**Preamble**

I was slightly reluctant to add to the deluge of charts about the COVID-19 outbreak, but on the other hand making charts is one of the ways I relax and try to understand what’s going on around me. So first, to get out of the way my only advice at this point:

* wash hands frequently, for 20 seconds at a time, with plenty of soap
* work at home if you can
* limit any face to face social activities to very small groups of people and stay 1.5 metres apart
* if you’re a government, encourage people do the above; while resourcing and preparing the health system properly; and trying to cushion the economy (and particularly the most vulnerable people in it) from the shock.

For a educational tool on why social distancing works, I particularly recommend this beautifully put together [story and simulations from The Washington Post](https://www.washingtonpost.com/graphics/2020/world/corona-simulator/).

**Evolving understanding over time of case fatality rate**

I have a professional interest in uncertainty and in how we work with partial information and the drip feed of new information. So I was interested to create this chart, showing the [case fatality rate](https://en.wikipedia.org/wiki/Case_fatality_rate) (deaths from this disease, divided by all people diagnosed with this disease, in a given period of time) of COVID-19 over time.

What I’m showing here is the cumulative case fatality rate, based on all observations up to a given point in time. Eventually there will be a single number for the case fatality rate of the COVID-19 pandemic of 2020 to 2021, but not yet; the future is uncertain.

An edifying point here for statisticians to note is that if we treated deaths from diagnosed cases as a Bernoulli variable drawn from a population of the ‘true’ case fatality rate and naively estimated sampling error, for most of the chart above it would be negligible. Yet clearly there is a *lot* of uncertainty about where the rate will end up. There are many sources of uncertainty other than sampling error. Bear this in mind when next you consider an opinion poll.

Obviously a key driver of the change over time is the move of the disease into different populations, and particularly demographically older countries. Another key driver is the degree of testing, which provides the denominator of the case fatality rate. More testing means more known cases, driving the rate down. We see the impact of both these drivers when we decompose the case fatality rate by country:

The USA (grey line in the chart above, label somewhat hidden by France’s) provides a stark example of that second driver – testing was very slow to get off the ground, and as the extent of infections in that country is revealed by more testing the apparent case fatality rate is headed downwards. Italy – one of the oldest countries in the world – provides the exemplar of the first driver although this reveals itself in the high position on the latest data (7.5%!) rather than the trend in Italy’s time series. The increase over time in fatality rate there might be a result of the increasing strain on the health system, or simply an early-outbreak phenomenon from a lag in infections leading to deaths.

Here’s the code for those two simple plots. All the downstream data wrangling is done for me by Johns Hopkins and by Rami Krispin (who took the Johns Hopkins data and tidied it into an R package). All the upstream data collection, wrangling and reporting is done by the governments of various countries under great strain, mostly reporting to the WHO.

#--------------- Setup--------------------

library(coronavirus)

library(tidyverse)

library(scales)

the\_caption = "Source: WHO and many others via Johns Hopkins University and Rami Krispin's coronavirus R package.\nAnalysis by <http://freerangestats.info>"

top\_countries <- coronavirus %>%

filter(type == "confirmed") %>%

group\_by(Country.Region) %>%

summarise(cases = sum(cases)) %>%

top\_n(8, wt = cases)

#---------------------------global total-------------------

first\_non\_china\_d <- coronavirus %>%

filter(Country.Region != "China" & type == "death" & cases > 0) %>%

arrange(date) %>%

slice(1) %>%

pull(date)

first\_italy\_d <- coronavirus %>%

filter(Country.Region == "Italy" & type == "death" & cases > 0) %>%

arrange(date) %>%

slice(1) %>%

pull(date)

d1 <- coronavirus %>%

group\_by(date, type) %>%

summarise(cases = sum(cases)) %>%

arrange(date) %>%

spread(type, cases) %>%

ungroup() %>%

mutate(cfr\_today = death / confirmed,

cfr\_cumulative = cumsum(death) / cumsum(confirmed))

d1b <- d1 %>%

filter(date %in% c(first\_italy\_d, first\_non\_china\_d))

ac <- "steelblue"

d1c <- d1 %>%

mutate(cc = cumsum(confirmed)) %>%

summarise(`10000` = min(date[cc > 10000]),

`100000` = min(date[cc > 100000])) %>%

gather(variable, date) %>%

left\_join(d1, by = "date") %>%

mutate(label = paste0(format(as.numeric(variable), big.mark = ",", scientific = FALSE), "\ncases"))

d1 %>%

ggplot(aes(x = date, y = cfr\_cumulative)) +

geom\_line() +

scale\_y\_continuous(label = percent\_format(accuracy = 0.1)) +

expand\_limits(y = 0) +

geom\_point(data = d1b, colour = ac, shape = 1, size = 2) +

annotate("text", x = first\_italy\_d,

y = filter(d1, date == first\_italy\_d)$cfr\_cumulative - 0.001,

label = "First death in Italy",

hjust = 0, size = 3, colour = ac) +

annotate("text", x = first\_non\_china\_d,

y = filter(d1, date == first\_non\_china\_d)$cfr\_cumulative + 0.001,

label = "First death outside China",

hjust = 0, size = 3, colour = ac) +

geom\_text(data = d1c, aes(label = label),

size = 3, colour = "grey70",

hjust = 0.5, lineheight = 0.9, nudge\_y = -0.002) +

labs(caption = the\_caption,

x = "",

y = "Observed case fatality rate",

title = "Steadily increasing case fatality rate of COVID-19 in early 2020",

subtitle = "Increase probably reflects move of the disease into older populations.

Note that actual case fatality is likely to be much lower due to undiagnosed surviving cases.")

#-----------------Country-specific totals------------------------

d2 <- coronavirus %>%

group\_by(date, Country.Region, type) %>%

summarise(cases = sum(cases)) %>%

group\_by(date, Country.Region) %>%

spread(type, cases) %>%

arrange(date) %>%

group\_by(Country.Region) %>%

mutate(cfr\_cumulative = cumsum(death) / cumsum(confirmed)) %>%

filter(![is.na](http://is.na)(cfr\_cumulative)) %>%

ungroup() %>%

inner\_join(top\_countries, by = "Country.Region")

d2 %>%

ggplot(aes(x = date, y = cfr\_cumulative, colour = Country.Region)) +

geom\_line() +

geom\_text(data = filter(d2, date == max(date)), aes(label = Country.Region),

hjust = 0, check\_overlap = FALSE, size = 3) +

scale\_y\_continuous(label = percent\_format(accuracy = 1), limits = c(0, 0.2)) +

scale\_colour\_brewer(palette = "Set2") +

expand\_limits(x = max(d2$date) + 4) +

labs(caption = the\_caption,

x = "",

y = "Observed case fatality rate",

title = "Country-specific case fatality rate of COVID-19 in early 2020",

subtitle = "Eight countries with most diagnosed cases; Iran's early values truncated.\nA high level of uncertainty reflecting rapidly changing denominators as well as many unresolved cases.") +

theme(legend.position = "none")