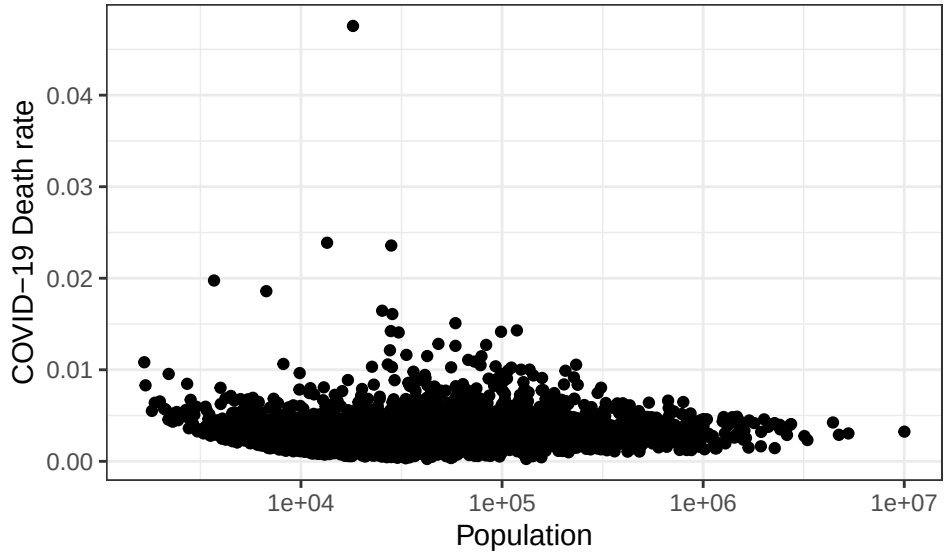
Announcements

Make sure to sign in on the google form.

Pset 8...

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.

First, the death rates are not demographically adjusted, so counties with mostly older people will likely have higher death rates regardless of how effective the response was. Second, counties with fewer people will have more variable death rates (as in the kidney cancer example), but their responses might not have been particularly good or bad.

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?

The prior density is

$$\frac{1}{\Gamma(a)}(b\lambda_i)^a e^{-b\lambda_i}\lambda_i^{-1}$$

The likelihood is

$$L(\lambda_i|Y_i) = e^{-cn_i\lambda_i}\lambda_i^{Y_i}$$

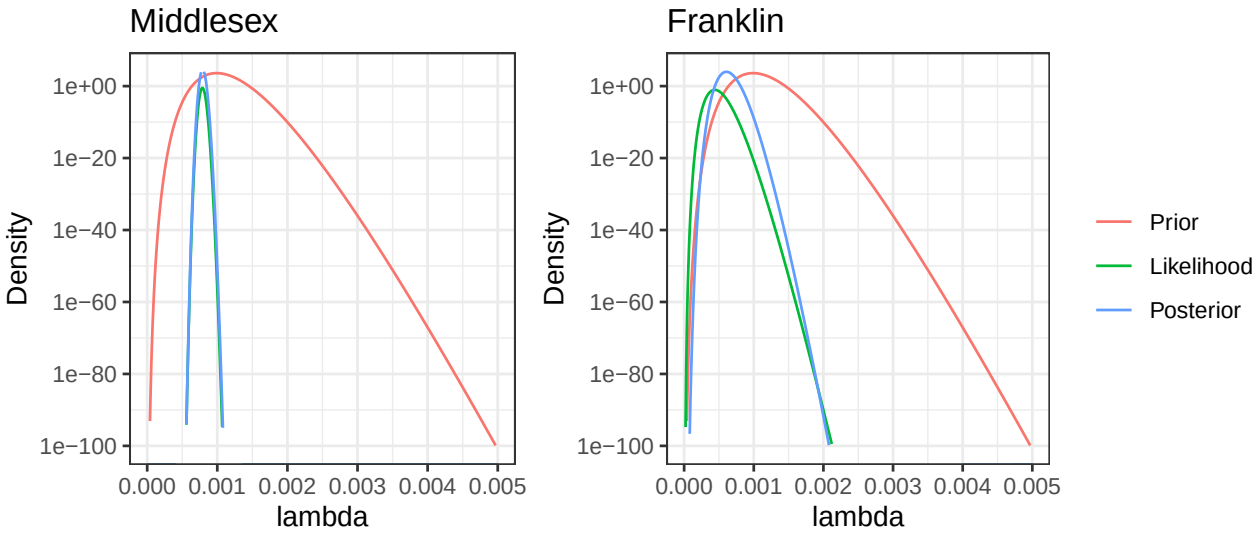
The posterior density is

$$\begin{aligned} f(\lambda|Y_i) &\propto P(Y_i = y_i|\lambda_i)f(\lambda) \\ &= e^{-cn_i\lambda_i}\lambda_i^{Y_i}\lambda_i^a e^{-b\lambda_i}\lambda_i^{-1} \\ &= e^{-(cn_i+b)\lambda_i}\lambda_i^{Y_i+a}\lambda_i^{-1} \end{aligned}$$

Pattern matching shows that this is proportional to the Gamma$(Y_i + a, b + cn_i)$ PDF, so

$$\lambda_i | Y_i \sim \text{Gamma}(Y_i + a, b + cn_i)$$

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.



Both plots show that the posterior is between the prior and the likelihood. However, since Middlesex is much larger, its posterior is much closer to its likelihood than its prior.

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.

From the expectation of a Gamma distribution,

$$E(\lambda_i | Y_i) = \frac{Y_i + a}{b + cn_i} = \frac{Y_i}{cn_i} \frac{cn_i}{b + cn_i} + \frac{a}{b} \frac{b}{b + cn_i}$$

so the posterior mean is a weighted average of $Y_i/(cn_i)$ with weight $\frac{cn_i}{b+cn_i}$ and a/b with weight $\frac{b}{b+cn_i}$. When b is large, our pseudocount prior population is large, so the prior mean accounts for most of the posterior mean. When n_i is large, our observed data is large, so the observed mean accounts for most of the posterior mean.

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.

By the Poisson-Gamma-Negative-Binomial conjugacy, with $\lambda | Y \sim \text{Gamma}(a + Y, b + cn)$ and $Y' \sim \text{Pois}(n\lambda)$, $Y' \sim \text{NBin}\left(a + Y, \frac{b+cn}{b+cn+n}\right)$.

$$E(Y') = E(E(Y'|\lambda)) = E(n\lambda) = n\frac{a + Y}{b + cn}$$

and the mean of the negative binomial distribution is

$$(a + Y)\frac{n}{b + cn + n}\frac{b + cn + n}{b + cn} = n\frac{a + Y}{b + cn}$$

$$\begin{aligned}\text{Var}(Y') &= E(\text{Var}(Y'|\lambda)) + \text{Var}(E(Y'|\lambda)) \\ &= E(n\lambda) + \text{Var}(n\lambda) \\ &= n\frac{a+Y}{b+cn} + n^2\frac{a+Y}{(b+cn)^2}\end{aligned}$$

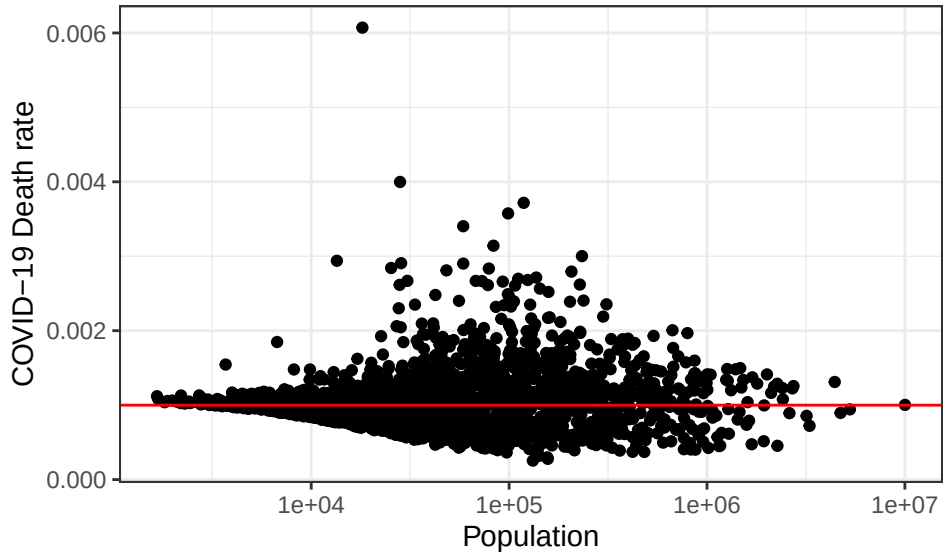
and the variance of the negative binomial distribution is

$$(a+Y)\frac{n}{b+cn+n}\left(\frac{b+cn+n}{b+cn}\right)^2 = n\frac{(a+Y)(b+cn+n)}{(b+cn)^2} = n\frac{a+Y}{b+cn} + n^2\frac{a+Y}{(b+cn)^2}$$

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



Though some of the counties with the highest death rates before still have the highest death rates after, all the estimated rates are much lower. Interestingly, Martinsville and Winchester had about the same rate before, but Winchester has a significantly higher rate after because its population size is larger. We also see that the two very small counties at the end of the raw list are dropped and replaced with much larger counties in the adjusted list.

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 $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$.

Let y be the number of questions Chat GPT-4 gets right in a row. By the Beta-Binomial conjugacy, if the system gets y questions correct in a row, we will update p to $p \sim \text{Beta}(y+1,1)$. We need our maximum c credible interval to have its lower bound at least at τ:

$$Q_{\text{Beta}(y+1,1)}(1-c) \geq \tau \implies F_{\text{Beta}(y+1,1)}(\tau) \leq 1-c$$

Using the fact that $\Gamma(y+2)/\Gamma(y+1) = y+1$,

$$\int_0^\tau (y+1)p^y dp = p^{y+1}\Big|_0^\tau = \tau^{y+1} \leq 1-c \implies y \geq \frac{\log(1-c)}{\log(\tau)} - 1$$

so our answer is the smallest integer larger than $\frac{\log(1-c)}{\log(\tau)} - 1$. (The inequality flip in the equation above comes from the fact that we're dividing by $\log(\tau)$ which is negative.)

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.

The posterior mean is the mean of the Beta$(1+m,1)$ distribution: $\frac{m+1}{m+2}$. The posterior median is τ such that

$$\int_0^\tau (m+1)p^m dp = p^{m+1}\Big|_0^\tau = \tau^{m+1} = 0.5 \implies \tau = \left(\frac{1}{2}\right)^{\frac{1}{m+1}}$$

The posterior mode is the maximum of the density $f(m) = (m+1)p^m$ or equivalently the maximum of the log density: $\log(m+1) + m\log(p)$. Differentiating with respect to p and setting to 0 gives $\frac{m}{p} = 0 \implies p = \infty$. However, p is constrained to $[0,1]$, so we must check the density at the bounds. At $p=0$, $f(m)=0$, and at $p=1$, $f(m) = m+1$, so the maximum is at $p=1$. The MLE of a binomial is $\hat{p} = Y/n$, which is 1 in this case, so the MAP is the same as the MLE.

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

Let Y be the indicator of the system answering the next question correctly. We have $Y \sim \text{Bern}(p)$ with $p \sim \text{Beta}(1+m,1)$. By the fundamental bridge and Adam's law,

$$P(Y=1) = E(Y) = E(E(Y|p)) = E(p) = \frac{1+m}{2+m}$$

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 $\left(\frac{1+m}{2+m}\right)^n$.

If all the questions were independent and we didn't update p between questions, this would be the case. However, when the system gets a question right, it is more likely that p is high, so the probability of getting the next question correct increases.

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

Let Y_i be an indicator of getting the i^{th} question correct of the n remaining. Throughout, we will implicitly condition on the first m being correct. We are solving for:

$$P(Y_1 = 1, ..., Y_n = 1) = P(Y_n = 1 | Y_{n-1} = 1, ..., Y_1 = 1) \cdot ... \cdot P(Y_1 = 1)$$

By the same reasoning as in 3, after getting the first $i-1$ questions correct, $p \sim \text{Beta}(m+i, 1)$, so

$$P(Y_i = 1 | Y_{i-1} = 1, ..., Y_1 = 1) = \frac{m+i}{1+m+i}$$

Therefore,

$$P(Y_1 = 1, ..., Y_n = 1) = \prod_{i=1}^{n} \frac{m+i}{1+m+i} = \frac{m+1}{n+m+1}$$

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

If $m = 0$, we have no observations, so we have a Bayes' Billiards situation where any number of correct responses from 0 to n is equally likely, each with probability $\frac{1}{n+1}$.

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.

Our posterior for p is $\text{Beta}(a+1, b+1)$. By the law of total probability, the fact that the Beta PDF integrates to 1, properties of the Γ function, and the fact that a and b are non-negative integers,

$$\begin{aligned}
P(Y = y) &= \int_0^1 P(Y = y|p) f(p) dp \\
&= \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)} p^a (1-p)^b dp \\
&= \binom{n}{y} \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)} \int_0^1 p^{y+a} (1-p)^{n-y+b} dp \\
&= \binom{n}{y} \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)} \frac{\Gamma(y+a+1)\Gamma(n-y+b+1)}{\Gamma(n+a+b+2)} \int_0^1 \frac{\Gamma(n+a+b+2)}{\Gamma(y+a+1)\Gamma(n-y+b+1)} p^{y+a} (1-p)^{n-y+b} dp \\
&= \binom{n}{y} \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)} \frac{\Gamma(y+a+1)\Gamma(n-y+b+1)}{\Gamma(n+a+b+2)} \\
&= \frac{n!}{y!(n-y)!} \frac{(a+b+1)!}{a!b!} \frac{(y+a)!(n-y+b)!}{(n+a+b+1)!} \\
&= \frac{\binom{y+a}{y}\binom{n-y+b}{n-y}}{\binom{n+a+b+1}{n}}
\end{aligned}$$

We can check this is a valid PMF by ensuring

$$\sum_{y=0}^{n} \frac{\binom{y+a}{y}\binom{n-y+b}{n-y}}{\binom{n+a+b+1}{n}} = 1 \iff \sum_{y=0}^{n} \binom{y+a}{y}\binom{n-y+b}{n-y} = \binom{n+a+b+1}{n}$$

We will prove this using the story proof offered by Peter Luo on Ed: First, note that

$$\sum_{y=0}^{n} \binom{y+a}{y}\binom{n-y+b}{n-y} = \sum_{y=0}^{n} \binom{y+a}{a}\binom{n-y+b}{b}$$

Consider $n+a+b+1$ people standing in a line, and suppose we want to select $a+b+1$ of them. Clearly, there are $\binom{n+a+b+1}{a+b+1} = \binom{n+a+b+1}{n}$ ways to do this. Alternatively, we can first choose the $(a+1)^{st}$ person, then choose a people to his left and b people to his right. The $(a+1)^{st}$ person must be in position $a+1$ through $n+a+1$ if there are a people to his left and b to his right in a line with total length $n+a+b+1$. Let $y+a+1$ represent the position of the $(a+1)^{st}$ person, so y ranges from 0 to n. Then, there are $y+a$ people to the left, from whom we choose a, and there are $n-y+b$ people to his right, from whom we choose b. This gives the summation, completing the proof.