BACK-OF-THE-EVELOPE CALCULATIONS SESSIONS

2024-03-05

## 1. Load package

# install.packages("tidyverse")
# install.packages("patchwork")
# install.packages("zoo")

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr 1.1.2 ✔ readr 2.1.4
## ✔ forcats 1.0.0 ✔ stringr 1.5.0
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0
## ✔ purrr 1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag() masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(patchwork)

## Warning: package 'patchwork' was built under R version 4.2.3

library(zoo)

##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
## as.Date, as.Date.numeric

## 2. Load Brazil data

## Set the working directory: update this to the path to the folder you saved the data file in
setwd("/Users/kjia/Documents/DW/")

## Read in the data
WHO\_data <- read.csv("WHO-COVID-19-global-data.csv")
WHO\_data$Date\_reported <- as.Date(WHO\_data$Date\_reported)

## Subset for only Brazil
brazil\_data <- WHO\_data %>% filter(Country == "Brazil")
head(brazil\_data)

## Date\_reported Country\_code Country WHO\_region New\_cases Cumulative\_cases
## 1 2020-01-03 BR Brazil AMRO 0 0
## 2 2020-01-04 BR Brazil AMRO 0 0
## 3 2020-01-05 BR Brazil AMRO 0 0
## 4 2020-01-06 BR Brazil AMRO 0 0
## 5 2020-01-07 BR Brazil AMRO 0 0
## 6 2020-01-08 BR Brazil AMRO 0 0
## New\_deaths Cumulative\_deaths
## 1 0 0
## 2 0 0
## 3 0 0
## 4 0 0
## 5 0 0
## 6 0 0

## 3. Explore raw cases and deaths

## Look at variable names
colnames(brazil\_data)

## [1] "Date\_reported" "Country\_code" "Country"
## [4] "WHO\_region" "New\_cases" "Cumulative\_cases"
## [7] "New\_deaths" "Cumulative\_deaths"

## Have a look at the data
p\_cases <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_reported,y=New\_cases),col="red")+
 geom\_line(aes(x=Date\_reported,y=New\_deaths),col="blue")+
 scale\_y\_continuous(labels = scales::label\_number\_si())

## Warning: `label\_number\_si()` was deprecated in scales 1.2.0.
## ℹ Please use the `scale\_cut` argument of `label\_number()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was
## generated.

p\_cases



# zoom in on deaths
p\_deaths <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_reported,y=New\_deaths),col="blue")+
 scale\_y\_continuous(labels = scales::label\_number\_si())

p\_cases/p\_deaths



### a) What striking features can you discern from visualizing the data? What might generate these?

* Cases and deaths vary within weeks (e.g., in July 2020, more died at the weekend than on Tuesday / Wednesday).
* Throughout the whole period, cases and deaths had similar trends in week-to-week variations. That is, cases and deaths both peaked at July 2020 and at Jan 2021.
* However, at the beginning of the observation period, the death-to-case ratio was higher.

## 4.1 Smooth the death data

* Use moving average to reduce the noise within weeks due to the Day-of-Week effect.

brazil\_data <- brazil\_data %>% mutate(New\_cases\_rollmean=zoo::rollmean(New\_cases,7,na.pad=TRUE))
brazil\_data <- brazil\_data %>% mutate(New\_deaths\_rollmean=zoo::rollmean(New\_deaths,7,na.pad=TRUE))

p\_cases\_smoothed <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_reported,y=New\_cases),col="red", size =0.3)+
 geom\_line(aes(x=Date\_reported,y=New\_cases\_rollmean),col="purple", size =1)+
 scale\_y\_continuous(labels = scales::label\_number\_si())

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was
## generated.

p\_cases\_smoothed

## Warning: Removed 6 rows containing missing values (`geom\_line()`).



## 4.2 Question: can deaths provide a more accurate description of SARS-CoV-2 incidence than cases?

### b) Why do we think that deaths might be more accurate than cases? What problems might there be with the way case data is generated?

* Deaths might be more accurate than cases as an indicator of the trends in SARS-CoV-2 incidence because death is a more severe outcome and is less likely to be under-reported than cases. Cases are subject to under-reporting because newly infected individuals experience mild symptoms may not test, seek care, or report their test result.

### c) Discuss what you know/could find out that is important to understanding these data. How are the numbers linked? How are they generated?

* Important to know the representativeness of the data in terms of (1) whether they are in-hospital or out-hospital deaths, (2) if in hospital deaths, what are the catchment population of the hospital, (3) who has access to the hospital services (for death) and test services (for cases)?
* Also important to understand to what extent people who tested positive would attend the hospitals and show up in the death data.
* Lastly, important to know reporting process–how people get into the line list (if there is one). Hospitals have to be a reportable disease, reporting infrastructure, with lags.

### d) What other data or assumptions do we need to know to estimate infections from deaths? How much precision might we have in those assumptions?

* We need clinical data informing the infection-to-fatality ratio (IFR) specific to the particular population.
* Regarding the precision, we need to acknowledge that IFR changes over time.

### e) What are some other biases or problems that we haven’t accounted for?

* There is a time delay between infection and death (e.g., 21 days), so that deaths occurring today correspond to infections that have occurred, e.g., 21 days ago, accounting for the infection-to-death period.

## 4.3 Back calculating SARS-CoV-2 infections from deaths

IFR <- 0.01 ## Proportion of people infected that die
time\_shift <- 21 ## Average interval, in days, between infection time and death time

## Shift dates back
brazil\_data <- brazil\_data %>% mutate(Date\_shifted= Date\_reported-time\_shift) ## << Calculate date of infection from date of death>>

## scale up death numbers
brazil\_data <- brazil\_data %>% mutate(Infections\_estimated= New\_deaths\_rollmean/IFR) ## << Calculate cases from deaths using IFR

## Plot against cases
p\_cases\_calculated <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_shifted,y=Infections\_estimated,col="Estimated infections")) +
 geom\_line(aes(x=Date\_reported,y=New\_cases\_rollmean,col="Reported cases")) +
 scale\_color\_manual(values=c("Estimated infections"="blue","Reported cases"="red")) +
 ylab("Estimated infections")+
 scale\_y\_continuous(labels = scales::label\_number\_si())
p\_cases\_calculated

## Warning: Removed 6 rows containing missing values (`geom\_line()`).
## Removed 6 rows containing missing values (`geom\_line()`).



## 4.4 Compare reported vs. estimated infections

sum(brazil\_data$New\_cases\_rollmean, na.rm = T)

## [1] 8320676

sum(brazil\_data$Infections\_estimated, na.rm = T)

## [1] 20695971

* Estimated infection is 2.5 times higher than reported cases given a IFR of 0.01. What does this tell us about underreporting?

p\_report\_rate <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_reported,y=Infections\_estimated/New\_cases\_rollmean),col="purple",size=1) +
 ylab("Underreporting rate")+
 scale\_y\_continuous(labels = scales::label\_number\_si())
p\_report\_rate

## Warning: Removed 54 rows containing missing values (`geom\_line()`).



## 4.5 Sensitivity analysis

## Try different IFRs ##
IFR\_low <- 0.001
IFR\_high <- 0.03
brazil\_data <- brazil\_data %>% mutate(Date\_shifted=Date\_reported-time\_shift,
 Infections\_estimated\_low=New\_deaths\_rollmean/IFR\_low,
 Infections\_estimated\_high=New\_deaths\_rollmean/IFR\_high)

## Plot against cases
p\_cases\_calculated2 <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_shifted,y=Infections\_estimated\_high,col="Estimated infections (high IFR)")) +
 geom\_line(aes(x=Date\_reported,y=New\_cases\_rollmean,col="Reported cases")) +
 scale\_color\_manual(values=c("Estimated infections"="blue","Reported cases"="red",
 "Estimated infections (low IFR)"="purple","Estimated infections (high IFR)"="darkgreen"))+
 scale\_y\_continuous(labels = scales::label\_number\_si())
p\_cases\_calculated3 <- ggplot(data=brazil\_data) +
 geom\_line(aes(x=Date\_shifted,y=Infections\_estimated\_low,col="Estimated infections (low IFR)")) +
 geom\_line(aes(x=Date\_reported,y=New\_cases\_rollmean,col="Reported cases")) +
 scale\_color\_manual(values=c("Estimated infections"="blue","Reported cases"="red",
 "Estimated infections (low IFR)"="purple","Estimated infections (high IFR)"="darkgreen"))+
 scale\_y\_continuous(labels = scales::label\_number\_si())
p\_cases\_calculated2/p\_cases\_calculated3

## Warning: Removed 6 rows containing missing values (`geom\_line()`).
## Removed 6 rows containing missing values (`geom\_line()`).
## Removed 6 rows containing missing values (`geom\_line()`).
## Removed 6 rows containing missing values (`geom\_line()`).



* More severe under-reporting of cases when IFR is low and less severe when IFR is high.

## 4.6 Look at implied total infections

## Population of brazil is ~200,000,000
brazil\_pop <- 200000000
sum(brazil\_data$Infections\_estimated,na.rm=TRUE)/brazil\_pop

## [1] 0.1034799

* Around 10% of the total Brazilian population has been infected by 2021-01-19.

### f) Is 3% a reasonable IFR? Note that cumulative cases reported should always be less than cumulative infections. If cumulative cases exceeds estimated infections, then our assumed 3% IFR is not reasonable.

sum(brazil\_data$Infections\_estimated\_high,na.rm=TRUE)/brazil\_pop

## [1] 0.03449329

sum(brazil\_data$New\_cases,na.rm=TRUE)/brazil\_pop

## [1] 0.0424405

* Note that 4.2% of the whole population had been reported cases, which exceeded the estimated total infections (3.4%), so the assumed IFR = 3% is not reasonable and the true IFR is certainly not that high.

## g) Is 0.01% a reasonable IFR? Note that cumulative infections should always be less than total population for the observation period. If cumulative infections exceeds the total population, then our assumed 0.01% IFR is not reasonable.

sum(brazil\_data$Infections\_estimated\_low,na.rm=TRUE)/brazil\_pop

## [1] 1.034799

* It can’t be that low because that would mean literally everyone in Brazil was infected.