Math for Data Science: Problem Set 2

Group: Carl Harper, Eduardo Monte Jorge Hey Martins, Sofia Breganni

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.

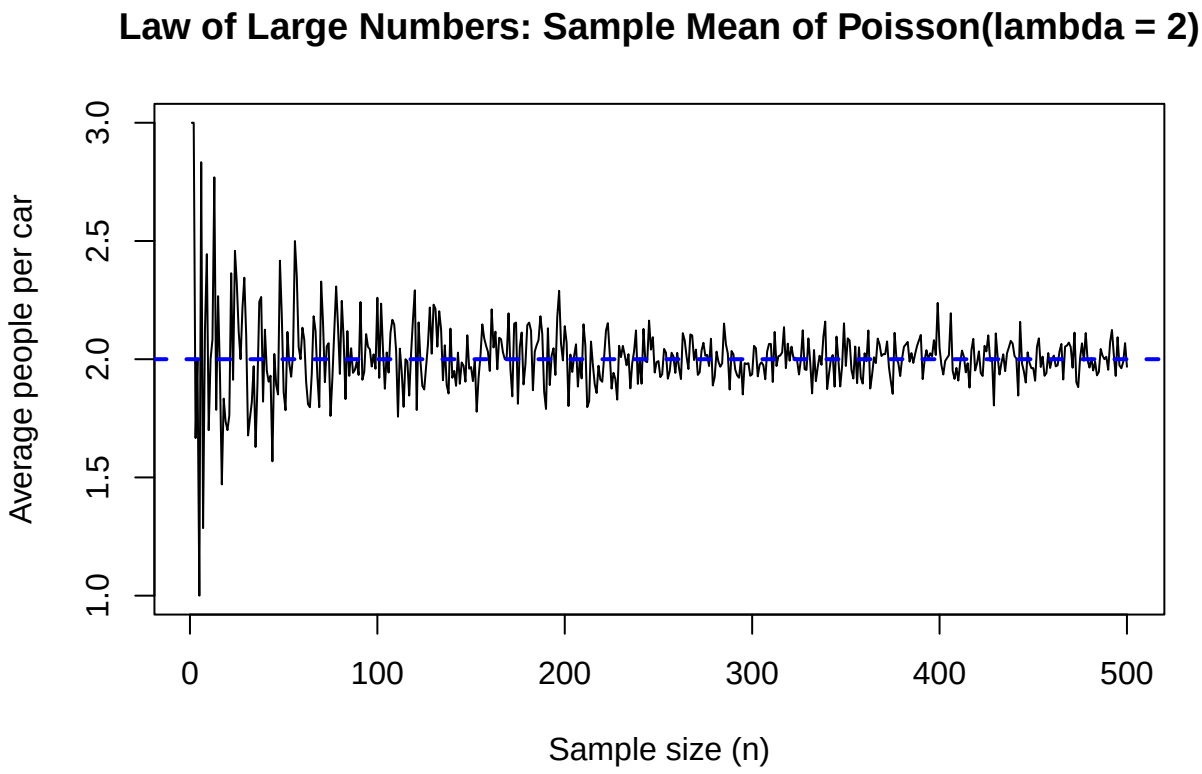
Collecting data on people per car is costly and you'd love to minimize how many data points you have to collect. However, you're also familiar with the Law of Large Numbers and know that the sample mean converges to the true mean as the sample size n grows large.

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.

```
# Set parameters
set.seed(666)
lambda <- 2
max_n <- 500

# Generate sample means for each n
sample_means <- sapply(1:max_n, function(n) {
  mean(rpois(n, lambda))
})

# Plot results
plot(1:max_n, sample_means, type = "l",
```

[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.

```
     xlab = "Sample size (n)",
     ylab = "Average people per car",
     main = "Law of Large Numbers: Sample Mean of Poisson(lambda = 2)")
abline(h = lambda, col = "blue", lwd = 2, lty = 2)
```



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.

We observe the number of people per car for a sample of $n = 100$ cars. Let

$$X_i = \text{number of people in car } i, \quad i = 1, \ldots, 100,$$

and suppose that

$$X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda = 2).$$

For a Poisson(λ) random variable, we have

$$\mathbb{E}[X_i] = \lambda = 2 \quad \text{and} \quad \text{Var}(X_i) = \lambda = 2.$$

The sample mean number of people per car is

$$\bar{X}_{100} = \frac{1}{100}\sum_{i=1}^{100} X_i.$$

By the Central Limit Theorem, the sampling distribution of the sample mean is approximately Normal:

$$\bar{X}_n \approx \mathcal{N}\left(\mathbb{E}[X_i],\ \frac{\mathrm{Var}(X_i)}{n}\right).$$

Substituting $\mathbb{E}[X_i] = 2$, $\mathrm{Var}(X_i) = 2$, and $n = 100$, we obtain

$$\bar{X}_{100} \approx \mathcal{N}\left(2,\ \frac{2}{100}\right) = \mathcal{N}(2,\ 0.02).$$

Equivalently, the standard error of the sample mean is

$$\mathrm{SE}(\bar{X}_{100}) = \sqrt{\frac{2}{100}} = \sqrt{0.02} \approx 0.1414.$$

Thus, the approximate distribution of the average number of people per car in a sample of 100 cars is

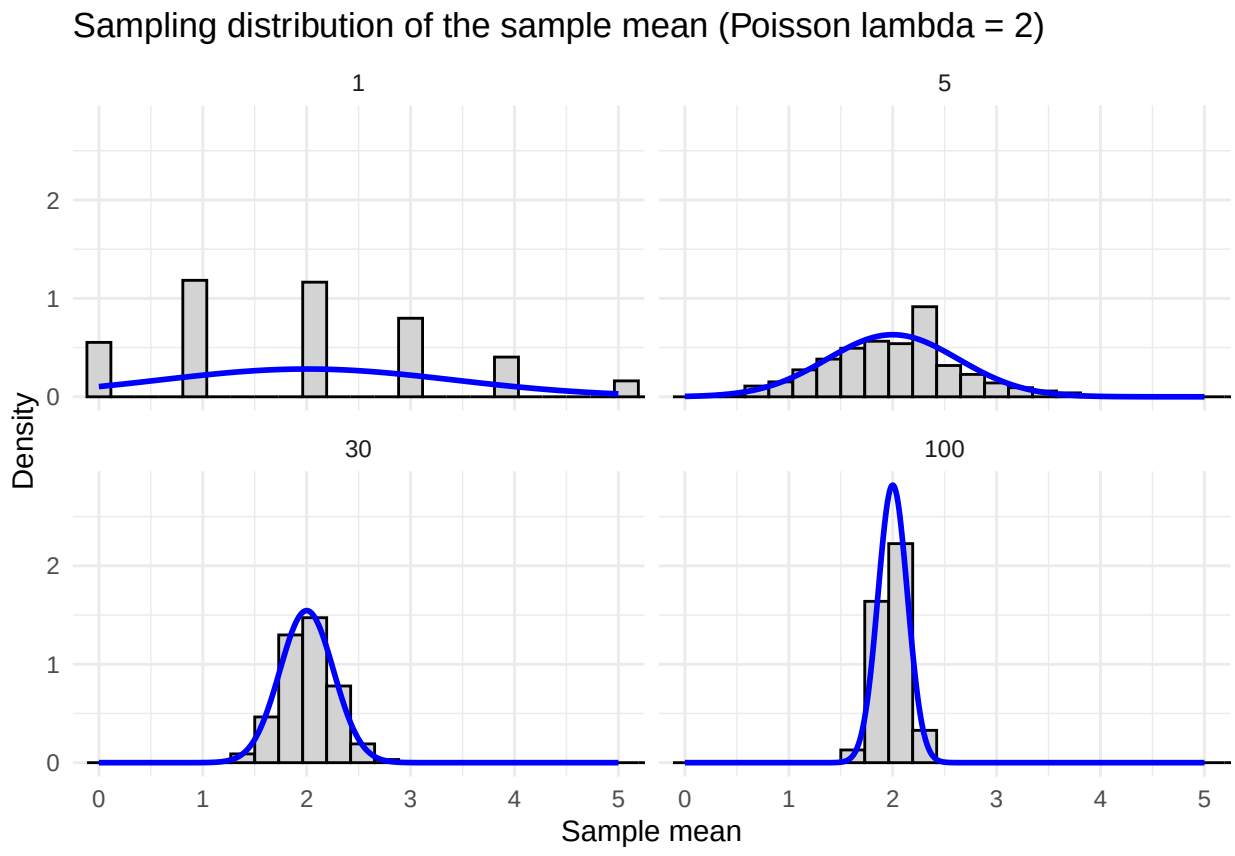
$$\boxed{\bar{X}_{100} \approx \mathcal{N}(2,\ 0.02).}$$

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.

```r
library(ggplot2)

# Setup
set.seed(666)
lambda <- 2
reps   <- 10000
Ns     <- c(1, 5, 30, 100)

# Simulate sample means using replicate()
sim_mat <- sapply(Ns, function(n) replicate(reps, mean(rpois(n, lambda))))
colnames(sim_mat) <- paste0("n=", Ns)

# Tidy for plotting
df <- data.frame(
  mean = as.vector(sim_mat),
  n    = factor(rep(Ns, each = reps), levels = Ns)
)

# Normal curves to overlay
xgrid <- seq(0, 5, by = 0.01)
norm_curves <- do.call(rbind, lapply(Ns, function(n) {
```

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

```
  data.frame(
    x = xgrid,
    y = dnorm(xgrid, mean = lambda, sd = sqrt(lambda / n)),
    n = factor(n, levels = Ns)
  )
}))

ggplot(df, aes(x = mean)) +
  geom_histogram(aes(y = after_stat(density)), bins = 40, fill = "lightgray", color = "black") +
  geom_line(data = norm_curves, aes(x = x, y = y), linewidth = 1, color = "blue") +
  facet_wrap(~ n, nrow = 2) +
  coord_cartesian(xlim = c(0, 5)) +
  labs(title = "Sampling distribution of the sample mean (Poisson lambda = 2)",
       x = "Sample mean", y = "Density") +
  theme_minimal()
```



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%?

```
# X_i ~ Poisson(lambda = 2)
# Under the CLT, the sample mean is approximately Normal:
# Xbar_n ~ Normal(mean=lambda, variance = lambda / n)
```

```r
# For n = 100:
lambda <- 2
n <- 100
mean_true <- lambda
sd_true <- sqrt(lambda / n)

# The city will act if the sample mean equal or less to 1.7
threshold <- 1.7

# Compute the probability of observing a mean equal or less 1.7
p_value <- pnorm(threshold, mean = mean_true, sd = sd_true)
p_value
```

```
## [1] 0.01694743
```

```r
# app. 0.0169 or 1.7%

# Interpretation output:
```

Using the Central Limit Theorem, we approximate the sampling distribution of the sample mean as

$$\bar{X}_{100} \approx \mathcal{N}\left(2,\ \frac{2}{100}\right).$$

The probability that the observed mean is 1.7 or lower, even though the true mean is 2, is

$$\Pr(\bar{X}_{100} \leq 1.7) = \Phi\left(\frac{1.7 - 2}{\sqrt{2/100}}\right) = 0.0169.$$

This corresponds to about 1.7%, meaning there is a 1.7% chance of falsely concluding that the mean is below 1.7.

```r
# To ensure this false-alarm probability is below 1%, we solve for n:
target_alpha <- 0.01
z_alpha <- qnorm(1 - target_alpha)
n_required <- ceiling(lambda / ((0.3 / z_alpha)^2))
n_required
```

```
## [1] 121
```

```r
# Interpretation output:
```

To reduce this probability below 1%, we need at least $n_{\text{required}} = 121$ observations.

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

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

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.

The probability function for a Poisson distribution is:

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Therefore, the likelihood function for the series of events is:

$$L(\lambda) = \prod_{t=1}^{n} P(X_t = x_t) = \prod_{t=1}^{n} \left(\frac{e^{-\lambda}\lambda^{x_t}}{x_t!} \right).$$

Adding in the values for x:

$$\begin{aligned} L(\lambda) = \prod_{t=1}^{5} \left(\frac{e^{-\lambda}\lambda^{x_t}}{x_t!} \right) &= \left(\frac{e^{-\lambda}\lambda^1}{1!} \right) \left(\frac{e^{-\lambda}\lambda^3}{3!} \right) \left(\frac{e^{-\lambda}\lambda^1}{1!} \right) \left(\frac{e^{-\lambda}\lambda^2}{2!} \right) \left(\frac{e^{-\lambda}\lambda^0}{0!} \right) \\ &= \left(\frac{e^{-\lambda}\lambda}{1} \right) \left(\frac{e^{-\lambda}\lambda^3}{6} \right) \left(\frac{e^{-\lambda}\lambda}{1} \right) \left(\frac{e^{-\lambda}\lambda^2}{2} \right) \left(\frac{e^{-\lambda}}{1} \right) \\ &= \left(\frac{e^{-5\lambda}\lambda^7}{12} \right) \end{aligned}$$

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

$$\begin{aligned} \ell(\lambda) &= \log L(\lambda) \\ &= \log \left(\prod_{t=1}^{n} \frac{e^{-\lambda}\lambda^{x_t}}{x_t!} \right) \\ &= \sum_{t=1}^{n} \log \left(\frac{e^{-\lambda}\lambda^{x_t}}{x_t!} \right) \\ &= \sum_{t=1}^{n} (-\lambda + x_t \log \lambda - \log(x_t!)) \\ &= -n\lambda + \log \lambda \sum_{t=1}^{n} x_t - \sum_{t=1}^{n} \log(x_t!) \end{aligned}$$

Adding in the values for x:

$$\begin{aligned}\ell(\lambda) &= \log\lambda \sum_{t=1}^{5} x_t - 5\lambda - \sum_{t=1}^{5} \log(x_t!) \\ &= -5\lambda + \log\lambda(1+3+1+2+0) - (\log(1) + \log(6) + \log(1) + \log(2) + \log(1)) \\ &= -5\lambda + 7\log\lambda - 2.48\end{aligned}$$

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

In order to maximize the log-likelihood function we must set $\frac{d\ell(\lambda)}{d\lambda} = 0$
Where $\ell(\lambda) = -n\lambda + \log\lambda \sum_{t=1}^{n} x_t - \sum_{t=1}^{n} \log(x_t!)$

$$\frac{d\ell(\lambda)}{d\lambda} = -n + \frac{1}{\lambda}\sum_{t=1}^{n} x_t = 0$$

$$\frac{1}{\lambda}\sum_{t=1}^{n} x_t = n$$

$$\frac{1}{\lambda} = \frac{n}{\sum_{t=1}^{n} x_t}$$

$$\lambda = \frac{\sum_{i=1}^{n} x_t}{n}.$$

Adding in the values for x:

$$\lambda = \frac{\sum_{i=1}^{n} x_t}{n} = \frac{1+3+1+2+0}{7} = \frac{7}{5} = 1.4$$

The maximum likelihood estimator is $\hat{\lambda} = \frac{\sum_{t=1}^{n} x_t}{n} = 1.4$.

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?

$\hat{\lambda}$ is the sample mean of the counts and (for the Poisson model) the MLE of the Poisson rate: it estimates the average number of major floods per five-year period.

Because each observation represents a five-year period, the estimated rate of 1.4 floods per period corresponds to roughly 0.28 major floods per year.

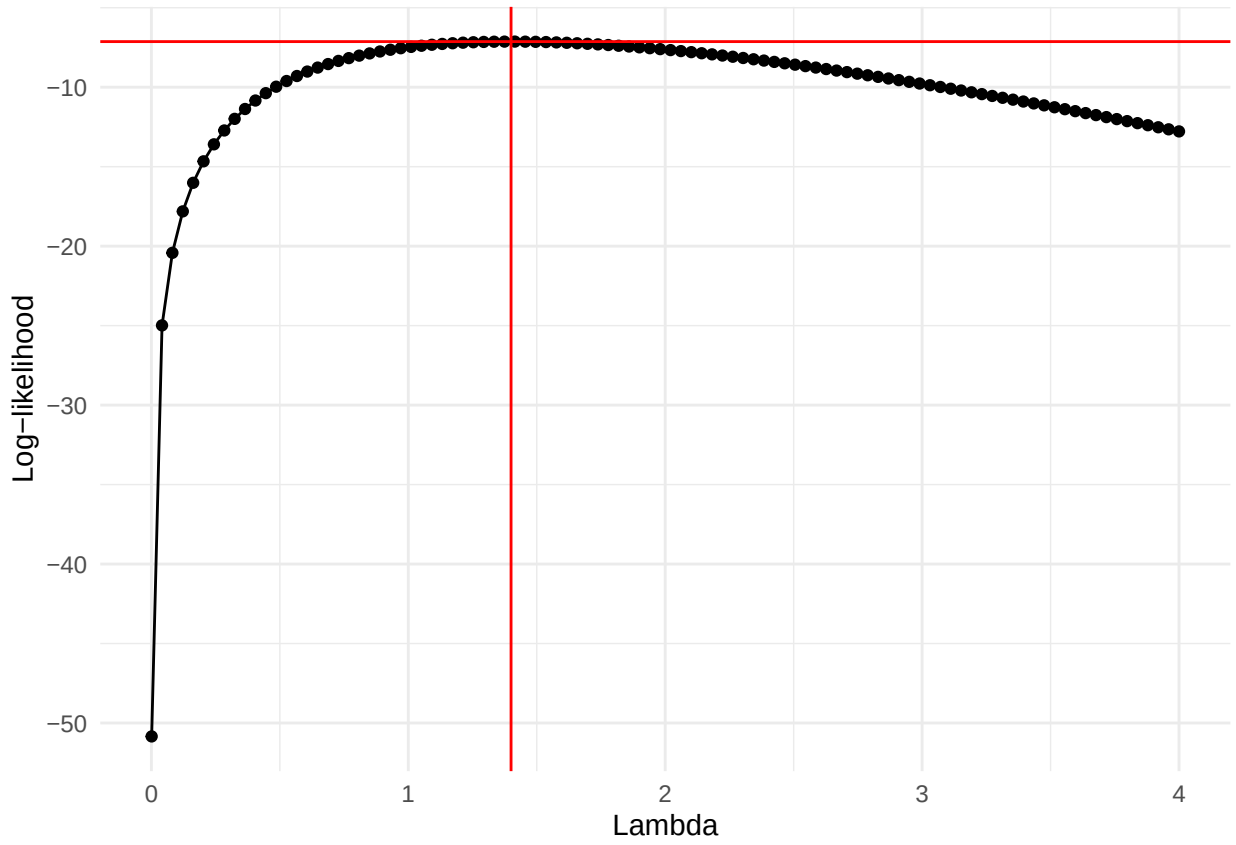
e. Show that you found the MLE by plotting the log likelihood on the y-axis against a series of candidate values for λ ranging from 0 to 4 on the x-axis.

```
# Setup
set.seed(666)
flood_count <- c(1, 3, 1, 2, 0)

# Log likelihood function
likelihood <- function(x, l){
  lik <- sum(x * log(l) - l - log(factorial(x)))
```

```r
  return(lik)
}

likelihood(flood_count, l = 1.4) # -7.129601
```

```
## [1] -7.129601
```

```r
# Plot setup
lik_df <- data.frame(lambda = seq(0.001, 4, length.out = 100), # 0.001 to avoid NaN
                     likelihood = NA)
lik_df$likelihood <- sapply(X = lik_df$lambda, FUN = likelihood, x = flood_count)

# Log likelihood plot
library(ggplot2)
ggplot(lik_df, aes(x = lambda, y = likelihood)) +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = max(lik_df$likelihood), color = "red") +
  geom_vline(xintercept = 1.4, color = "red") +
  xlab("Lambda") +
  ylab("Log-likelihood") +
  theme_minimal()
```



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.

If prior distribution is a beta distribution, so is the posterior

$$\begin{aligned} p(\theta|y,n) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1} \\ p(\theta|44,274) &\propto \theta^{2+44-1}(1-\theta)^{10+274-44-1} \\ &\propto \theta^{45}(1-\theta)^{239} \end{aligned}$$

this is the kernel of the beta distribution with updated parameters

$$\begin{aligned} &\alpha = \alpha + y,\ \beta = \beta + n - y \\ &\alpha = 2 + 44,\ \beta = 10 + 274 - 44 \\ &\alpha = 46,\ \beta = 240 \\ &\quad = Beta(\theta; 46, 240) \end{aligned}$$

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

$$E(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

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

$$\text{Beta}(1, 1) \quad \text{(uniform)}$$

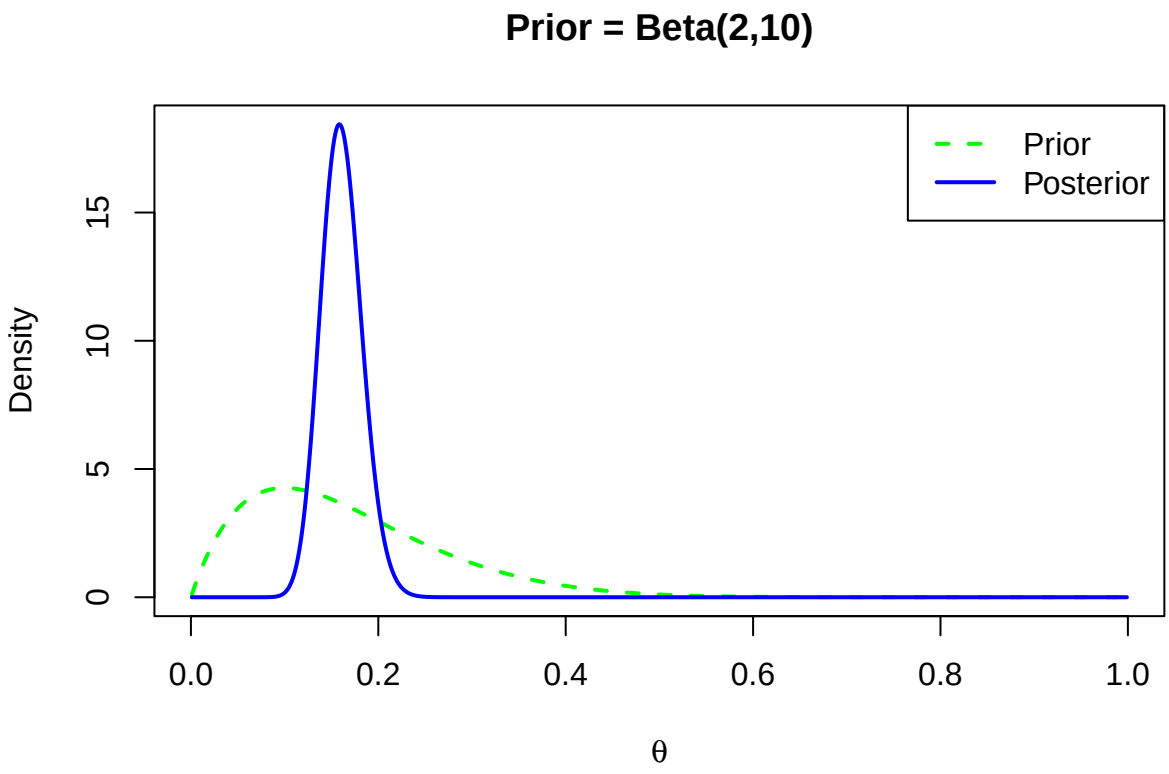
$$\text{Beta}(0.5, 0.5) \quad \text{(Jeffreys-type weak prior)}$$

$$\text{Beta}(100, 2) \quad \text{(strongly informative, favoring large } \theta\text{)}$$

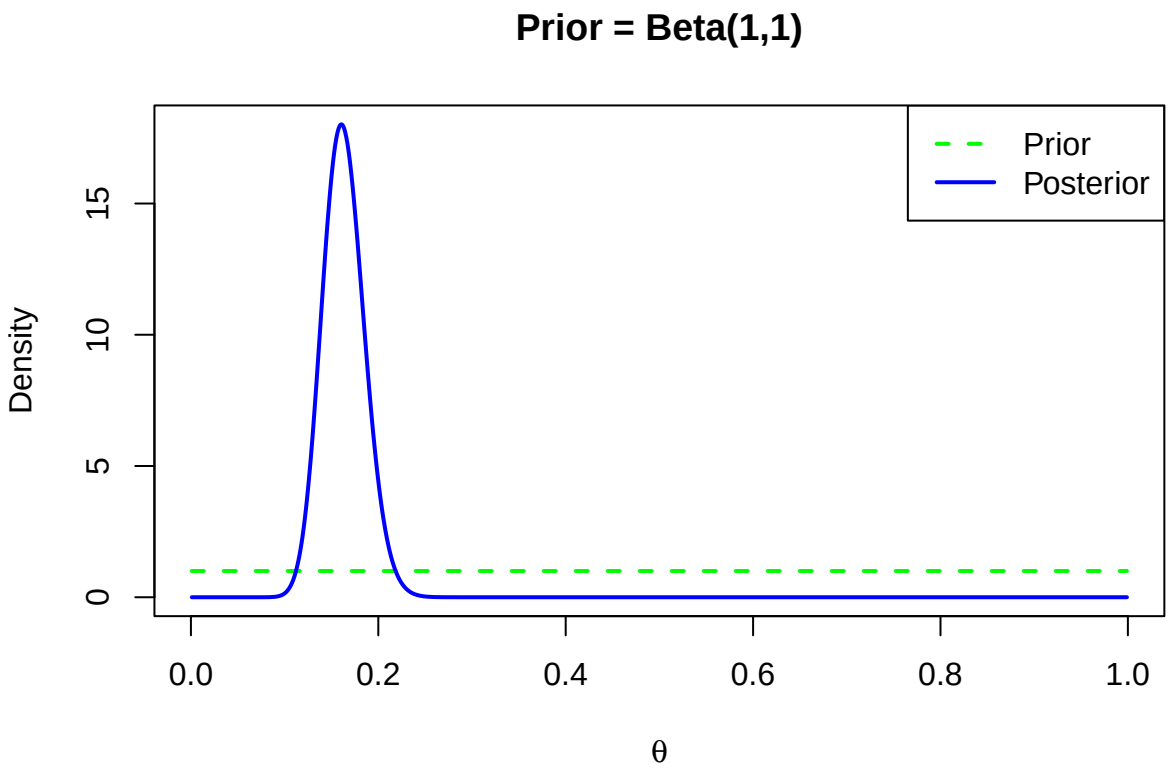
`Please plot, on a common $\theta \in [0,1]$ axis, the posterior densities for the four priors (the baseline prior in part (a) and the alternative priors above).`

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

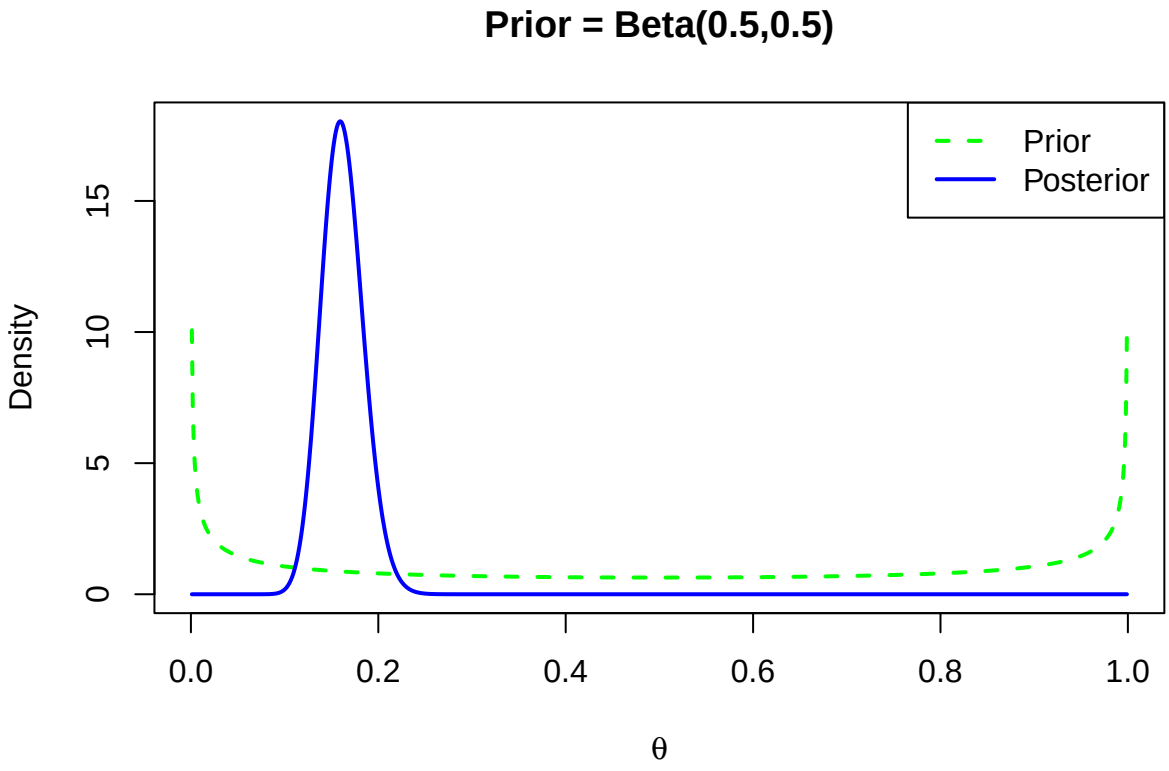
```r
S <- 44  # number of sites with algae
n <- 274 # total number of sites


theta <- seq(0.001, 0.999, length.out = 1000) #allows for jeffrey's type prior

#plotting the prior and the posterior
plot_posterior_distr <- function(a, b){
  # Compute prior density
  prior_dens <- dbeta(theta, a, b)

  # Compute posterior density
  post_dens <- dbeta(theta, a + S, b + n - S)

  # Prepare empty plot with appropriate limits
  plot(theta, prior_dens, type="n",
       ylim=c(0, max(prior_dens, post_dens)),
       xlab=expression(theta),
       ylab="Density",
       main=paste("Prior = Beta(", a, ",", b,")", sep=""))

  # Plot prior (dashed line)
  lines(theta, prior_dens, col="green", lty=2, lwd=2)

  # Plot posterior (solid line)
  lines(theta, post_dens, col="blue", lty=1, lwd=2)

  # Add legend
  legend("topright",
         legend=c("Prior", "Posterior"),
         col=c("green", "blue"),
         lty=c(2, 1),
         lwd=c(2, 2))
}

# Plot with baseline prior
plot_posterior_distr(2, 10)
```



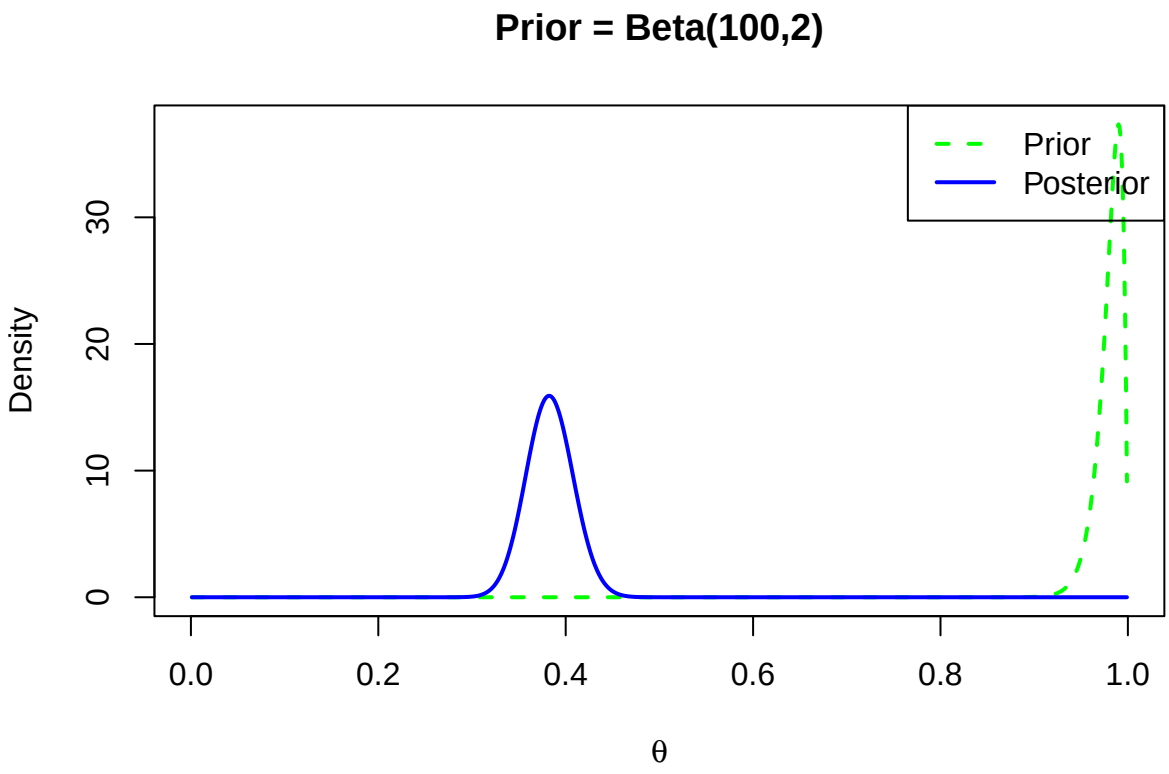

```
plot_posterior_distr(1,1)
```




```
plot_posterior_distr(0.5,0.5)
```




```
plot_posterior_distr(100,2)
```




```r
a_post <- 46
b_post <- 240

# 95% Credible Interval, helps answer question
lower <- qbeta(0.025, a_post, b_post)
upper <- qbeta(0.975, a_post, b_post)

cat("95% Credible Interval: [", round(lower, 3), ",", round(upper, 3), "]\n")
```

```
## 95% Credible Interval: [ 0.121 , 0.206 ]
```

```r
# Posterior mean
post_mean <- a_post / (a_post + b_post)
cat("Posterior Mean:", round(post_mean, 3), "\n")
```

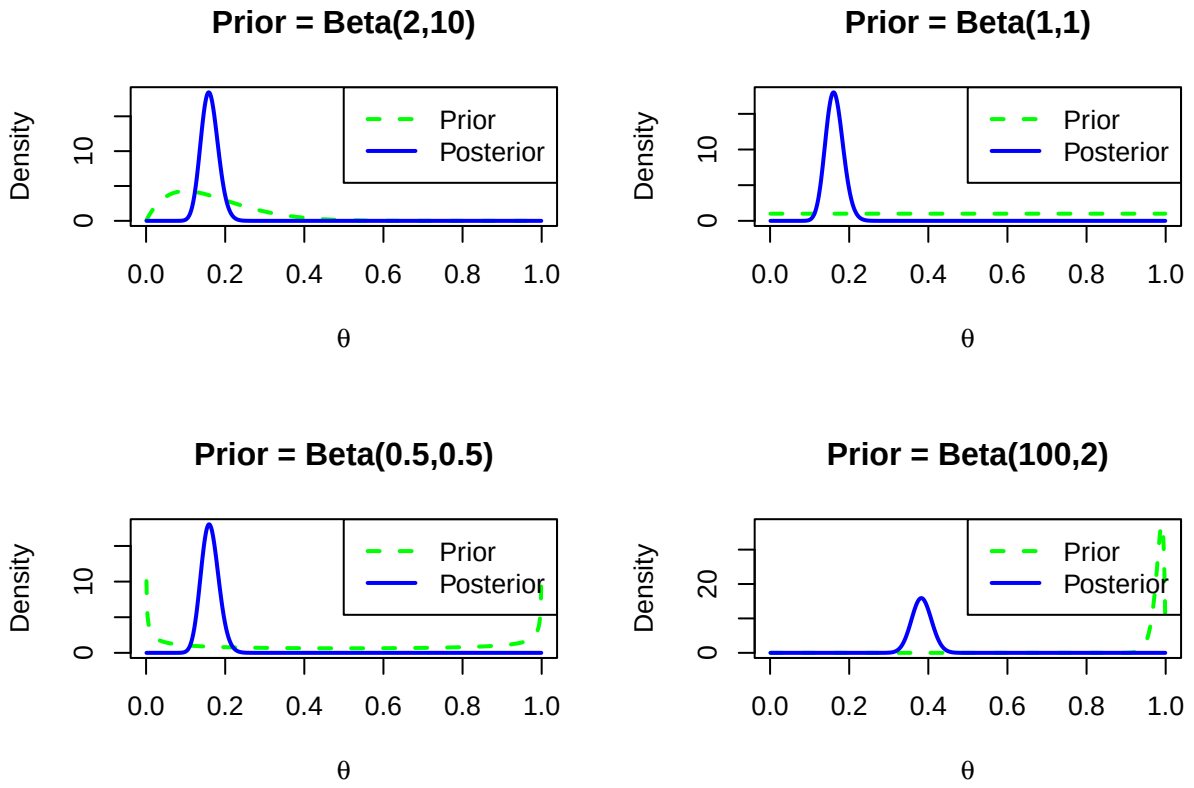
```
## Posterior Mean: 0.161
```

```r
# Posterior mode (peak of distribution)
post_mode <- (a_post - 1) / (a_post + b_post - 2)
cat("Posterior Mode:", round(post_mode, 3), "\n")
```

```
## Posterior Mode: 0.158
```

Plotting all the distributions side-by-side:

```r
par(mfrow=c(2,2))

# Plot all four priors
plot_posterior_distr(2, 10)      # Baseline
plot_posterior_distr(1, 1)       # Uniform
plot_posterior_distr(0.5, 0.5)   # Jeffreys-type
plot_posterior_distr(100, 2)     # Strongly informative
```




```r
# Reset to single plot
par(mfrow=c(1,1))
```

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? From comparing the four priors, **prior strength** (the sum $\alpha + \beta$) has the most substantial impact on the posterior. The three weak priors Beta$(2, 10)$, Beta$(1, 1)$, and Beta$(0.5, 0.5)$ all have small values of $\alpha + \beta$ (12, 2, and 1 respectively) and yield nearly identical posteriors centered around the data proportion $44/274 \approx 0.16$. Their posterior means all lie close to 0.16, despite having different prior means.This demonstrates that when priors are weak, the data dominates the inference. In other words, the fewer assumptions you encode in the prior, the more the data speaks for itself. In contrast, the strong prior Beta$(100, 2)$, with $\alpha + \beta = 102$, produces a noticeably different posterior, with a posterior mean of approximately 0.38. which shifts the distribution considerably to the right. The three weak priors are the most defensible in this context. They align with an objective/ reference prior (e.g., Jeffreys prior) and allow the data to drive the inference, consistently suggesting that algae are present at roughly $12\% - 20\%$ of sites. The large Beta$(100, 2)$ prior would only be defensible if there were strong external evidence indicating a very

high (a figure close to its high prior mean) and true algae prevalence, that is not reflected in the data here.