Johannes Bracher (johannes.bracher@kit.edu) 15 April 2024
NOZGEKA infectious disease modelling short course

Exercise sheet *Nowcasting, short-term forecasting and R-values*

Background: We will analyse a subset of the data set on the 2003 SARS epidemic in Hong Kong contained in the EpiEstim R package (type `?SARS2003` in `R`); see also Cori et al (2009, `https://doi.org/10.1371/journal.pcbi.1000471`). Some parts of this exercise sheet, like the reporting triangle, are not actually available and have been generated artificially by me for exercise purposes.

Access to files: Please download the files made available at `https://github.com/jbracher/NOZGEKA`. You can do that by clicking on **Code** → **Download ZIP**.

Use of software: I tried to conceive this exercise sheet in a way that it can be solved using `R` or purely the EpiEstim web app and a calculator. You may choose your preferred tool. Solutions refer to `R`, but the results should be the same using the web app.

Solutions: Solutions are available in the file `exercise1_solution.pdf`. Do not hesitate to consult the solution if you are stuck somewhere.

1. Read the file `data_day_50_realtime.csv` into `R` or the EpiEstim web app at `https://shiny.dide.imperial.ac.uk/epiestim/`. Note: the latter may throw a security warning in your browser.
   If you are using the web app: You will need to tick the box "Header" when using the provided file. Also, you will only get a visualization once you are also estimating R_t. To this end you need to select "Distributional Estimate" and "Parametric without uncertainty (offset gamma)" in order to be able to specify the generation time distribution via its mean and standard deviation.

2. Generate a visualization (epicurve) of the data.

3. Load the package `EpiEstim` and estimate the effective reproductive number R_t. Use a parametric (gamma) serial interval distribution with mean 8.4 days and standard deviation 3.8 days. For the window size use 4 days. Plot the result. (Note: as R_t fluctuates quite a bit, plotting it on a log scale may be helpful).
   Hint: the start and end dates for 4-day windows can be specified as follows.

   ```
   # generate start and end days for 4-day windows
   t_start <- seq(2, 47) # as there are 50 days in the data
   t_end <- t_start + 3 # note: end days are included, thus +3 rather than +4
   ```

4. Check the results visually and think about the two following questions:

   (a) Around day 24, the reproductive number is estimated to be above 10. According to Wikipedia (`https://en.wikipedia.org/wiki/SARS`), "R_0, ranges from 2 to 4 depending on different analyses" for SARS-COV-1. Discuss some possible explanations for these extreme estimates.

   (b) For the most recent days (47–50), the estimated R_t falls considerably below 1. Which potential issue do you see with this pattern?

5. As the epidemiologists in charge are aware of the delay problem commonly occurring in surveillance data, they preserved daily snapshots of the epicurve and regularly generate reporting triangles. Here is the reporting triangle they generated on day 49:

t	$d=0$	$d=1$	$d=2$	$d=3$
41	22	23	11	2
42	16	19	8	5
43	9	16	4	4
44	5	12	5	3
45	14	18	7	4
46	16	12	4	5
47	12	10	6	?
48	6	10	?	
49	5	?		
50	?			

On day 50 the cells containing question marks can be filled in. Use the time series data as available on days 49 and 50 to compute the missing cells:

t	41	42	43	44	45	46	47	48	49	50
data as of day 49	58	48	33	25	43	37	28	16	5	
data as of day 50	58	48	33	25	43	37	30	26	14	7

6. The empty cells in your updated reporting triangle can be filled in using the simple nowcasting method from the slides. Perform these computations using a calculator or R. **Note:** The reporting triangle for day 50 is available in the file `reporting_triangle_day50.csv`, but to keep things simple you can just note your results by hand. The incidence time series with nowcasted and complete data are available in `data_day50_nowcast.csv` and `data_day50_complete.csv` for further use. So you are not obliged to perform the computations for all six cells if you feel you got the principle.

7. Use the incidence time series with nowcasting correction(file `data_day50_nowcast.csv`) and the final time series (file `data_day50_complete.csv`) to re-run the nowcasting computations. Compare the results to those from question 4.

8. Re-run the R_t estimation using the complete data and an adapted serial interval distribution with mean 5 days and standard deviation 3 days. How do the results compare to those from the previous question?

9. Re-run the R_t estimation using a window size of 7 rather than 4 days. How do the results compare to those from question 8?

10. The incidence time series is aggregated by the date of symptom onset. It is assumed that the mean incubation period of SARS-COV1 is around 5 days. How can you take this into account when interpreting the results? Use the value 0.62 (uncertainty interval 0.5–0.71) estimated for day 50 as an example.