Math for Data Science: Problem Set 2

Group: Nick Bornemann, Jishnu Verma

2025-11-11

Due Date: Thursday, November 20 by the end of the day. (The Moodle submission link will become inactive at midnight of November 21.)

Instructions: Please submit one solution set per group and include your group members' names at the top. This time, please write your solutions within this Rmd file, under the relevant question. Please submit the knitted output as a pdf. Make sure to show all code you used to arrive at the answer. However, please provide a brief, clear answer to every question rather than making us infer it from your output, and please avoid printing unnecessary output.

1. The Law of Large Numbers and the Central Limit Theorem

You are an urban planner interested in finding out how many people enter and leave the city using personal vehicles every day. (You're not interested in the number of *cars*; you're interested in the number of *people* who use cars to get to work.) To do this, you decide to collect data from a few different points around the city on how many people there are per car. You already have reliable satellite data on the number of cars that come into the city, so if you get a good estimate of people per car you'll be in good shape.

Collecting data on people per car is costly and you'd love to minimize how many data points you have to collect. However, you're also familiar with the Law of Large Numbers and know that the sample mean converges to the true mean as the sample size n grows large.

a. Let's illustrate this with a small simulation. Suppose the number of people in a car is distributed Poisson with a rate of $\lambda = 2$ people per car.[1] Construct 500 samples from this distribution, with the first sample having $n = 1$ cars, the second $n = 2$ cars, and so on. Compute the average number of people per car in each sample. Plot this on the y-axis against the sample size on the x-axis and run a horizontal blue line through the true mean. Comment on what you see.

b. You collect data on 100 cars and compute the average number of people per car in this sample. Use the Central Limit Theorem to write down the approximate distribution of this quantity.

c. Let's examine this distribution more closely. Generate 10,000 replicates of the sample mean with $n = 100$ and plot a histogram.[2] Are you convinced that the Normal approximation you found in the previous question is good enough? Compare this to $n = 1$, $n = 5$, and $n = 30$, generating a histogram for each. (We're aiming to recreate the second row of Figure 10.5 from Slide 47 of Lecture 4.) Comment on what you observe.

d. Suppose the city government will enact measures to regulate the number of people allowed per car during rush hour if they think the mean is below 1.7 people per car. Using the Normal approximation from part (b) above, find the probability that you get a mean of 1.7 or less in your sample of 100, even though the true mean is 2. (Please give the theoretical answer, not a simulation. You can use `R` as a calculator.) What should you do to ensure that this probability stays below 1%?

[1] I should have mentioned that you're an urban planner in San Francisco, where it's rare but possible to have 0 people in a car.

[2] Try using the `replicate` function rather than a loop, as this will speed things up considerably.

2. Maximum Likelihood

Bangladesh, home to 163 million people, is the world's most populous delta region; one-fourth of the country's land mass is only seven feet above sea level.[3] Although the communities in Bangladesh's low-lying coastal regions have always been vulnerable to catastrophic flooding events, this seems to be happening with growing frequency. Is climate change increasing the occurrence of flooding in Bangladesh?

We often use the Poisson distribution to model (rare) climate events such as earthquakes and hurricanes. So let X_t be the number of major floods in Bangladesh in time period t, and let X_t be distributed:

$$X_t \sim \text{Poisson}(\lambda)$$

a. We observe the following number of floods in Bangladesh per five-year period for the first quarter of the 21st century:

$$\begin{bmatrix} 1 & 2000-2004 \\ 3 & 2005-2009 \\ 1 & 2010-2014 \\ 2 & 2015-2019 \\ 0 & 2020-2024 \end{bmatrix}$$

Please write down the likelihood of this series of events for some unknown λ, assuming the floods in each period are independent and identically distributed.

b. Take the log of the likelihood you wrote down in part (a). Show all steps.

c. Maximize the log-likelihood from part (b) to derive an MLE estimator for λ. Show all steps.

d. Interpret the $\hat{\lambda}$ you found in part (c) in your own words. What is this quantity conceptually, and how do you get it from the data?

e. Show that you found the MLE by plotting the log likelihood on the y-axis against a series of candidate values for λ ranging from 0 to 4 on the x-axis.

(a) Likelihood

Given the flood counts $x = (1, 3, 1, 2, 0)$ and assuming $X_t \sim$ iid Poisson(λ), the likelihood is:

$$L(\lambda \mid x) = \prod_{t=1}^{5} \frac{e^{-\lambda}\lambda^{x_t}}{x_t!}.$$

Substituting the data:

$$L(\lambda) = \frac{e^{-\lambda}\lambda^{1}}{1!} \cdot \frac{e^{-\lambda}\lambda^{3}}{3!} \cdot \frac{e^{-\lambda}\lambda^{1}}{1!} \cdot \frac{e^{-\lambda}\lambda^{2}}{2!} \cdot \frac{e^{-\lambda}\lambda^{0}}{0!}.$$

(b) Log-likelihood

The log-likelihood is:

$$\ell(\lambda) = \sum_{t=1}^{5} \left(-\lambda + x_t \log(\lambda) - \log(x_t!)\right).$$

[3] https://www.nrdc.org/stories/bangladesh-country-underwater-culture-move

Since $\sum x_t = 7$ and $n = 5$, this becomes:

$$\ell(\lambda) = -5\lambda + 7\log(\lambda) - \sum_{t=1}^{5} \log(x_t!).$$

(c) Maximum Likelihood Estimator

Differentiate:

$$\frac{d\ell}{d\lambda} = -5 + \frac{7}{\lambda}.$$

Solve for zero:

$$-5 + \frac{7}{\lambda} = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{7}{5} = 1.4.$$

Second derivative:

$$\frac{d^2\ell}{d\lambda^2} = -\frac{7}{\lambda^2} < 0.$$

(d) Interpretation

$\hat{\lambda} = 1.4$ is the **average number of major floods per 5-year period** based on the years 2000–2024.

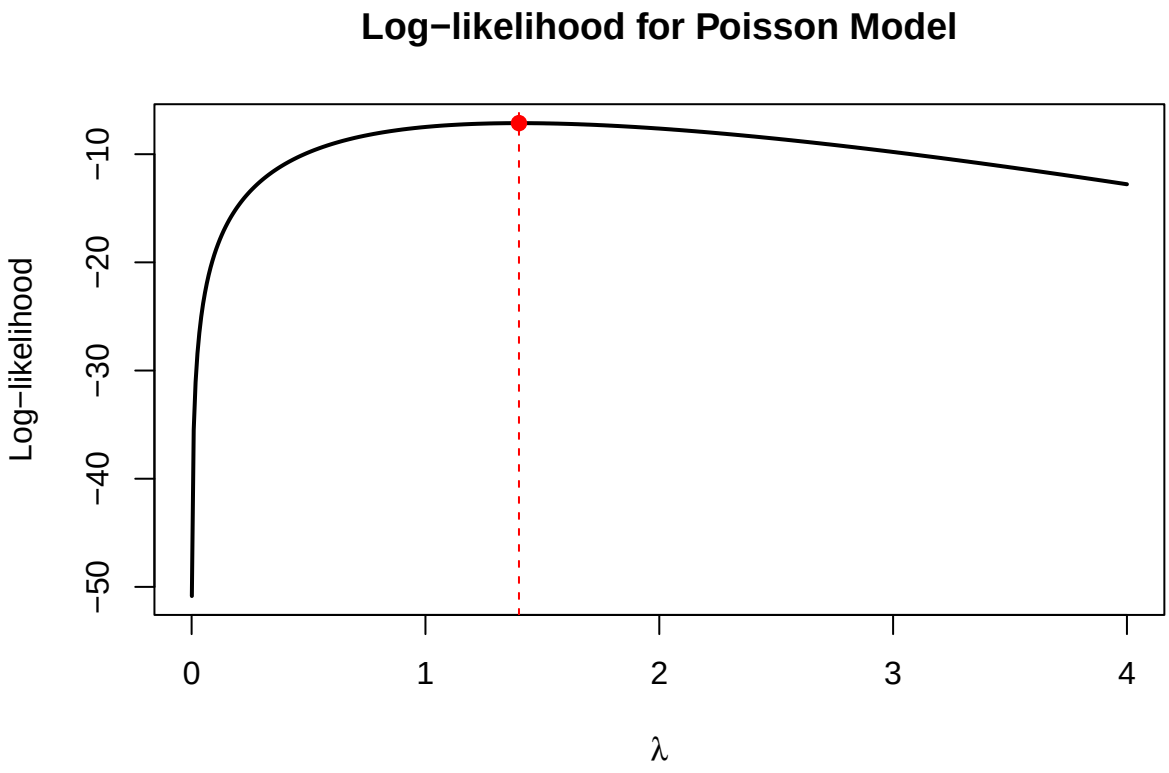
Since the MLE of a Poisson rate equals the sample mean:

$$\hat{\lambda} = \frac{7}{5} = 1.4,$$

this means the observed flood rate is about **1.4 floods per five years**.

```r
x <- c(1,3,1,2,0)

loglik <- function(lambda, x) {
  n <- length(x)
  S <- sum(x)
  const <- sum(lgamma(x + 1))
  -n * lambda + S * log(lambda) - const
}

lambda_grid <- seq(0.001, 4, length.out = 500)
loglik_vals <- sapply(lambda_grid, loglik, x = x)

lambda_hat <- mean(x)

plot(lambda_grid, loglik_vals, type="l", lwd=2,
     xlab=expression(lambda),
     ylab="Log-likelihood",
     main="Log-likelihood for Poisson Model")
```

```r
abline(v = lambda_hat, col="red", lty=2)
points(lambda_hat, loglik(lambda_hat, x), pch=19, col="red")
```



3. Bayesian Analysis

You monitor the presence of a blue–green algae species across freshwater sites. In a sample of $n = 274$ sites, you observe algae present at $y = 44$ sites. Let $\theta \in (0, 1)$ denote the true probability that a randomly selected site has detectable algae.[4]

a. Assume:

$$y \sim \text{Binomial}(n, \theta), \qquad \theta \sim \text{Beta}(\alpha, \beta).$$

Take as a baseline prior $\alpha = 2$, $\beta = 10$. Using Beta–Binomial conjugacy, write down the posterior $p(\theta \mid y, n)$ and identify its parameters.

b. Give the expression (in terms of α, β, y, n) for the posterior mean of θ.

c. Alternatively, we may consider using the priors below:

$\text{Beta}(1, 1)$ (uniform)
$\text{Beta}(0.5, 0.5)$ (Jeffreys-type weak prior)
$\text{Beta}(100, 2)$ (strongly informative, favoring large θ)

Please plot, on a common $\theta \in [0, 1]$ axis, the posterior densities for the four priors (the baseline prior in part

[4]This question is adapted from https://avehtari.github.io/BDA_course_Aalto/assignments/assignment2.html.

(a) and the alternative priors above).

a)

We model the data as

$$y \sim \text{Binom}(n, \theta), \qquad \theta \sim \text{Beta}(\alpha, \beta).$$

b)

By Beta–Binomial conjugacy, the posterior is also Beta–distributed:

$$\theta \mid y, n \sim \text{Beta}(\alpha + y,\ \beta + n - y).$$

---

c)

For our observed data $n = 274$ and $y = 44$, the posterior parameters under each prior are:

Prior	α	β	Posterior α'	Posterior β'	Posterior mean $E[\theta \mid y, n]$
Baseline Beta$(2, 10)$	2	10	46	240	0.1608
Uniform Beta$(1, 1)$	1	1	45	231	0.1630
Jeffreys Beta$(0.5, 0.5)$	0.5	0.5	44.5	230.5	0.1618
Informative Beta$(100, 2)$	100	2	144	232	0.3830

The general expression for the posterior mean is

$$E[\theta \mid y, n] = \frac{\alpha + y}{\alpha + \beta + n}.$$

```
library(tidyverse)
```

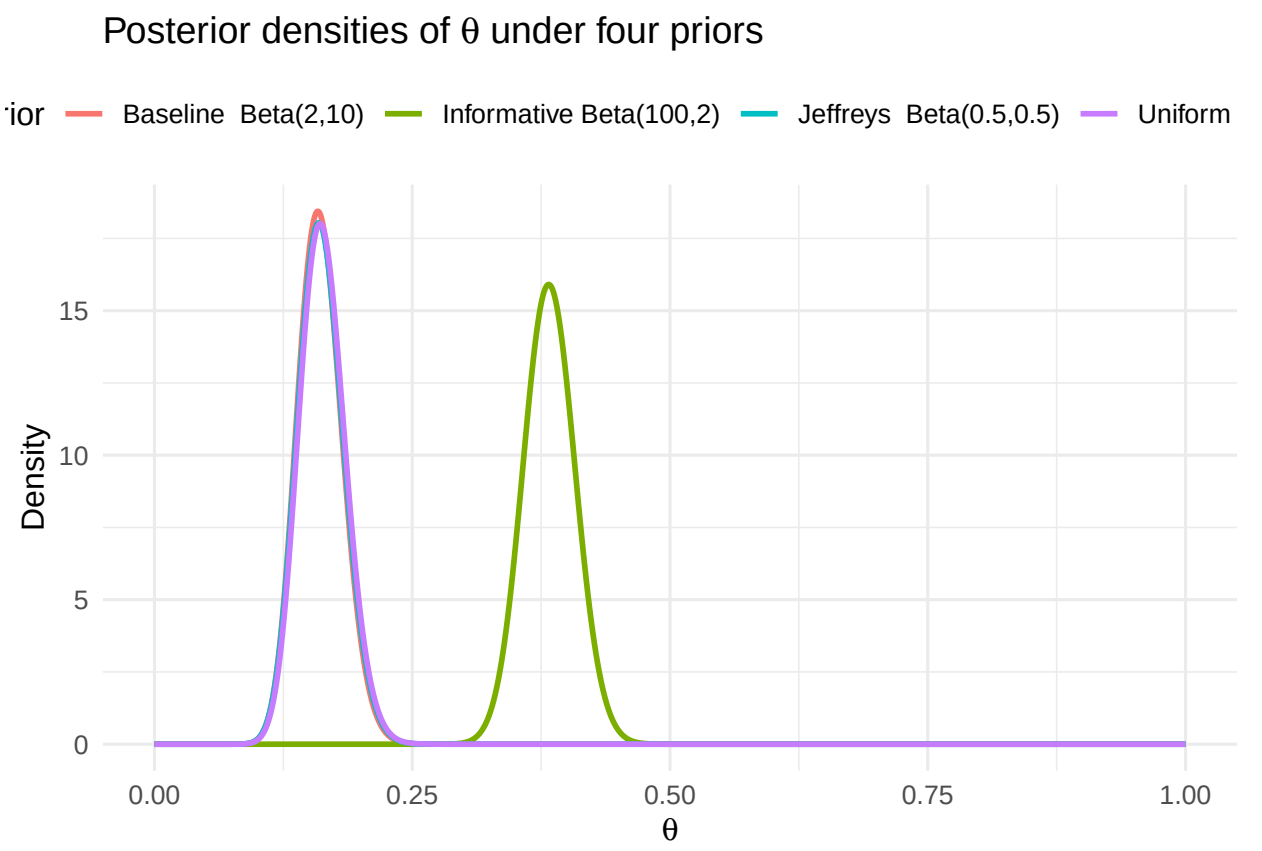
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Data
n <- 274
y <- 44

# Priors
priors <- tribble(
  ~name,                       ~alpha, ~beta,
  "Baseline  Beta(2,10)",        2,     10,
  "Uniform   Beta(1,1)",         1,      1,
  "Jeffreys  Beta(0.5,0.5)",     0.5,    0.5,
  "Informative Beta(100,2)",   100,      2
)

# Posterior params and means
post <- priors %>%
  mutate(
    alpha_post = alpha + y,
    beta_post  = beta  + (n - y),
    post_mean  = (alpha + y) / (alpha + beta + n)
  )

print(post)
```

```
## # A tibble: 4 x 6
##   name                     alpha  beta alpha_post beta_post post_mean
##   <chr>                    <dbl> <dbl>      <dbl>     <dbl>     <dbl>
## 1 Baseline  Beta(2,10)       2    10         46        240      0.161
## 2 Uniform   Beta(1,1)        1     1         45        231      0.163
## 3 Jeffreys  Beta(0.5,0.5)    0.5   0.5       44.5      230.     0.162
## 4 Informative Beta(100,2)  100     2        144        232      0.383
```

```r
# Grid and densities
theta_grid <- seq(0, 1, length.out = 2001)

dens <- post %>%
  rowwise() %>%
  mutate(d = list(tibble(
    theta = theta_grid,
    density = dbeta(theta_grid, alpha_post, beta_post)
  ))) %>%
  unnest(d) %>%
  ungroup()

# Plot
ggplot(dens, aes(theta, density, color = name)) +
  geom_line(linewidth = 1) +
  labs(
    title = expression("Posterior densities of " * theta * " under four priors"),
    x = expression(theta),
    y = "Density",
    color = "Prior"
  ) +
  theme_minimal(base_size = 12) +
```

```
  theme(legend.position = "top")
```



d. In a few sentences, interpret how prior shape and strength influence the posterior relative to the data. Which prior(s) seem the most defensible in this context? If you were interested in monitoring algae presence, what would be your takeaway from this analysis?

The comparison shows that the weak or moderately informative priors (Baseline, Uniform, Jeffreys) have very little influence on the posterior. With a sample of 274 sites, the data dominate the update, and all three priors yield posterior means close to the empirical proportion of algae presence (~16%). The only clear deviation comes from the highly informative **Beta(100,2)** prior, which puts substantial weight on its strong prior belief and therefore pulls the posterior toward much higher values than the data suggest. This highlights that, in practice, prior *strength* matters far more than prior *shape* when the dataset is reasonably large.

In this context, the weakly informative or neutral priors are the most defensible, since there is no strong justification for imposing heavy prior assumptions about algae prevalence. For monitoring purposes, the main takeaway is that the estimated prevalence is quite robust across all reasonable priors. This gives confidence that the underlying detection rate is around 15–16%, and that management or monitoring decisions can reliably be based on this estimate.