Week 3: Intro to Bayes and git branches

YUHONG ZHANG

29/01/24

Branches on git

Branches on git are useful when you have more than one person working on the same file, or when you are experimenting with different code etc that may not work. So far we've just been pushing to the 'main' branch, but you can also create other branches within your repo, do some work, save and push, and then if you're happy, merge that work back into the 'main' branch. The idea is that the 'main' branch is always kept clean and working, while other branches can be tested and deleted.

Before merging work into the main branch, it's good practice to do a 'pull request' – this flags that you want to make changes, and alerts someone to review your code to make sure it's all okay.

For this week, I would like you to save this .qmd file to your class repo, then create a new branch to make your edits to the file. Then, once you are happy with this week's lab submission, on GitHub, create a 'pull request' and assign me to be the reviewer.

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.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.4     v tibble    3.2.1
```

```
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- 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 becor
```

```r
n <- 129
x <- 118
theta_hat <- x / n
se <- sqrt((theta_hat * (1 - theta_hat)) / n)
lower_bound <- theta_hat - 1.96 * se
upper_bound <- theta_hat + 1.96 * se
conf_interval <- c(lower_bound, upper_bound)
theta_hat
```

```
[1] 0.9147287
```

```r
conf_interval
```

```
[1] 0.8665329 0.9629244
```

The MLE estimate $\hat{\theta}$ is 0.9147287 and 95% confidence interval is (0.8665329,0.9629244).

Question 2

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

```r
alpha_prior <- 1
beta_prior <- 1
alpha_post <- alpha_prior + x
beta_post <- beta_prior + n - x
posterior_mean <- alpha_post / (alpha_post + beta_post)
credible_interval <- qbeta(c(0.025, 0.975), alpha_post, beta_post)
posterior_mean
```

```
[1] 0.9083969
```

```
credible_interval
```

```
[1] 0.8536434 0.9513891
```

The posterior mean for $\hat{\theta}$ is 0.9083969 and 95% credible interval is (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?

```
alpha_post1 <- 10 + x
beta_post1 <- 10 + n - x
posterior_mean1 <- alpha_post1 / (alpha_post1 + beta_post1)
credible_interval1 <- qbeta(c(0.025, 0.975), alpha_post1, beta_post1)
posterior_mean1
```

```
[1] 0.8590604
```

```
credible_interval1
```

```
[1] 0.7990363 0.9099708
```

The interpretation of prior on θ is that before data is observed, the unknown parameter is modeled as a random variable θ follows a distribution Beta(10,10), we assuming we know more in this question, since more people will be observed in this case (20-2=18 more women) and also in Question 2, the prior is Beta(1,1) which is flat and we consider all values of θ equally, so we will know more information when Beta(10,10) is prior on θ.

Question 4

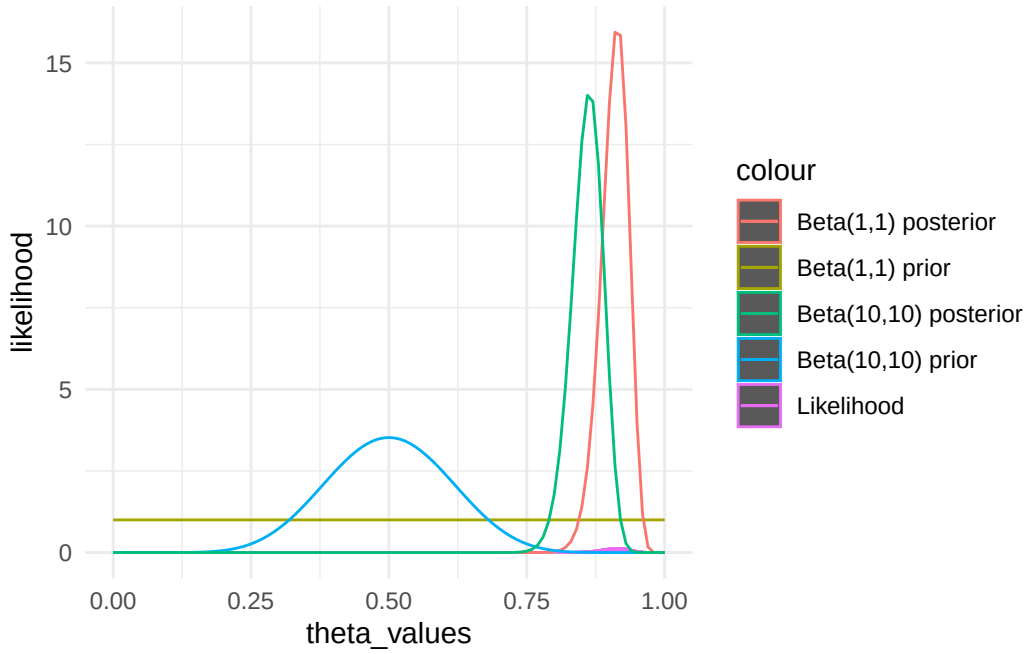
Create a graph in ggplot which illustrates

- 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.

```
theta_values <- seq(0, 1, by = 0.001)
likelihood <- dbinom(x, size = n, prob = theta_values)
data <- data.frame(theta = theta_values, likelihood = likelihood)
ggplot(data, aes(x = theta_values, y = likelihood)) +
  geom_histogram(stat = "identity",aes(color = "Likelihood")) +
  stat_function(fun = dbeta, args = list(shape1 = 1, shape2 = 1), aes(color = "Beta(1,1) p
  stat_function(fun = dbeta, args = list(shape1 = 118+1, shape2 = 1 + n - 118), aes(color
    stat_function(fun = dbeta, args = list(shape1 = 10, shape2 = 10), aes(color = "Beta(10
  stat_function(fun = dbeta, args = list(shape1 = 118+10, shape2 = 10 + n - 118), aes(colo
  theme_minimal()
```

```
Warning in geom_histogram(stat = "identity", aes(color = "Likelihood")):
Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```



From the plot, we can find that the peak of likelihood around 0.9 and shows a slightly left skewed trend. For Beta(1,1) prior, the plot is flat, which implies that before observing data, we consider all values of θ equally, however, for Beta(10,10) prior, we can find the peak of it is at θ=0.5, which shows that women indicating they are happy are most likely around 50% and be more informative. Moreover, we noticed that after observing data, both posteriors skewed and shifted to around 0.9 which is closer to the peak of likelihood, and the Beta(1,1) posterior is much closer to the likelihood compares to Beta(10,10) posterior.

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.

```
a <- 251527
b <- 241945
probability <- pbeta(0.5, a + 1, 1 + b)
probability
```

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

The posterior is still Beta(a+1, b+1), the probability is 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

The noninformative prior for θ is uniform(0,1), since the uniform(0,1) is flat and implies that we consider the same probability to all possible values of improvement, so uniform(0,1) should be noninformative prior, however, I think the uniform(-1,1) can also be the reasonable noninformative prior, since the average improvement in success probability can also be negative, ie, between -100% to 100%, since the uniform(-1,1) is flat and implies that we consider the same probability to all possible values of improvement, which implies uniform(-1,1) can also be the noninformative prior.

A subjective/informative prior for θ can be beta(3,3), for beta(3,3), we will observed 3 successes and 3 failures and the mean is 0.5 and the variance about 0.036, which means that it assumes that the most likely improvement in success is 50% and it will variate by 0.036, since practicing will increase the improvement in success probability, I think that will be an informative prior.