50.034 – Introduction to Probability and Statistics

January–May Term, 2021

Problem Set

Due by: Week 14 Monday (26 Apr 2021), 1pm.
Please submit your problem set online via eDimension.

Reminder: There will be a physically held final exam in Week 14.

In the second half of this course, you have learned/will learn various concepts in statistics, including conjugate priors, Bayes estimation, MLE, and linear regression.

For this problem set, you will work on a **real-world problem** that is currently of immense concern to everyone. Your goal is to apply all these concepts to estimate the basic reproduction number for COVID-19, based on actual data from Singapore's Ministry of Health (MOH), and publicly available data shared by European Centre for Disease Prevention and Control (ECDC).

The *basic reproduction number* of an infectious disease, typically denoted by R_0, is defined to be the expected number of people that an infected individual will go on to infect. Notice that R_0 is not a constant! Different countries have different R_0 values for COVID-19, depending on environment factors, social behaviour, etc., and the R_0 value can be reduced over time, with measures implemented by the country's government (e.g. social distancing).

Overview. We shall estimate the worldwide value of R_0 for COVID-19 in the early stages of the outbreak, by first estimating two parameters: the generation interval for COVID-19, and the growth rate for the number of COVID-19 cases.

- The *generation interval* of an infectious disease shall mean for us a random variable G representing the duration (in number of days) from the day an infected individual gets infected, to the day the infected individual infects another individual. For this problem set, we shall assume that the generation interval G of any infectious disease always follows a gamma distribution.[1]
- In the early stages of an epidemic, the number of infected cases would have an exponential growth over time. The *growth rate* for the number of infected cases shall mean the daily rate of growth r for this exponential growth. In other words, the number of infected cases on Day k equals approximately $n_0 e^{rk}$, where n_0 represents the number of infected cases on Day 0.[2]

Given that the gamma random variable G has parameters α and β, and assuming that every individual is susceptible, we can approximate[3] the R_0 of an infectious disease in the early exponential growth stage, via the following approximation:

$$R_0 \approx \left(1 + \frac{r}{\beta}\right)^{\alpha}.$$

In this messy real world, there is unfortunately only a limited number of infected cases with sufficient information that would help us estimate the parameters α and β for the generation interval of COVID-19 with high certainty. Hence, we shall use a Bayesian approach: Based on what we understand about SARS, we shall fix a constant value for α, and choose a suitable prior distribution for β. We then compute the posterior distribution for β based on our limited number of relevant COVID-19 observations. As for the growth rate r, we shall use linear regression to approximate r.

[1] The use of gamma distributions to model generation intervals is a common practice in epidemiology.
[2] In this problem set, we shall estimate the worldwide growth rate, and not the growth rate in Singapore.
[3] This approximation is based on the Euler-Lotka equation, which will not be covered in this course.

Step 1: Choosing a prior based on data from the SARS outbreak.

Let G_{SARS} and G_{COVID} be the generation intervals for SARS and COVID-19 respectively. Assume that G_{SARS} follows a gamma distribution with parameters α_{SARS} and β_{SARS}. Similarly, assume that G_{COVID} follows a gamma distribution with parameters α_{COVID} and β_{COVID}.

Question 1a:
Research studies on the SARS outbreak have reported that the observed values for the sample mean and sample standard deviation of G_{SARS} are 8.4 and 3.8 respectively, based on a large sample of SARS infected cases. Using the method of moments, give estimates for α_{SARS} and β_{SARS}. Explain in as much details as possible how you obtain your estimates. (5 marks) [Hint: Have you read Chapter 7.6 of the course textbook?]

Question 1b*: (Bonus question)
The values 8.4 and 3.8 given in Question 1a are obtained from a scientific paper on SARS, and these values are rounded off to the nearest 1 decimal place. The full data of all observed values of individual instances of G_{SARS} can be found in the appendix at the end of this problem set. Suppose $\widehat{\alpha}_{\text{SARS}}$ and $\widehat{\beta}_{\text{SARS}}$ are the maximum likelihood estimates of α_{SARS} and β_{SARS} respectively. Using the full data in the appendix, determine the values of $\widehat{\alpha}_{\text{SARS}}$ and $\widehat{\beta}_{\text{SARS}}$. Explain in as much details as possible, including full justifications of why your estimates are maximum likelihood estimates. (5 bonus marks)
[Hint: Have you read Chapter 7.6 of the course textbook?]

We have learned concepts such as the mean μ and the standard deviation σ of a random variable. In Homework 6, we saw another useful concept called the *coefficient of variation*[4], which is defined as $\frac{\sigma}{|\mu|}$. (This is defined only when $\mu \neq 0$.) In epidemiology, the coefficient of variation is used as a measure to categorize diseases and quantify health risk. Notice that both SARS and COVID-19 are diseases caused by coronaviruses, with "similar" health risk. For simplicity, we shall assume that both G_{SARS} and G_{COVID} have the same coefficient of variation. Our rationale for considering SARS is the hope that the generation intervals for both diseases are sufficiently similar to give us a "good" prior.

Question 1c:
Explain in detail why this assumption (that G_{SARS} and G_{COVID} have the same coefficient of variation) implies that $\alpha_{\text{SARS}} = \alpha_{\text{COVID}}$. (5 marks)

Assume that α_{COVID} is a fixed constant, whose value equals the value of α_{SARS} obtained in Question 1a. We shall treat β_{COVID} as a random variable. Assume that the prior of β_{COVID} follows a gamma distribution with prior hyperparameters 2 and λ.

Question 1d:
Within the same diagram, plot four graphs of the probability density functions for the gamma distribution with parameters 2 and θ, for the four values $\theta = 1, \theta = 2, \theta = 3, \theta = 4$. (The four graphs have to be in the same plot.) You should clearly label the four graphs so that there is no ambiguity as to which value for θ each graph corresponds to. How does the shape of the graphs change as the value of θ increases? (5 marks)

Question 1e:
The random variable β_{COVID} has a gamma prior distribution with prior hyperparameters 2 and λ. What should the value of λ be, so that the conditional expectation of $\frac{\alpha_{\text{COVID}}}{\beta_{\text{COVID}}}$ given this gamma prior equals 8.4? (5 marks)
[Note that 8.4 is the observed value for the sample mean of G_{SARS}.]

Henceforth, assume that the gamma prior hyperparameters for β_{COVID} are 2 and λ, where λ has the value you obtained in Question 1e.

[4]The coefficient of variation is sometimes also called the *relative standard deviation*. This concept is widely used in chemistry, economics, engineering, and physics.

Step 2: Computing a posterior based on data from Singapore's MOH.
In many real-world problems, data collection and data preparation are essential prerequisite steps before we can carry out any statistical inference.

For this step, we shall simulate the situation that we only have access to the press releases by Singapore's MOH for a 6-month time period, from the beginning of September 2020, up to and including 28th February 2021. The press releases can be found via the following link: https://www.moh.gov.sg/covid-19/past-updates

Question 2a:
Your goal is to find infected cases for which MOH has provided sufficient information, such that the observed values for G_{COVID} corresponding to these selected infected cases can be inferred. Search through these press releases in the Sept 2020–Feb 2021 period, and give a list of at least 10 infected cases, with their corresponding observed values for the generation interval G_{COVID}. Notice that every infected case has a uniquely assigned case number, so please indicate the case numbers clearly for your selected infected cases, and explain how you inferred their corresponding observed values for G_{COVID}. (20 marks)

In this course, we have seen the concept of conjugate priors. In Week 9's cohort class, we learned that if we have a statistical model consisting of i.i.d. observable random variables sampled from either a Poisson distribution with parameter θ or a exponential distribution with parameter θ, then the family of gamma distributions is a family of conjugate priors for this parameter θ. Hence, given a specific gamma prior for θ and given the observed values for the observable random variables, we can easily determine the gamma posterior for θ.

An analogous result holds if our i.i.d. observable random variables are sampled from a gamma distribution with parameters α_0 and θ, where α_0 (first parameter for this gamma distribution) is a known constant, and θ (second parameter for this gamma distribution) is a random variable.

Theorem. Consider a statistical model where $X_1, \ldots, X_n$ are observable random variables that each follows a gamma distribution with parameters α_0 and θ. Assume that α_0 is a fixed known constant, and treat θ as a random variable. Assume that $X_1, \ldots, X_n$ are conditionally i.i.d. given θ. If θ has a gamma prior with prior hyperparameters α and β, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the gamma distribution with posterior hyperparameters $\alpha' = \alpha + n\alpha_0$ and $\beta' = \beta + (x_1 + \cdots + x_n)$.

Question 2b:
Using the Bayesian approach, define a statistical model for the inference of the generation interval of COVID-19. In your model set-up, please clearly define your random variables. You should clearly specify whether the random variables are observable or latent, and you should also clearly specify all model assumptions. (5 marks)
[Hint: What is the definition of a statistical model?]

Question 2c:
Based on the observed values you found in Question 2a, and using the prior distribution for β_{COVID} from the end of Step 1, determine the posterior distribution for β_{COVID}. Justify your answer in as much details as possible. [Hint: Use the theorem above.] (5 marks)

Question 2d:
Using Bayes estimation, give an estimate for β_{COVID}. Please give all details, including any assumptions you make. (5 marks)

Question 2e:
The posterior distribution you computed in Question 2c is based on a specific choice of a prior gamma distribution with prior hyperparameters 2 and λ, where λ has the value you obtained in Question 1e. Perform sensitivity analysis by testing different values of λ (with the fixed value 2 for the first parameter). What is the range of values for λ you tested? Explain your choice for this range, and clearly state any assumptions you make. (20 marks)

Step 3: Computing the growth rate based on data from ECDC.
As of 1st April 2021, the United States has the most number of reported COVID-19 cases (30.46 millon!) among all countries. While there have been several waves of infection so far, the initial stages of the outbreak in March 2020 (last year) essentially had an exponential growth in the number of cases. We shall use publicly available data on the daily increases in the number of infected cases in the United States to estimate their growth rate, which serves as a proxy for approximating the worldwide growth rate during the early stages of the pandemic last year. The data provided by ECDC can be found here: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide (Although this site is no longer up-to-date, it still has the data in the early stages of the pandemic.) For each integer $1 \leq k \leq 31$, let z_k be the **cumulative** number of COVID-19 cases in the United States from the start of the pandemic up till the k-th day of March 2020, and let $y_k = \ln z_k$.

Question 3a:
We have assumed that the number of COVID-19 cases has an exponential growth over time. Explain in detail why this assumption implies that y_k grows linearly with respect to k. (5 marks)

Question 3b:
For each integer $1 \leq k \leq 31$, let $x_k = k$, and treat $(x_1, y_1), \ldots, (x_{31}, y_{31})$ as 31 points on the xy-plane. Find the best fit line $y = mx + c$ for these 31 points. What is a good approximation for the growth rate r for COVID-19? (10 marks)

Step 4: Estimating the R_0 for COVID-19.
From Steps 1–3, we have obtained estimates for α_{COVID}, β_{COVID}, and the growth rate r for the number of COVID-19 cases.

Question 4a:
What is your estimate for the basic reproduction number R_0 for COVID-19? (2 marks)

Question 4b:
Can you describe two limitations for the model assumptions used in Steps 1–3? (8 marks)

Final Remarks: You have all the tools and knowledge to continue monitoring the COVID-19 pandemic, including estimating the basic reproduction number R_0 for COVID-19 in any specific country. How important do you think social distancing is in helping to reduce R_0? This last question on social distancing is not part of the problem set, but it is something to think about.

Also, in Step 2, the process of searching through press releases is exactly what many COVID-19 researchers actually do to prepare their data for statistical inference! For statistical inference on G_{COVID}, it is interesting to note that many researchers indeed use Singapore's detailed press releases for this data preparation step.

Appendix: The following data is obtained from a 2003 paper published in Science.[5]

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$Num(k)$	3	1	14	15	10	11	19	21	22	18	9	12	8	6	3	2	3	1	1	1

Here, $Num(k)$ denotes the number of cases with observed values $G_{\text{SARS}} = k$.
There were no reported cases of any observed value $G_{\text{SARS}} > 20$.

[5]Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003.