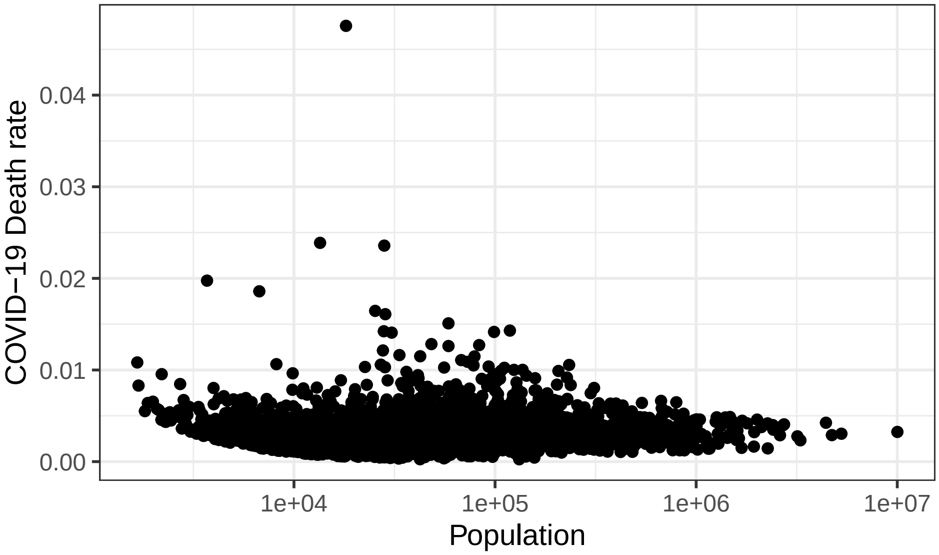
Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 8 due Friday 4/7

COVID-19 Impact

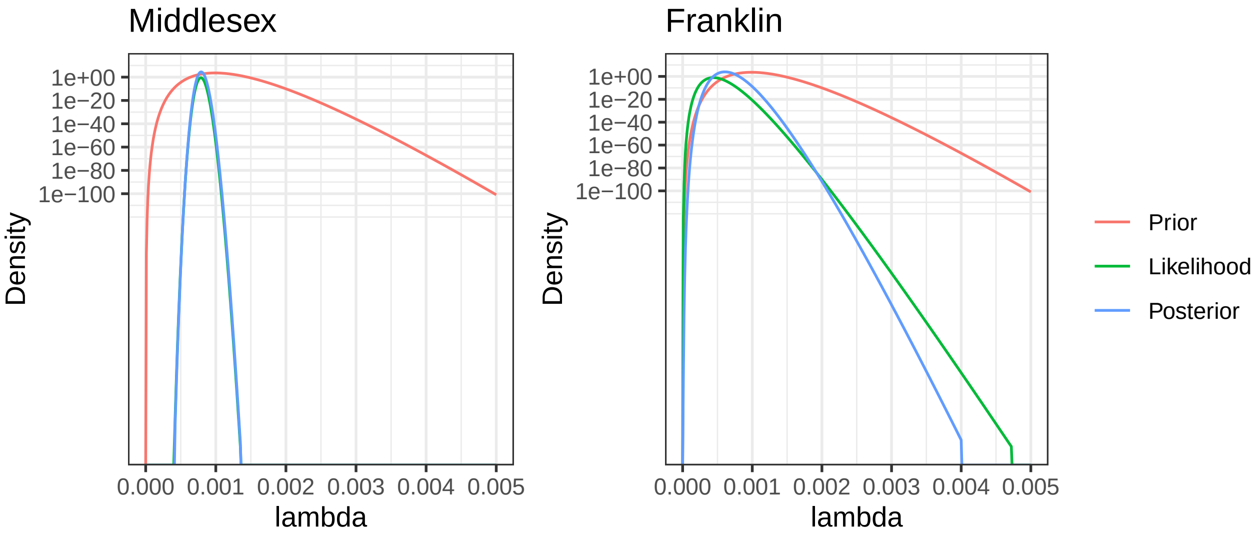
These questions will deal with a dataset listing deaths from COVID-19 per county from January 1st, 2020 to March 25th, 2023 available here and 2020 county population numbers available here. The CDC does not make data available for counties with fewer than 10 deaths. We could use a censored data approach, but for simplicity we will restrict our focus to counties with at least 10 deaths.



1. Suppose we wanted to know which counties had the best and worst COVID-19 responses. Name a few reasons we should not just look at the counties with the maximum and minimum raw death rates.

2. We will model the deaths in a particular county as $Y_i \sim \text{Pois}(c\lambda_i n_i)$ where $c = 3.23$ is the number of years included in the data set, λ_i is the county's annual death rate from COVID-19, and n_i is the county's population. Also, suppose we use the prior $\lambda_i \sim \text{Gamma}(a, b)$. Write the prior density for λ_i, the likelihood function for λ_i, and the posterior density for λ_i. What is the posterior distribution of λ_i?

3. The following plot shows the prior, the likelihood, and the posterior for $b = 100000$, $a = b \cdot 0.001$ for Middlesex, MA (the county that contains Harvard) and Franklin county, a county in western Massachusetts. Middlesex has a population of $n = 1632002$ while Franklin has a population of $n = 71035$. Interpret the plots.



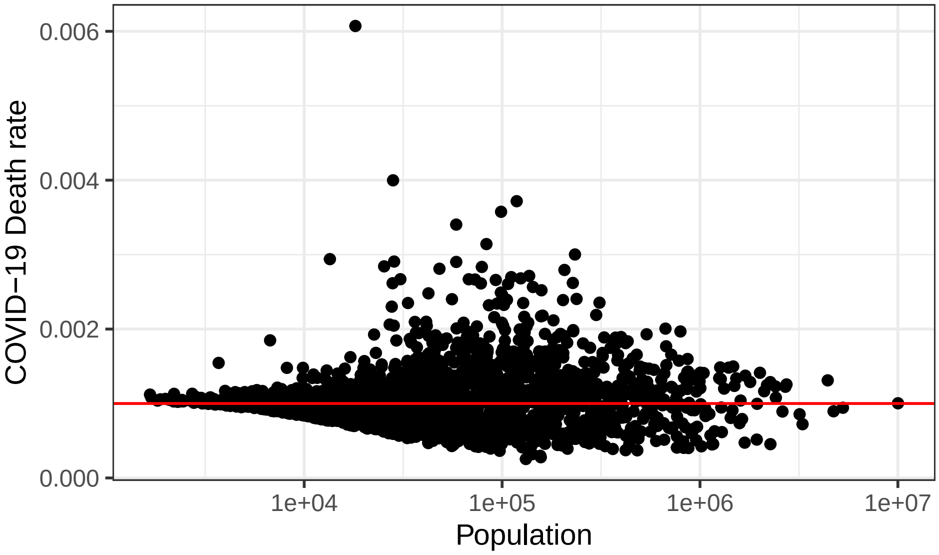
4. Show that the posterior mean $E(\lambda_i|Y_i)$ can be interpreted as a weighted average of the observed death rate and the prior mean. If we view a and b as "pseudocounts" of the number of deaths and the population, give an interpretation of the posterior mean for large b and for large n_i.

5. Suppose (wrongly) that the COVID-19 death rate does not change from year to year. What is the posterior predictive distribution of COVID-19 deaths for 2024 (Y') for a county with Y deaths from 2020 to 2023 and n people? Verify that the expected value and variance of this distribution agree with what we would obtain through Adam's and Eve's laws conditioning on λ.

6. Compare the MLE for λ_i to the posterior mean of λ_i for the counties with the highest COVID-19 death rates. The red line shows the prior mean.

County	Population	Rate (Unadjusted)
Montour County, Pennsylvania	18145	0.0147
Martinsville city, Virginia	13486	0.0074
Winchester city, Virginia	28122	0.0073
Norton city, Virginia	3696	0.0061
Galax city, Virginia	6725	0.0058

County	Population	Rate (Adjusted)
Montour County, Pennsylvania	18145	0.0061
Winchester city, Virginia	28122	0.004
Potter County, Texas	118527	0.0037
Madison County, Tennessee	98843	0.0036
Newton County, Missouri	58644	0.0034



Chat GPT-4 testing

1. You are testing Chat GPT-4's question answering abilities, and you want to evaluate the probability p of it answering a question correctly. To model your initial uncertainty about its abilities, you use the noninformative prior $p \sim \text{Unif}(0,1)$. Assume we have not yet performed any tests. How many questions would Chat GPT-4 need to get correct in a row before we will be c confident p is at least τ? Recall that the PDF of a Beta(a, b) random variable is $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$. Also, note that for an integer a, $\Gamma(a) = (a-1)!$

2. Now, suppose Chat GPT-4 has answered the first m questions correctly. Find the posterior mean, median, and mode (MAP) of p. Show that the MAP is equivalent to the MLE because we are using a flat prior.

3. What is the probability Chat GPT-4 gets the next question correct given it got the first m correct?

4. You have n more questions you plan to ask. Explain intuitively why the probability of it getting all of these n questions correct is not $\left(\frac{1+m}{2+m}\right)^n$.

5. What is the probability of Chat GPT-4 getting the next n questions correct given that it got the first m correct?

6. Why does this make sense in the special case of $m = 0$?

7. Now, suppose Chat GPT-4 has gotten a questions correct and b questions wrong. Updating from the original uniform prior, find the PMF of Y, the number of questions Chat GPT-4 will get correct out of the next n questions.