to be heart diseases. For Latvia, the main contributor for the ages between 10 and 50 were external causes while for people above 50, the decrease is due to decrease in neoplasm and increase is due to heart diseases and cerebrovascular diseases. For Poland, there does not seem to be one main contributor for the changes in ages between 10 and 70. However, above 70 the main causes of deaths are increase in heart diseases, circulatory disorders and cerebrovascular disease. Overall, we can say that the increase observed in the selected post communist countries are mostly in the 0-5 ages. This can be due to multiple reasons such as development in medicine during these times and increase in availability of early child care. However, it is possible to say that there does not seem to be a homogenous pattern of change in life expectancy by ages and causes of death in these three countries.