Lab Exercises 3

28/01/24

Question 1

Consider the happiness example from the lecture, with 118 out of 129 women indicating they are happy. We are interested in estimating θ, which is the (true) proportion of women who are happy. Calculate the MLE estimate $\hat{\theta}$ and 95% confidence interval.

```
happy_count <- 118
total_count <- 129


# MLE for the proportion theta
theta_hat <- happy_count / total_count

# 95% Confidence Interval
SE <- sqrt((theta_hat * (1 - theta_hat)) / total_count)
z <- qnorm(0.975)

CI_lower <- theta_hat - z * SE
CI_upper <- theta_hat + z * SE

# Output the results
cat("MLE of theta: ", theta_hat, "\n")
```

```
MLE of theta:  0.9147287
```

```
cat("95% Confidence Interval: [", CI_lower, ", ", CI_upper, "]\n")
```

```
95% Confidence Interval: [ 0.8665338 ,  0.9629236 ]
```

Question 2

Assume a Beta(1,1) prior on θ. Calculate the posterior mean for $\hat{\theta}$ and 95% credible interval.

```r
# Prior parameters for Beta distribution (Beta(1,1))
alpha_prior <- 1
beta_prior <- 1

# Posterior parameters
alpha_post <- alpha_prior + happy_count
beta_post <- beta_prior + total_count - happy_count

# Posterior mean
posterior_mean <- alpha_post / (alpha_post + beta_post)

# 95% Credible Interval
CI_lower <- qbeta(0.025, alpha_post, beta_post)
CI_upper <- qbeta(0.975, alpha_post, beta_post)

# Output the results
cat("Posterior mean of theta: ", posterior_mean, "\n")
```

```
Posterior mean of theta:  0.9083969
```

```r
cat("95% Credible Interval: [", CI_lower, ", ", CI_upper, "]\n")
```

```
95% Credible Interval: [ 0.8536434 ,  0.9513891 ]
```

Question 3

Now assume a Beta(10,10) prior on θ. What is the interpretation of this prior? Are we assuming we know more, less or the same amount of information as the prior used in Question 2?

A Beta(10, 10) prior suggests that, before observing any data, we believe that θ is likely to be around 0.5, with a moderate degree of certainty. This prior is symmetric and centered at 0.5, much like a Beta(1, 1) prior, but it is more concentrated around the center. The parameters of the new prior (both being 10) indicate that there is a stronger belief in θ being near the center of the distribution. This can be interpreted as if we have observed 10 successes and 10 failures in prior experiments or studies.

The Beta(10, 10) prior suggests a specific belief (centered around 0.5) with a higher degree of confidence than the Beta(1, 1) prior. Essentially, the Beta(10, 10) prior is less diffuse and more informative compared to the Beta(1, 1) prior. Therefore, we are assuming more information than beta(1,1).

Question 4

Create a graph in ggplot which illustrates

- The likelihood (easiest option is probably to use `geom_histogram` to plot the histogram of appropriate random variables)
- The priors and posteriors in question 2 and 3 (use `stat_function` to plot these distributions)

Comment on what you observe.

```
library(ggplot2)
library(scales)

# Set the constants
trials <- total_count
successes <- happy_count

# Prior parameters for Beta distributions
alpha_prior1 <- 1
beta_prior1 <- 1
alpha_prior2 <- 10
beta_prior2 <- 10

# Posterior parameters for Beta distributions
alpha_post1 <- alpha_prior1 + happy_count
beta_post1 <- beta_prior1 + total_count - happy_count
alpha_post2 <- alpha_prior2 + happy_count
beta_post2 <- beta_prior2 + total_count - happy_count

# Generating theta values
theta_values <- seq(0, 1, length.out = 1000)

# Likelihood function (Binomial) given successes
likelihood <- dbinom(successes, trials, theta_values)

# Prior distributions
```

```r
prior1 <- dbeta(theta_values, alpha_prior1, beta_prior1)
prior2 <- dbeta(theta_values, alpha_prior2, beta_prior2)

# Posterior distributions
posterior1 <- dbeta(theta_values, alpha_post1, beta_post1)
posterior2 <- dbeta(theta_values, alpha_post2, beta_post2)

# Create a data frame for plotting
plot_data <- data.frame(
  theta = rep(theta_values, 5),
  density = c(likelihood, prior1, prior2, posterior1, posterior2),
  distribution = factor(rep(c('Likelihood', 'Prior Beta(1,1)', 'Prior Beta(10,10)', 'Poste
)

# Plotting
ggplot(plot_data, aes(x = theta, y = density, color = distribution)) +
  geom_line() +
  scale_color_manual(values = c('black', 'blue', 'green', 'red', 'orange')) +
  labs(title = 'Likelihood, Priors, and Posteriors',
       x = 'Theta',
       y = 'Density') +
  theme_minimal() +
  theme(legend.title = element_blank())
```

Likelihood, Priors, and Posteriors
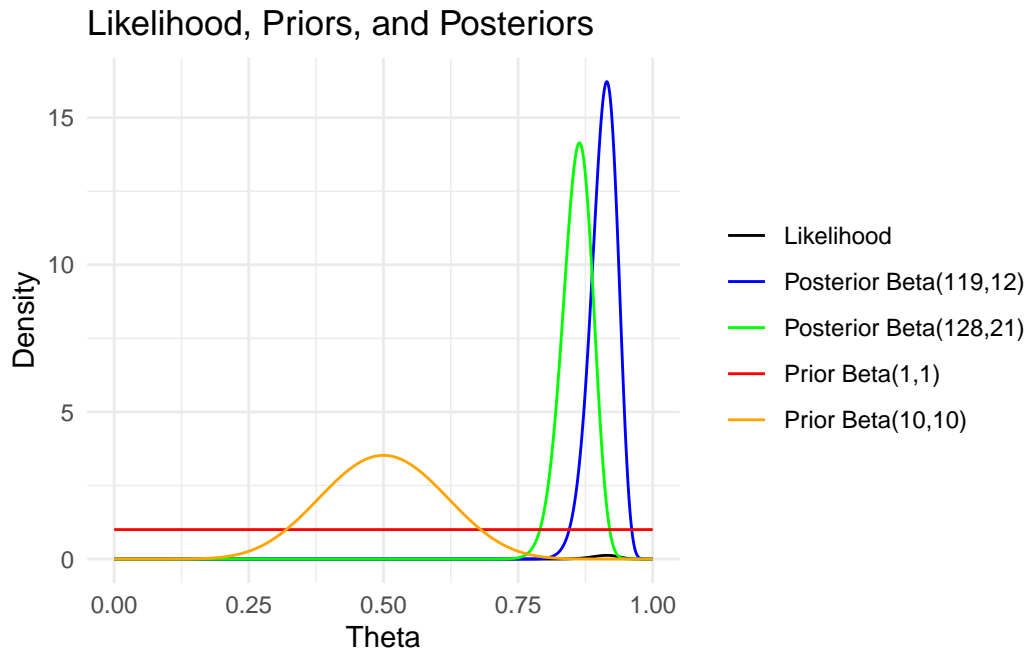
The prior with Beta(10,10) has more information than the other prior with Beta(1,1) because it has higher certainty about θ being around 0.5 which is more specific quantity. Therefore, with the same likelihood, the posterior based on Beta(10,10) moves less to the right compared with the posterior based on Beta(1,1).

Question 5

Laplace was interested in calculating the probability that observing a male birth was less than 0.5, given data he observed in Paris. Calculate this probability, assuming a uniform prior on observing a male birth and using data given in the slides.

```
# Number of observed male and female births
male_births <- 251527
female_births <- 241945

# Prior parameters for Beta distribution (uniform prior)
alpha_prior <- 1
beta_prior <- 1

# Posterior parameters (since uniform prior, just add the counts to 1)
alpha_post <- alpha_prior + male_births
```

```r
beta_post <- beta_prior + female_births

# Calculate the probability that observing a male birth is less than 0.5
prob <- pbeta(0.5, alpha_post, beta_post)

# Output the result
prob
```

```
[1] 1.146058e-42
```

Question 6

(No R code required) A study is performed to estimate the effect of a simple training program on basketball free-throw shooting. A random sample of 100 college students is recruited into the study. Each student first shoots 100 free-throws to establish a baseline success probability. Each student then takes 50 practice shots each day for a month. At the end of that time, each student takes 100 shots for a final measurement. Let θ be the average improvement in success probability. θ is measured as the final proportion of shots made minus the initial proportion of shots made.

Given two prior distributions for θ (explaining each in a sentence):

- A noninformative prior, and
- A subjective/informative prior based on your best knowledge

Noninformative Prior:

This type of prior distribution expresses no preference for any particular value of θ and aims to have minimal impact on the posterior inferences. It essentially allows the data to speak for itself. An example of a noninformative prior for a proportion like θ is a uniform distribution, which could be represented by a Beta(1,1) distribution, indicating equal probability for all values of θ between 0 and 1.

Subjective/Informative Prior:

A subjective or informative prior distribution incorporates specific knowledge or beliefs about the parameter before observing the current data. For instance, if the researcher has prior experience or existing data suggesting that the training program typically improves free-throw success by about 10%, they might choose a Beta distribution centered around 0.1, such as Beta(2,18), reflecting a belief that the improvement is likely to be around 10% but allowing for some variation based on the new data.