IC 252 Lab 6

Gaurav Bhutani

The present assignment requires you to analyse Coronavirus disease 2019 (COVID-19) data from various data sources including data from India and around the world. You will see yourself unearthing very interesting information about COVID-19 "hidden" in the data.

Problem 1. The data file *Covid19IndiaData_30032020.xlsx* presents the Indian patient-level data until 30th March 2020. Source for latest Indian COVID-19 data: https://api.rootnet.in/.

(i) Calculate and plot the probability mass function (pmf) of the age of infected patients — this includes *Hospitalized*, *Recovered* and *Dead*. Evaluate the expected age of an infected patient from this pmf and the variance of the pmf. Does the variance seem *high* or *low*? What can you say about the expectation calculated in light of the variance of the pmf?

(ii) Calculate and plot the pmfs of the age of *Recovered* and *Dead* patients. Calculate the expectation and variance of the pmfs. Are the expectations the same as in Case (i)? What can you say about COVID-19 by comparing the expectation values?

(iii) Find the conditional pmf of the age of all infected patients conditional to the *gender* of the patient. Are the pmfs identically distributed? Compare the expectations and comment of the possible reason(s) for any difference.

Problem 2. The data file *linton_supp_tableS1_S2_8Feb2020.xlsx* presents patient-level case data from China and other parts of the world. This includes the following information — *Exposure date (E)*, *Symptoms onset date (O)*, *Hospitalisation date (H)*, and *Date of death (X)* in case of deceased patients. The data also includes whether the patient is/was a resident of Wuhan (China). Please note that details of surviving and deceased patients (until 31st January 2020) are included as different Sheets. The data can be downloaded from http://www.mdpi.com/2077-0383/9/2/538/s1 and may require some *cleaning* before being used. You can compare your results to the results published in the Journal of Clinical Magazine in February 2020 (https://www.mdpi.com/2077-0383/9/2/538/pdf).

(i) Calculate and plot the pmf of the incubation period, which is defined as the duration between the date of infection exposure (E) and the date of onset of symptoms (O). Calculate the mean incubation period and the variance of the distribution. Please use the left exposure data as the date of infection.

(ii) Now calculate the expected incubation period by excluding Wuhan residents and compare the values with part (i). What can you comment based on the comparison?

(iii) Calculate the pmfs of the onset to hospitalization ($H - O$) for dead patients, onset to death ($X - O$) and hospitalization to death ($X - H$). Do you see a similarity in the distribution? Comment. Also, compare the $H - O$ pmf for surviving and dead patients; comment on the difference.

Problem 3 (optional and ungraded). The data file *time-series-19-covid-combined_csv.csv* presents daily timeseries data of cumulative region-wise cases—*Confirmed*, *Recovered* and *Deaths*. The latest data can be downloaded from https://datahub.io/core/covid-19.

(i) Calculate and plot the pmf of the increase rate of Confirmed cases (totalled over the world), which is calculated as the difference in the Confirmed cases between two consecutive days. Please note that you will have to reorganise the data with one row for each date, for the present analysis. Calculate the expectation of the increase rate using the pmf along with its variance.

(ii) Plot the pmfs of the increase rate of Recovered and Dead patients, along with their expectations. Do you see any difference in the expectation of the increase rates of Confirmed, Recovered and Dead patients?

(iii) Find the expectation (and variance) of conditional pmfs of the increase rate of Confirmed cases region-wise and month-wise.

(iv) Evaluate the joint pmf of the increase rate of Confirmed cases (C) and the increase rate of Recovered cases (R). Are C and R independent?

(v) Would expectation be the best indicator for the present timeseries data or do you propose a more useful way to predict the increase rate of the next day? *Hint: increase rate for any given day may be a function of the past dew days; see* time series *analysis.*

(vi) Can you think of any other random variables that may be useful in analysing this data?

More COVID-19 related data can be found at the following sources. Please feel free to analyse the data and mine useful information out of it. You could then also run your codes on a daily basis on the latest data, updating COVID-19 predictions. These latest findings (in the forms of graphs, plots and summaries) can be presented and shared on plus.google.com, which is an internal social media only accessible through IIT Mandi email accounts. Use the hashtag #ic252covid19.

- Real-time case information: https://www.nature.com/articles/s41597-020-0448-0.pdf
- Latest Indian COVID-19 data: https://api.rootnet.in/
- Latest worldwide time series listing of COVID-19 cases: https://datahub.io/core/covid-19