COS-D407. *Scientific Modeling and Model Validation*

Hands-on excercises

Week 5

University of Helsinki, Finland

01.11.2021–15.12.2021

Lecturer: Christina Bohk-Ewald

Source: https://github.com/christina-bohk-ewald/2021-COS-D407-scientific-modeling-and-model-validation

Table of content:

1. Some preparations in R

1.1 Open a new script for week 5 in R (e.g., *week-5.R*) and save it to a folder of your choice (e.g., *course-COS-D407*).

1.2 Create a filepath to this folder from where you would like to load data and to where you would like to save your outcome. For example,

```
the_course_COS_D407_path <- c("C:/course-COS-D407")
```

1.3 You can then set the working directory to this path

```
setwd(the_course_COS_D407_path)
```

2. Download, load, and prepare required input data

In week 5 we analyze the robustness of the demographic scaling model's COVID-19 infection estimates for Finland with respect to IFR estimates. We will start with reading and comparing infection fatality rate estimates from different sources, continue with estimating COVID-19 infections for Finland over time with the demographic scaling model based on these different IFR estimates, and finally compare the resulting COVID-19 infection estimates in order to analyze the impact of these different IFR estimates.

Note that you know already the steps 2.1 through 2.5 from the previous lab session, steps 2.6 and 2.7 are new.

2.1 Download confirmed cases and reported deaths attributable to COVID-19 as of today

Please go to the website of the Johns Hopkins University CSSE. The files

- *time_series_covid19_confirmed_global.csv*
- *time_series_covid19_deaths_global.csv*

contain confirmed cases and reported deaths, respectively, for many countries on a daily basis since January 22, 2020. Please download these two files and save them in your project folder.

2.2 Load COVID-19 data

Please load the numbers of confirmed cases and reported deaths from COVID-19 in R using the function *read.csv* of the R-package *openxlsx*.

```
require(openxlsx)

confirmed <- read.csv("time_series_covid19_confirmed_global.csv",header=TRUE,
stringsAsFactors = FALSE)
confirmed[1:2,1:6]
```

```
##   Province.State Country.Region      Lat     Long X1.22.20 X1.23.20
## 1                  Afghanistan 33.93911 67.70995        0        0
## 2                      Albania 41.15330 20.16830        0        0
```

```
deaths <- read.csv("time_series_covid19_deaths_global.csv",header=TRUE,
stringsAsFactors = FALSE)
deaths[1:2,((ncol(deaths)-5):ncol(deaths))]
```

```
##   X9.10.20 X9.11.20 X9.12.20 X9.13.20 X9.14.20 X9.15.20
## 1     1420     1420     1420     1420     1425     1426
## 2      324      327      330      334      338      340
```

2.3 Download and load abridged life tables

Please go to the UNWPP2019 website, download abridged life tables for both sexes together, save them into your project folder, and then load them into R.

```
lt_1950_2020 <- read.xlsx("WPP2019_MORT_F17_1_ABRIDGED_LIFE_TABLE_BOTH_SEXES.xlsx",
sheet = 1,startRow = 17)
```

Brief data description. The data object *lt_1950_2020* contains abridged life tables for both sexes for all UN countries.

Explore this data object and find out how large Finnish remaining life expectancy at birth has been 1950-55 through 2015-19.

```
## lt_1950_2020[1:2,]

colnames(lt_1950_2020)
```

```
##  [1] "Index"
##  [2] "Variant"
##  [3] "Region,.subregion,.country.or.area.*"
##  [4] "Notes"
##  [5] "Country.code"
##  [6] "Type"
##  [7] "Parent.code"
##  [8] "Period"
##  [9] "Age.(x)"
## [10] "Age.interval.(n)"
## [11] "Central.death.rate.m(x,n)"
## [12] "Probability.of.dying.q(x,n)"
## [13] "Probability.of.surviving.p(x,n)"
## [14] "Number.of.survivors.l(x)"
## [15] "Number.of.deaths.d(x,n)"
## [16] "Number.of.person-years.lived.L(x,n)"
## [17] "Survival.ratio.S(x,n)"
## [18] "Person-years.lived.T(x)"
## [19] "Expectation.of.life.e(x)"
## [20] "Average.number.of.years.lived.a(x,n)"
```

```
lt_1950_2020[which(lt_1950_2020[,"Region,.subregion,.country.or.area.*"]=="Finland" &
  lt_1950_2020["Age.(x)"]==0),c("Period","Expectation.of.life.e(x)")]
```

```
##          Period Expectation.of.life.e(x)
## 4823  1950-1955       66.401073999999994
## 10349 1955-1960       68.189361000000005
## 15875 1960-1965       69.067266000000004
## 21401 1965-1970       69.716247999999993
## 26927 1970-1975       70.930030000000002
## 32453 1975-1980       72.714505000000003
## 37979 1980-1985       74.326365999999993
## 43505 1985-1990       74.788223000000002
## 49031 1990-1995       75.841427999999993
## 54557 1995-2000       77.144569000000004
## 60083 2000-2005       78.400789000000003
## 65609 2005-2010       79.548135000000002
## 71135 2010-2015       80.705350999999993
## 76661 2015-2020       81.646075999999994
```

2.4 Load global pattern over age of COVID-19 deaths

Dudel et al. (2020) provide data on age-specific death counts attributable to COVID-19 in their supplementary material. These data have served as a basis for calculating a global average pattern over age for total death counts as input for the demographic scaling model. You can download this *global average pattern over age* from the GitHub repository for this course.

```
global_age_dist_deaths <- source("global_age_dist_deaths.R")
## global_age_dist_deaths
```

Brief data description. The data object *global_age_dist_deaths* contains the global pattern over 10-year age groups of COVID-19 deaths.

Note that we follow here the original methodology of the demographic scaling model that has been introduced in the paper of Bohk-Ewald et al. (2020). Another way is to use COVID-19-related death counts by age that have been reported to the COVerAGE database (as presented in course week 4 on Monday).

2.5 Load infection fatality rates from Verity et al. (2020)

Verity and colleagues (2020) report infection fatality rates by 10-year age groups for Hubei province, China, on page 5. Please create a data object *ifr_by_age_china_verity* that contains these data or download it from the GitHub repository for this course.

```
ifr_by_age_china_verity <- read.table("infection-fatality-rates-by-age-china-Verity.txt",
header=FALSE, stringsAsFactors = FALSE)

ifr_by_age_china_verity
```

```
##   V1      V2       V3       V4
## 1  0 1.6e-05 1.85e-06 0.000249
## 2 10 7.0e-05 1.50e-05 0.000500
## 3 20 3.1e-04 1.40e-04 0.000920
## 4 30 8.4e-04 4.10e-04 0.001850
## 5 40 1.6e-03 7.60e-04 0.003200
## 6 50 6.0e-03 3.40e-03 0.013000
## 7 60 1.9e-02 1.10e-02 0.039000
## 8 70 4.3e-02 2.50e-02 0.084000
## 9 80 7.8e-02 3.80e-02 0.133000
```

Brief data description. The data object *ifr_by_age_china_verity* contains the modal estimate as well as the lower and upper bound of the 95 percent credible interval of the infection fatality rates of Hubei province, China, by 10-year age groups.

2.6 Load infection fatality rates from Salje et al. (2020)

Salje and colleagues (2020) report infection fatality rates by 10-year age groups for France. Please create a data object *ifr_by_age_france_salje* that contains these data or download it from the GitHub repository for this course.

```
ifr_by_age_france_salje <- read.table("infection-fatality-rates-by-age-france-Salje.txt",
header=FALSE, stringsAsFactors = FALSE)

ifr_by_age_france_salje <- ifr_by_age_france_salje/100

ifr_by_age_france_salje
```

```
##    V1      V2      V3      V4
## 1 0.0 0.00001 1.0e-06 0.00002
## 2 0.1 0.00001 1.0e-06 0.00002
## 3 0.2 0.00007 3.0e-05 0.00010
## 4 0.3 0.00020 1.0e-04 0.00040
## 5 0.4 0.00060 3.0e-04 0.00090
## 6 0.5 0.00200 1.0e-03 0.00360
## 7 0.6 0.00900 5.0e-03 0.01400
## 8 0.7 0.02400 1.4e-02 0.03700
## 9 0.8 0.10100 6.0e-02 0.15600
```

Brief data description. The data object *ifr_by_age_france_salje* contains the modal estimate as well as the lower and upper bound of the 95 percent credible interval of the infection fatality rates of France by 10-year age groups.

2.7 Infection fatality rates from Levin et al. (2020)

Levin and colleagues (2020) propose an exponential model function to determine infection fatality rates by single years of age. Please create a data object *ifr_by_age_levin* that contains these data for 10-year age groups.

```
ifr_by_age_levin <- c(0)
current_ifr_sum <- c(0)
exp_IFR <- exp(-7.53 + 0.119 * (seq(0.5,90.5,1))) / 100
for(group in 1:8){
    pos <- (1+10*(group-1)):(10+10*(group-1))
    current_ifr_sum[group] <- sum(exp_IFR[pos])
    ifr_by_age_levin[group] <- current_ifr_sum[group]/10
} ## group
current_ifr_sum[9] <- sum((exp_IFR[(pos[length(pos)]+1):(length(exp_IFR))]))
ifr_by_age_levin[9] <- current_ifr_sum[9]/length((pos[length(pos)]+1):(length(exp_IFR)))

round(ifr_by_age_levin,5)
```

```
## [1] 0.00001 0.00003 0.00011 0.00037 0.00120 0.00396 0.01300 0.04275 0.15094
```

Brief data description. The data object *ifr_by_age_levin* contains the central estimate of infection fatality rates by 10-year age groups.

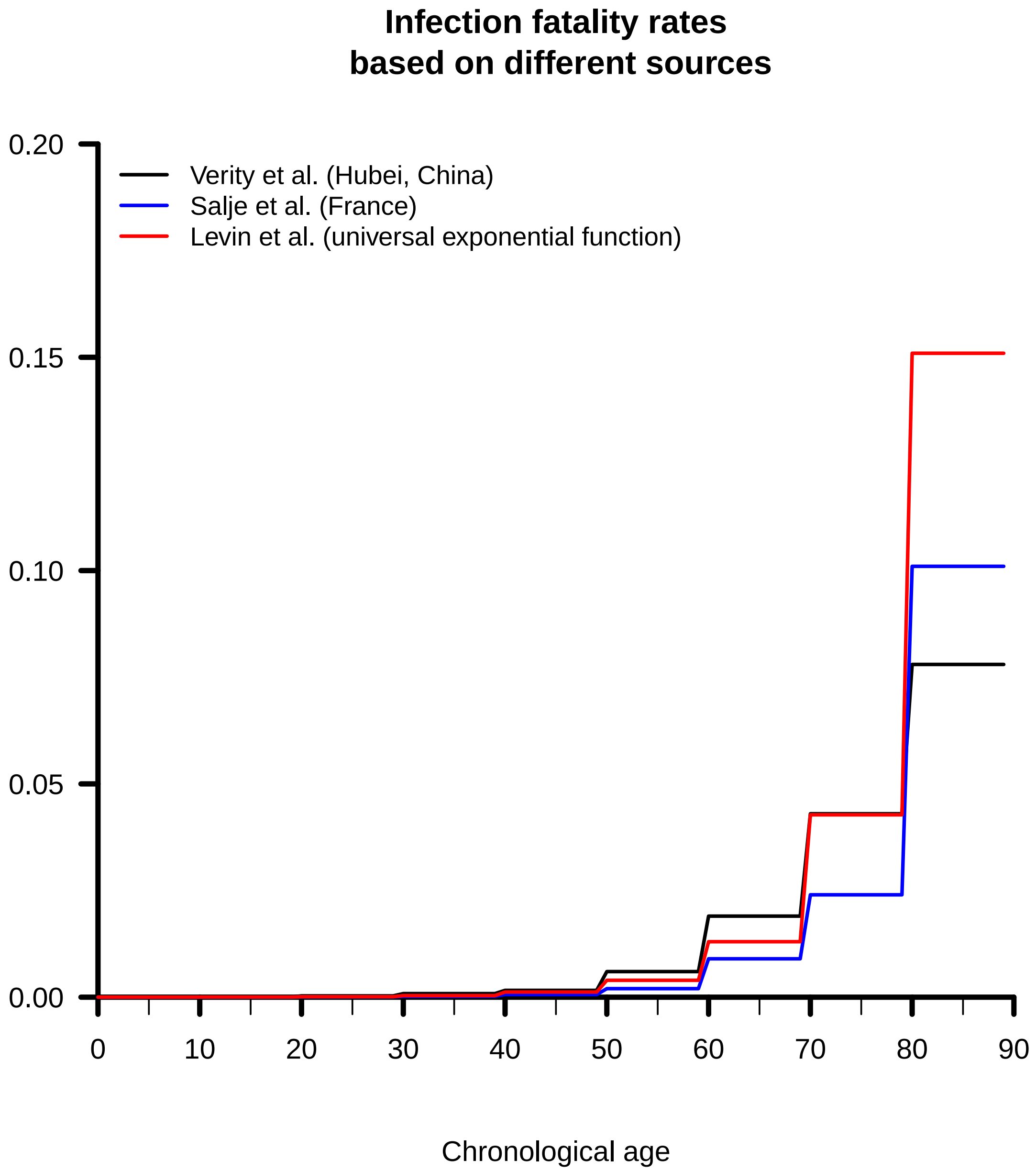
3. Analyze and compare IFR estimates from different sources

3.1. Visualize original IFR estimates from different sources

An easy way to compare the IFR estimates from Verity et al. (2020), Salje et al. (2020), and Levin et al. (2020), is to plot them:

```
par(fig = c(0,1,0,1), las=1, mai=c(0.8,0.8,0.8,0.4))

plot(x=-100,y=-100,xlim=c(0,90),ylim=c(0,0.2),xlab="Chronological age",ylab="",
    main="Infection fatality rates\n based on different sources",axes=FALSE)

axis(side=1,at=seq(0,90,5),labels=FALSE,lwd=1,pos=0)
axis(side=1,at=seq(0,90,10),labels=TRUE,lwd=3,pos=0)
axis(side=2,at=seq(0,0.2,0.05),labels=TRUE,lwd=3,pos=0)

lines(0:89,rep(ifr_by_age_china_verity[,2],each=10),col="black",lwd=2)
lines(0:89,rep(ifr_by_age_france_salje[,2],each=10),col="blue",lwd=2)
lines(0:89,rep(ifr_by_age_levin,each=10),col="red",lwd=2)
```

```
legend(0,0.2,c("Verity et al. (Hubei, China)","Salje et al. (France)",
"Levin et al. (universal exponential function)"),col=c("black","blue","red"),
bty="n",lwd=2,lty=1,cex=0.9)
```



Please describe and compare the IFR estimates of Verity et al. (2020), Salje et al. (2020), and Levin et al. (2020). What IFR estimates are particularly low in younger ages, and what IFR estimates are particularly large in older ages?

What do you think about these IFR estimates? How plausible are they?

3.2 Scale original IFR estimates from a reference country onto a country of interest

You can now scale the original IFR estimates of Verity et al. (2020) and Salje et al. (2020) to better match the context in Finland with respect to age structure, preconditions, and medical services, adopting the scaling procedure based on remaining life expectancy of the demographic scaling model.

3.2.1 Source basic functions of the demographic scaling model You can find the basic functions of the demographic scaling model in the file *basic-functions-week-5.R* in the GitHub repository for this course. They contain the functions:

- *to_ungroup* and *to_ungroup_spar* to interpolate IFR estimates of Verity et al. (2020) and Salje et al. (2020) into single years of age
- *get_ungrouped_ex_2015_2020* to ungroup remaining life expectancy
- *map_ifr_betw_ref_and_one_coi_thanatAge* and *map_ifr_betw_assigned_ref_and_one_coi_thanatAge* to scale IFRs from a reference country (here: China; France) onto a country of interest (here: Finland) based on remaining life expectancy
- *aggregate_mapped_ifr_10y* to aggregate scaled IFRs into 10-year age groups
- *disaggregate_deaths_one_coi_10y* to disaggregate total deaths into 10-year age

You may have a look at these basic functions if you wish. But it is also fine to just source (or load) them via the file *basic-functions-week-3.R*:

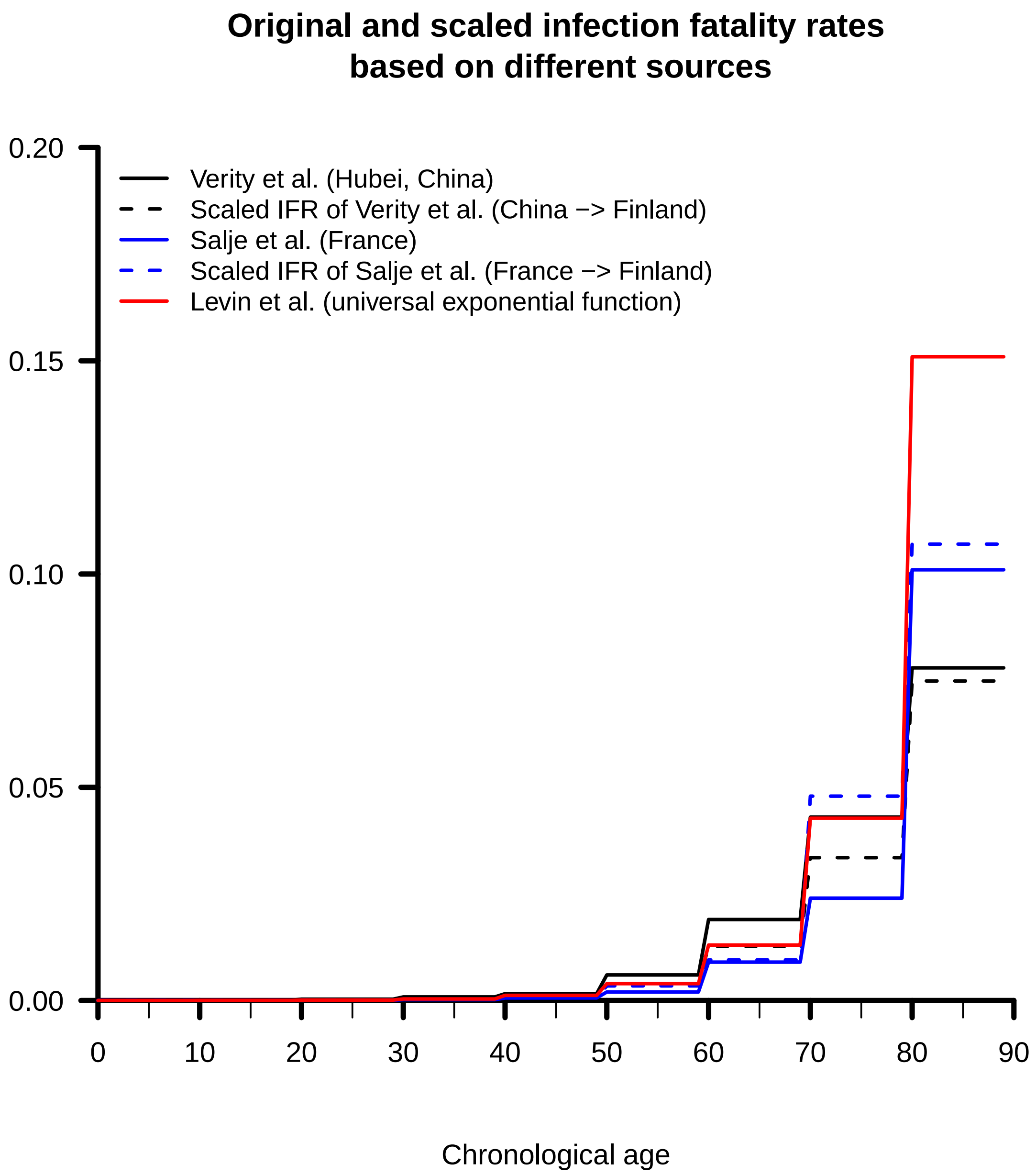
```r
source("basic-functions-week-5.R")
```

3.2.2 Scale original IFR estimates of Verity et al. (2020) and Salje et al. (2020) for Finland You can scale the original IFR estimates based on remaining life expectancy following the corresponding steps of the demograhic scaling model:

```r
#
## 1. Ungroup original IFR estimates:
#

ungrouped_mode_ifr_by_single_age_china_sp <- to_ungroup(to_ungroup=
        ifr_by_age_china_verity[,2],nr_grouped_years=10)

ungrouped_mode_ifr_by_single_age_france_sp <- to_ungroup_spar(to_ungroup=
        ifr_by_age_france_salje[,2],spar=0.195,nr_grouped_years=10)

#
## 2. Scale original IFRs onto a COI (here: Finland) via remaining life expectancy:
#

mapped_mode_ifr_thanatAge_verity <- map_ifr_betw_ref_and_one_coi_thanatAge(coi="Finland",
    lt_1950_2020=lt_1950_2020,
    ungrouped_ifr_by_single_age_china_sp=ungrouped_mode_ifr_by_single_age_china_sp)

mapped_mode_ifr_thanatAge_salje <- map_ifr_betw_assigned_ref_and_one_coi_thanatAge(ref="France",
    coi="Finland",deaths=deaths,lt_1950_2020=lt_1950_2020,
    ungrouped_ifr_by_single_age_china_sp=ungrouped_mode_ifr_by_single_age_france_sp)

## and fill in the few NA values:

pos_na <- which(is.na(mapped_mode_ifr_thanatAge_verity[1,]))
    if(length(pos_na)>0){
        for(pos in 1:length(pos_na)){
            if(pos_na[pos] < 6){
                mapped_mode_ifr_thanatAge_verity[1,pos_na[pos]] <-
                    min(mapped_mode_ifr_thanatAge_verity[1,],na.rm=TRUE)
            }
            if(pos_na[pos] >= 6){
                mapped_mode_ifr_thanatAge_verity[1,pos_na[pos]] <-
                    mapped_mode_ifr_thanatAge_verity[1,pos_na[pos]-1]
            }
        } ## for pos
    } ## if

pos_na <- which(is.na(mapped_mode_ifr_thanatAge_salje[1,]))
    if(length(pos_na)>0){
        for(pos in 1:length(pos_na)){
            if(pos_na[pos] < 6){
                mapped_mode_ifr_thanatAge_salje[1,pos_na[pos]] <-
                    min(mapped_mode_ifr_thanatAge_salje[1,],na.rm=TRUE)
            }
            if(pos_na[pos] >= 6){
                mapped_mode_ifr_thanatAge_salje[1,pos_na[pos]] <-
                    mapped_mode_ifr_thanatAge_salje[1,pos_na[pos]-1]
            }
        } ## for pos
    } ## if

#
## 3. Put scaled IFRs into 10-year age groups:
#

mapped_mode_ifr_thanatAge_verity_10y <- aggregate_mapped_ifr_10y(disaggregated_mapped_ifr=
                        mapped_mode_ifr_thanatAge_verity)

mapped_mode_ifr_thanatAge_salje_10y <- aggregate_mapped_ifr_10y(disaggregated_mapped_ifr=
                        mapped_mode_ifr_thanatAge_salje)
```

3.2.3 Visualize original and scaled IFR estimates from different sources You can now depict the original IFR estimates from different sources and compare them with the scaled IFR estimates for Finland:

```
par(fig = c(0,1,0,1), las=1, mai=c(0.8,0.8,0.8,0.4))

plot(x=-100,y=-100,xlim=c(0,90),ylim=c(0,0.2),xlab="Chronological age",ylab="",
    main="Original and scaled infection fatality rates\n based on different sources",
    axes=FALSE)

axis(side=1,at=seq(0,90,5),labels=FALSE,lwd=1,pos=0)
axis(side=1,at=seq(0,90,10),labels=TRUE,lwd=3,pos=0)
axis(side=2,at=seq(0,0.2,0.05),labels=TRUE,lwd=3,pos=0)

lines(0:89,rep(ifr_by_age_china_verity[,2],each=10),col="black",lwd=2)
lines(0:89,rep(mapped_mode_ifr_thanatAge_verity_10y,each=10),col="black",lty=2,lwd=2)

lines(0:89,rep(ifr_by_age_france_salje[,2],each=10),col="blue",lwd=2)
lines(0:89,rep(mapped_mode_ifr_thanatAge_salje_10y,each=10),col="blue",lty=2,lwd=2)

lines(0:89,rep(ifr_by_age_levin,each=10),col="red",lwd=2)

legend(0,0.2,c("Verity et al. (Hubei, China)",
"Scaled IFR of Verity et al. (China -> Finland)",
"Salje et al. (France)","Scaled IFR of Salje et al. (France -> Finland)",
"Levin et al. (universal exponential function)"),
col=c("black","black","blue","blue","red"),
bty="n",lwd=2,lty=c(1,2,1,2,1),cex=0.9)
```



Please describe and compare the original and scaled IFR estimates (for Finland). How do you explain the differences between original and scaled IFRs: China → Finland and France → Finland?

4. Estimate COVID-19 infections in Finland based on different IFR estimates

You can now estimate the COVID-19 infections in Finland over time based on these different IFR estimates, following the steps of the demographic scaling model.

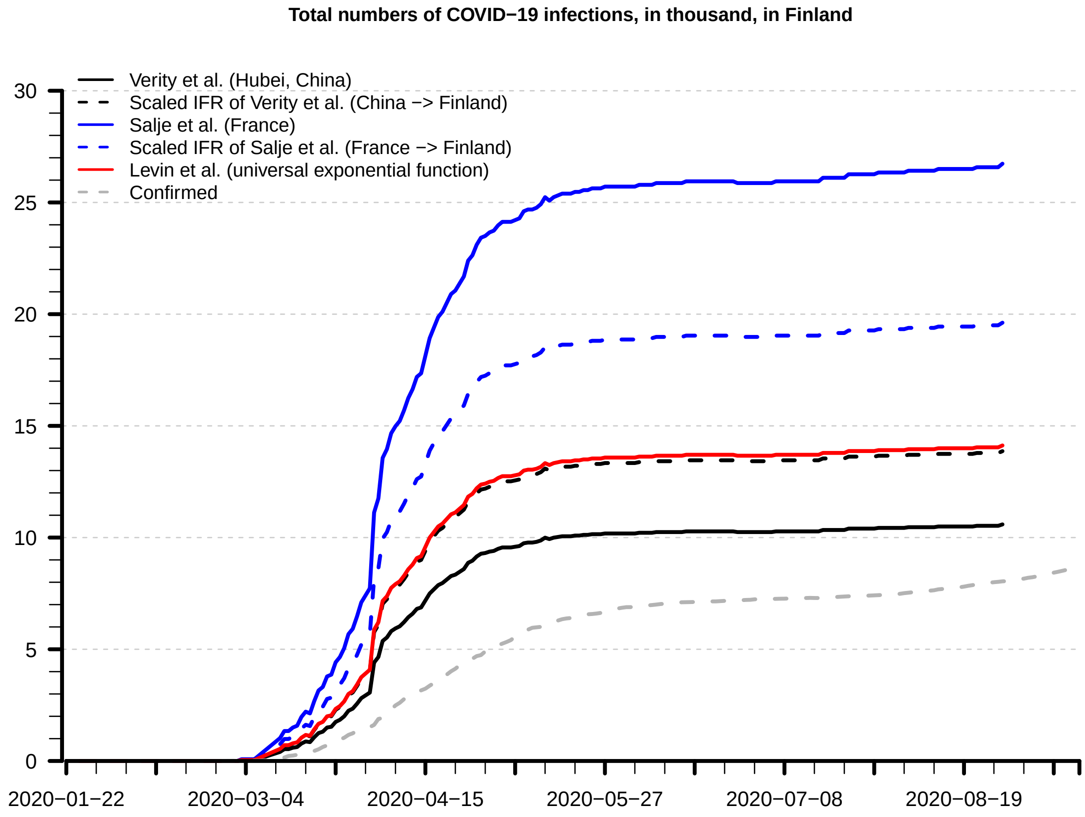
```
#
## 1. Disaggregate total COVID-19-related deaths into 10-year age groups:
#

deaths_by_age <- disaggregate_deaths_one_coi_10y(coi="Finland")

#
## 2. Estimate COVID-19 infections over time:
#

inf_mode_scaled_verity <- colSums( deaths_by_age / mapped_mode_ifr_thanatAge_verity_10y )
inf_mode_scaled_salje <- colSums( deaths_by_age / mapped_mode_ifr_thanatAge_salje_10y )
inf_mode_levin <- colSums( deaths_by_age / ifr_by_age_levin )
inf_mode_verity <- colSums( deaths_by_age / ifr_by_age_china_verity[,2] )
inf_mode_salje <- colSums( deaths_by_age / ifr_by_age_france_salje[,2] )
```

The data objects *inf_mode_levin*, *inf_mode_verity*, and *inf_mode_salje* contain the COVID-19 infection for Finland estimates based on the central estimates of the original IFR estimates of Levin et al., Verity et al., and Salje et al., respectively. The data objects *inf_mode_scaled_verity* and *inf_mode_scaled_salje* contain the COVID-19 infection estimates for Finland based on the scaled central estimates of the IFR for China and France, respectively.

You can now visualize the numbers of COVID-19 infections in Finland over time, based on different IFR estimates. When comparing these Finnish COVID-19 estimates with the numbers of confirmed cases, it is a good idea to also account for an average time to death of approximately 18 days.

```
dates <- seq(as.Date("22/01/2020", format = "%d/%m/%Y"),
by = "days", length = (ncol(deaths)-4) )
```

```
par(fig = c(0,1,0,1), las=1, mai=c(0.4,0.8,0.8,0.4))

plot(x=-100,y=-100,xlim=c(0,length(5:ncol(deaths))),ylim=c(0,30),xlab="Date",ylab="",
cex.main=0.9,main="Total numbers of COVID-19 infections, in thousand, in Finland",axes=FALSE)

segments(x0=rep(0,4),x1=rep(length(5:ncol(deaths)),4),y0=seq(5,30,5),y1=seq(5,30,5),
lty=2,col=grey(0.8))

lines(x=1:length(5:ncol(deaths)),y=c(inf_mode_scaled_verity[-c(1:18)],rep(NA,18))/1000,
col="black",lty=2,lwd=3)
lines(x=1:length(5:ncol(deaths)),y=c(inf_mode_verity[-c(1:18)],rep(NA,18))/1000,
col="black",lty=1,lwd=3)

lines(x=1:length(5:ncol(deaths)),y=c(inf_mode_scaled_salje[-c(1:18)],rep(NA,18))/1000,
col="blue",lty=2,lwd=3)
lines(x=1:length(5:ncol(deaths)),y=c(inf_mode_salje[-c(1:18)],rep(NA,18))/1000,
col="blue",lty=1,lwd=3)

lines(x=1:length(5:ncol(deaths)),y=c(inf_mode_levin[-c(1:18)],rep(NA,18))/1000,
col="red",lty=1,lwd=3)


lines(x=1:length(5:ncol(deaths)),y=confirmed[which(confirmed[,"Country.Region"]=="Finland"),
5:ncol(confirmed)]/1000,col=gray(0.7),lty=2,lwd=3)

legend(0,31.5,c("Verity et al. (Hubei, China)",
"Scaled IFR of Verity et al. (China -> Finland)",
"Salje et al. (France)","Scaled IFR of Salje et al. (France -> Finland)",
"Levin et al. (universal exponential function)",
"Confirmed"),
col=c("black","black","blue","blue","red",gray(0.7)),
bty="n",lwd=2,lty=c(1,2,1,2,1,2),cex=0.9)

axis(side=1,at=seq(1,length(5:ncol(deaths)),7),labels=FALSE,lwd=1,pos=0)

axis(side=1,at=c(seq(1,length(5:ncol(deaths)),21),length(5:ncol(deaths))),
labels=dates[c(seq(1,length(5:ncol(deaths)),21),
length(5:ncol(deaths)))],lwd=3,pos=0)

axis(side=2,at=seq(0,30,1),labels=FALSE,lwd=1,pos=0)

axis(side=2,at=seq(0,30,5),labels=TRUE,lwd=3,pos=0)
```



Please describe and compare the level and the temporal development of these estimated numbers of COVID-19 infections in Finland. Please also compare them to (1) the numbers of confirmed cases in Finland and, perhaps, to (2) the corresponding figures in other countries.

Please think about how you could extend this R-code in order to estimate the total numbers of COVID-19 infections for other countries and how to provide uncertainty estimates for them.

4. Time for you to think both creatively and critically about the robustness of the COVID-19 infection estimates in general and for Finland in particular.

How would you evaluate the process of the demographic scaling model?

How robust are the COVID-19 infection estimates with respect to taking infection fatality rates from different sources?

How plausible are the COVID-19 infection estimates, also considering the time (in)variance of input data?

In what direction would the infection fatality rates need to be adjusted in order to better account for different settings or conditions during the coronavirus pandemic?