COS-D407. *Scientific Modeling and Model Validation*

Hands-on exercises

Week 2

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 2 in R (e.g., *week-2.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 explore COVID-19 data

In week 2 we explore trends of the COVID-19 pandemic. We will start with downloading freely available data on COVID-19 for multiple countries. We will then continue analyzing numbers and trends of COVID-19-related cases, deaths, and case fatality rates, and finally start to come up with alternative explanations for possible cross-country differences through practicing creative and critical thinking.

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

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[1:2,1:10]
```

```
##   Province.State Country.Region      Lat     Long X1.22.20 X1.23.20 X1.24.20
## 1                    Afghanistan 33.93911 67.70995        0        0        0
## 2                        Albania 41.15330 20.16830        0        0        0
##   X1.25.20 X1.26.20 X1.27.20
## 1        0        0        0
## 2        0        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
```

Describe these data. For which countries and states are they available, for which dates are they available?

2.3 Explore data objects *confirmed* and *deaths*.

How many confirmed cases and reported deaths are there for Italy and for China most recently?

```
confirmed[which(deaths[,"Country.Region"]=="Italy"),c(1:4,ncol(confirmed))]
```

```
##     Province.State Country.Region      Lat     Long X9.15.20
## 150                         Italy 41.87194 12.56738   289990
```

```
deaths[which(deaths[,"Country.Region"]=="Italy"),c(1:4,ncol(deaths))]
```

```
##     Province.State Country.Region      Lat     Long X9.15.20
## 150                         Italy 41.87194 12.56738    35633
```

```
confirmed[which(deaths[,"Country.Region"]=="China"),c(1:4,ncol(confirmed))]
```

```
##    Province.State Country.Region     Lat     Long X9.15.20
## 57          Anhui          China 31.8257 117.2264      991
## 58        Beijing          China 40.1824 116.4142      935
## 59      Chongqing          China 30.0572 107.8740      584
## 60         Fujian          China 26.0789 117.9874      392
## 61          Gansu          China 35.7518 104.2861      170
## 62      Guangdong          China 23.3417 113.4244     1783
## 63        Guangxi          China 23.8298 108.7881      258
## 64        Guizhou          China 26.8154 106.8748      147
## 65         Hainan          China 19.1959 109.7453      171
## 66          Hebei          China 39.5490 116.1306      365
## 67   Heilongjiang          China 47.8620 127.7615      948
## 68          Henan          China 37.8957 114.9042     1277
## 69      Hong Kong          China 22.3000 114.2000     4975
## 70          Hubei          China 30.9756 112.2707    68139
## 71          Hunan          China 27.6104 111.7088     1019
## 72 Inner Mongolia          China 44.0935 113.9448      261
## 73        Jiangsu          China 32.9711 119.4550      665
## 74        Jiangxi          China 27.6140 115.7221      935
## 75          Jilin          China 43.6661 126.1923      157
## 76       Liaoning          China 41.2956 122.6085      264
## 77          Macau          China 22.1667 113.5500       46
## 78        Ningxia          China 37.2692 106.1655       75
## 79        Qinghai          China 35.7452  95.9956       18
## 80        Shaanxi          China 35.1917 108.8701      382
## 81       Shandong          China 36.3427 118.1498      831
## 82       Shanghai          China 31.2020 121.4491      950
## 83         Shanxi          China 37.5777 112.2922      203
## 84        Sichuan          China 30.6171 102.7103      670
## 85        Tianjin          China 39.3054 117.3230      234
## 86          Tibet          China 31.6927  88.0924        1
## 87       Xinjiang          China 41.1129  85.2401      902
## 88         Yunnan          China 24.9740 101.4870      205
## 89       Zhejiang          China 29.1832 120.0934     1282
```

```
deaths[which(deaths[,"Country.Region"]=="China"),c(1:4,ncol(deaths))]
```

```
##    Province.State Country.Region     Lat     Long X9.15.20
```

```
## 57          Anhui          China 31.8257 117.2264        6
## 58        Beijing          China 40.1824 116.4142        9
## 59      Chongqing          China 30.0572 107.8740        6
## 60         Fujian          China 26.0789 117.9874        1
## 61          Gansu          China 35.7518 104.2861        2
## 62      Guangdong          China 23.3417 113.4244        8
## 63        Guangxi          China 23.8298 108.7881        2
## 64        Guizhou          China 26.8154 106.8748        2
## 65         Hainan          China 19.1959 109.7453        6
## 66          Hebei          China 39.5490 116.1306        6
## 67   Heilongjiang          China 47.8620 127.7615       13
## 68          Henan          China 37.8957 114.9042       22
## 69      Hong Kong          China 22.3000 114.2000      102
## 70          Hubei          China 30.9756 112.2707     4512
## 71          Hunan          China 27.6104 111.7088        4
## 72 Inner Mongolia          China 44.0935 113.9448        1
## 73        Jiangsu          China 32.9711 119.4550        0
## 74        Jiangxi          China 27.6140 115.7221        1
## 75          Jilin          China 43.6661 126.1923        2
## 76       Liaoning          China 41.2956 122.6085        2
## 77          Macau          China 22.1667 113.5500        0
## 78        Ningxia          China 37.2692 106.1655        0
## 79        Qinghai          China 35.7452  95.9956        0
## 80        Shaanxi          China 35.1917 108.8701        3
## 81       Shandong          China 36.3427 118.1498        7
## 82       Shanghai          China 31.2020 121.4491        7
## 83         Shanxi          China 37.5777 112.2922        0
## 84        Sichuan          China 30.6171 102.7103        3
## 85        Tianjin          China 39.3054 117.3230        3
## 86          Tibet          China 31.6927  88.0924        0
## 87       Xinjiang          China 41.1129  85.2401        3
## 88         Yunnan          China 24.9740 101.4870        2
## 89       Zhejiang          China 29.1832 120.0934        1
```

```
sum(confirmed[which(confirmed[,"Country.Region"]=="China"),ncol(confirmed)])
```

```
## [1] 90235
```

```
sum(deaths[which(deaths[,"Country.Region"]=="China"),ncol(deaths)])
```

```
## [1] 4736
```

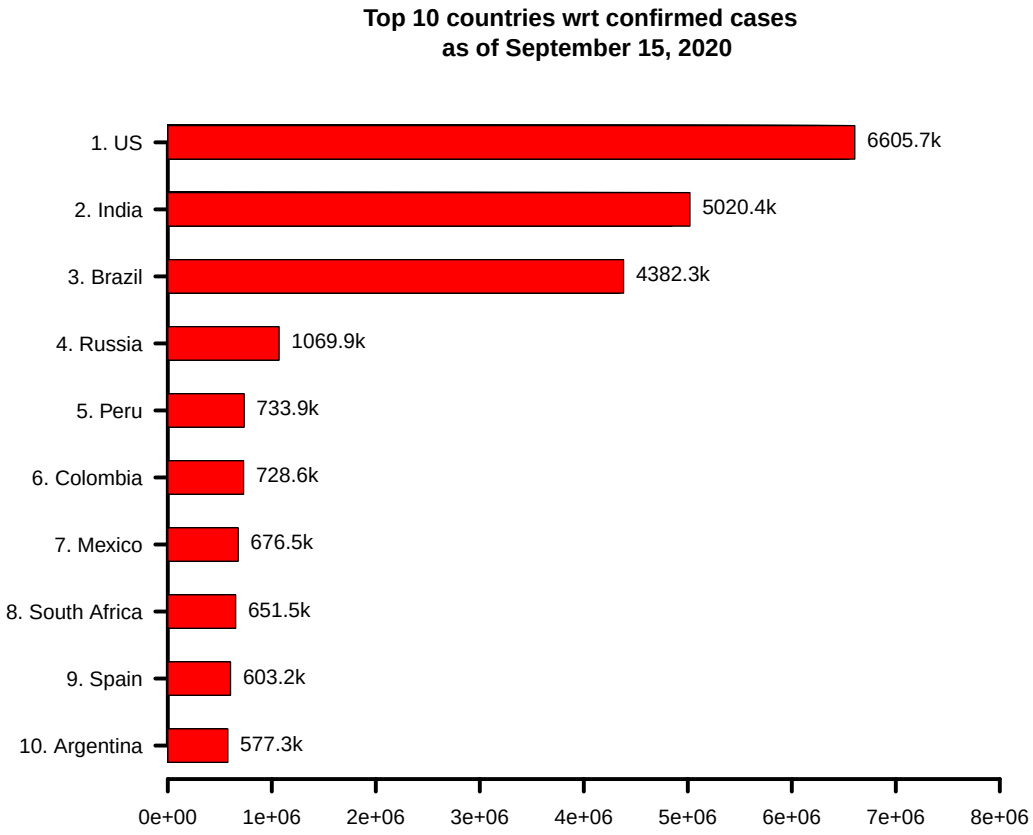
3. Plot confirmed cases and reported deaths attributable to COVID-19

We now want to visualize the numbers of confirmed cases and reported deaths from COVID-19. We focus on the ten countries with the most confirmed cases or reported deaths so far.

We start with the numbers of confirmed cases:

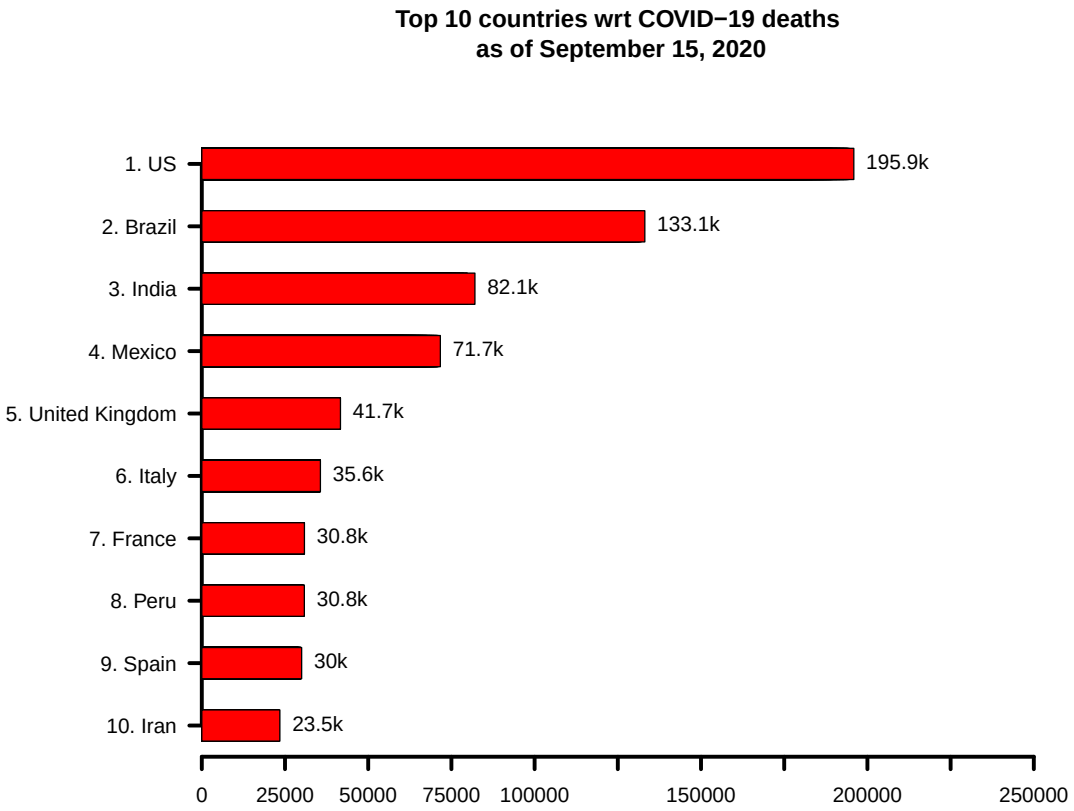
```
par(fig = c(0,1,0,1), las=1, mai=c(0.4,2.4,0.8,0.4))

plot(x=-100,y=-100,xlim=c(0,8000000),ylim=c(0,10),xlab="",ylab="",
    main="Top 10 countries wrt confirmed cases \n as of September 15, 2020",axes=FALSE)

country_labels <- c(0)
for(pop in 1:10){
```

```
    current_pop <- confirmed[order(confirmed[,ncol(confirmed)],decreasing=TRUE),][pop,1:2]
    if(!current_pop["Province.State"]==''){
        country_labels[pop] <- current_pop["Province.State"]
    }
    if(current_pop["Province.State"]==''){
        country_labels[pop] <- current_pop["Country.Region"]
    }
}

axis(side=1,at=seq(0,8000000,1000000),labels=TRUE,lwd=3,pos=0)
axis(side=2,at=seq(0.5,9.5,1),
labels=paste(rev(seq(1,10,1)),". ",rev(country_labels),sep=""),lwd=3,pos=0)

for(pop in 1:10){
    rect(xleft=0,xright=confirmed[order(confirmed[,ncol(confirmed)],
    decreasing=TRUE),][pop,5:ncol(confirmed)],ybottom=9.25-1*(pop-1),
    ytop=9.25-1*(pop-1)+0.5,col="red")

    text(confirmed[order(confirmed[,ncol(confirmed)],decreasing=TRUE),][pop,ncol(confirmed)],
    9.25-1*(pop-1)+0.25,paste(round(confirmed[order(confirmed[,ncol(confirmed)],
    decreasing=TRUE),][pop,ncol(confirmed)]/1000,1),"k",sep=""),pos=4)
}
```



. . . and continue with the numbers of reported deaths:

```
par(fig = c(0,1,0,1), las=1, mai=c(0.4,2.4,1.2,0.4))

    plot(x=-100,y=-100,xlim=c(0,250000),ylim=c(0,10),xlab="",ylab="",
        main="Top 10 countries wrt COVID-19 deaths\n as of September 15, 2020",axes=FALSE)

    country_labels <- c(0)
    country_row_number <- c(NA)
    for(pop in 1:10){
        current_pop <- deaths[order(deaths[,ncol(deaths)],decreasing=TRUE),][pop,1:2]
        country_row_number[pop] <- rownames(current_pop)
        if(!current_pop["Province.State"]==''){
            country_labels[pop] <- current_pop["Province.State"]
        }
        if(current_pop["Province.State"]==''){
            country_labels[pop] <- current_pop["Country.Region"]
        }
    }

    axis(side=1,at=seq(0,250000,25000),labels=TRUE,lwd=3,pos=0)
    axis(side=2,at=seq(0.5,9.5,1),labels=paste(rev(seq(1,10,1)),". ",
    rev(country_labels),sep=""),lwd=3,pos=0)

    for(pop in 1:10){
        rect(xleft=0,xright=deaths[order(deaths[,ncol(deaths)],decreasing=TRUE),]
        [pop,5:ncol(deaths)],ybottom=9.25-1*(pop-1),ytop=9.25-1*(pop-1)+0.5,col="red")

        text(deaths[order(deaths[,ncol(deaths)],decreasing=TRUE),][pop,ncol(deaths)],
        9.25-1*(pop-1)+0.25,paste(round(deaths[order(deaths[,ncol(deaths)],
        decreasing=TRUE),][pop,ncol(deaths)]/1000,1),"k",sep=""),pos=4)
    }
```



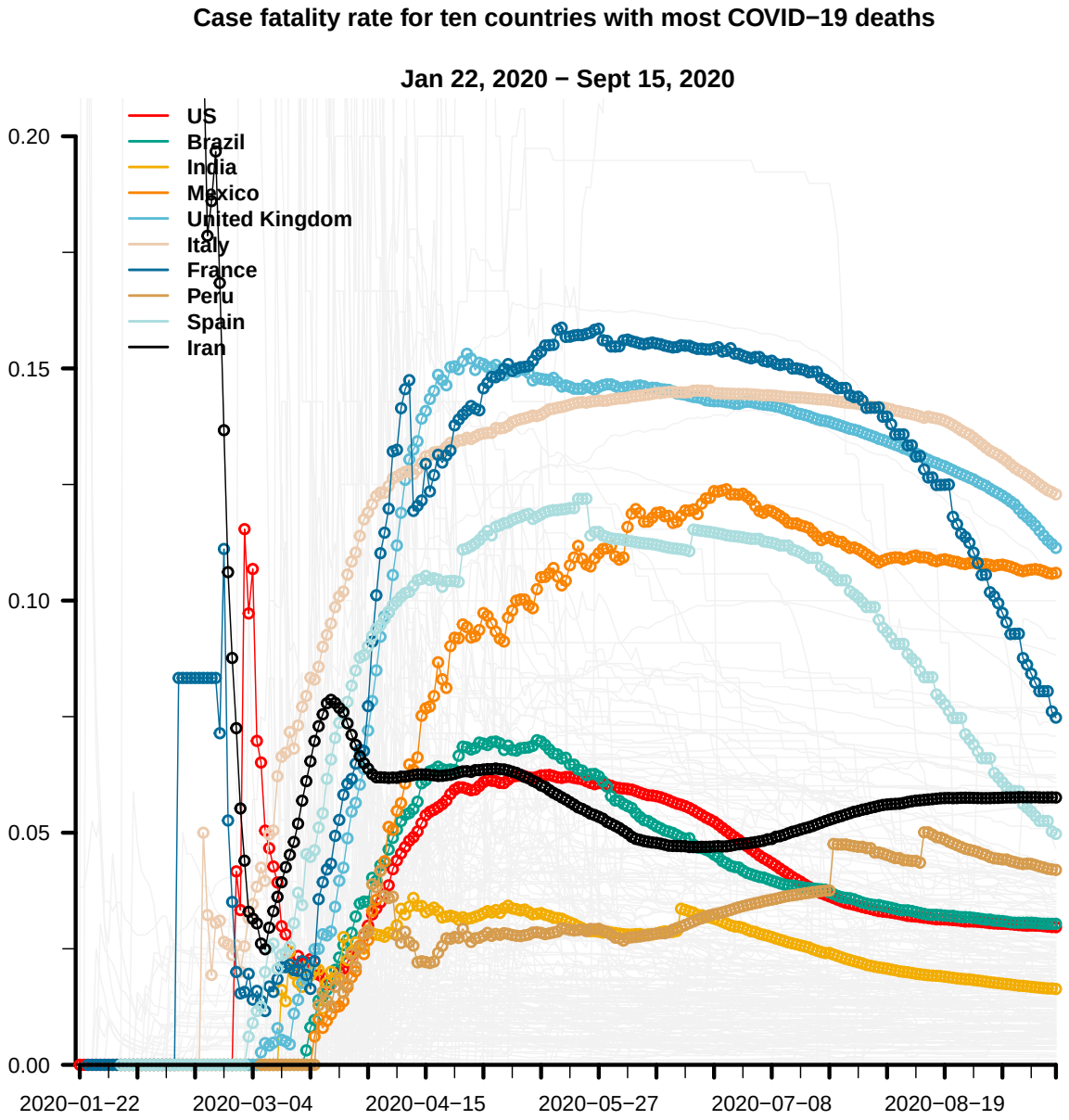
Compare the ranking of the top ten countries with respect to most confirmed cases and reported deaths. What similarities and differences do you observe?

4. Calculate and plot case fatality rates attributable to COVID-19

We now want to calculate and visualize the case fatality rates over time and highlight the development for the ten countries with most reported COVID-19 deaths.

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

cfr <- as.matrix(deaths[,5:length(deaths)] / confirmed[,5:length(confirmed)])
cfr[is.nan(cfr)] <- NA
```

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

  require(wesanderson)
  pal <- c(wes_palette("Darjeeling1"),wes_palette("Darjeeling2"))

  plot(x=-100,y=-100,xlim=c(1,length(dates)),ylim=c(0,0.2),xlab="Date",ylab="",
      main="Case fatality rate for ten countries with most COVID-19 deaths
          \nJan 22, 2020 - Sept 15, 2020",axes=FALSE)

  axis(side=2,at=seq(0,0.2,0.025),labels=FALSE,lwd=1,pos=0)
  axis(side=2,at=seq(0,0.2,0.05),labels=TRUE,lwd=3,pos=0)

  for(country in 1:nrow(deaths)){
      lines(x=1:length(dates),y=cfr[country,],col=gray(0.95),lwd=1)
  }


  for(country in 1:10){
      lines(x=1:length(dates),y=cfr[as.numeric(country_row_number[country]),],
          col=pal[country],lwd=1)
      points(x=1:length(dates),y=cfr[as.numeric(country_row_number[country]),],
          col=pal[country],lwd=2)
  }

  axis(side=1,at=seq(1,length(dates),7),labels=FALSE,lwd=1,pos=0)
  axis(side=1,at=c(seq(1,length(dates),14),length(dates)),
      labels=dates[c(seq(1,length(dates),14),length(dates))],lwd=3,pos=0)

  legend(x=length(dates)*0.035,y=0.21,country_labels,
      bty="n",col=c(pal[1:10]),lty=1,lwd=2,text.font=2)
```



Please describe and compare the levels and trends in the case fatality rates attributable to COVID-19 across the countries.

4. Time for you to think both creatively and critically about these empirical findings. What are possible explanations for the large cross-country differences in case fatality rates related to COVID-19?

There are so many things to explore and to think of here.

For example, how reliable are confirmed cases and reported deaths from COVID-19?

As a source of inspiration, you may want to have a look at the papers of Dowd et al. (2020) on *Demographic science aids in understanding the spread and fatality rates of COVID-19* and Dudel et al. (2020) on *Monitoring trends and differences in COVID-19 case-fatality rates using decomposition methods: Contributions of age structure and age-specific fatality*.