Week 3: Intro to Bayes and git branches

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.

The mle estimate for $\hat{\theta}$ is 0.9147287, and the 95% CI is (0.8665329 0.9629244).

```
n <- 129
y <- 118
theta.hat <- y / n #binomial
se <- sqrt((theta.hat * (1 - theta.hat)) / n)
lower<- theta.hat - 1.96 * se
```

```r
upper <- theta.hat + 1.96 * se
confint <- c(lower, upper)
theta.hat
```

```
[1] 0.9147287
```

```r
confint
```

```
[1] 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.

The posterior mean for $\hat{\theta}$ is 0.9083969, and the credible interval is (0.8536434, 0.9513891).

```r
#beta(1,1) is equivalent to uniform(0,1)
#y:number of success
#n: total number of trials
a.prior <- 1
b.prior <- 1
a.post <- a.prior + y
b.post <- b.prior + n - y
posterior_mean <- a.post / (a.post + b.post)
credible_interval <- qbeta(c(0.025, 0.975), a.post, b.post)
posterior_mean
```

```
[1] 0.9083969
```

```r
credible_interval
```

```
[1] 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?

The interpretation of this prior is: We think the the unknown parameter θ that underlies our data as the outcome of a random variable whose distribution follows Beta(10,10). We are assuming we know more information compared with the prior used in question 2.

```
a.prior <- 10
b.prior <- 10
a.post <- a.prior + y
b.post <- b.prior + n - y
posterior_mean <- a.post / (a.post + b.post)
credible_interval <- qbeta(c(0.025, 0.975), a.post, b.post)
posterior_mean
```

```
[1] 0.8590604
```

```
credible_interval
```

```
[1] 0.7990363 0.9099708
```

Question 4

Create a graph in ggplot which illustrates

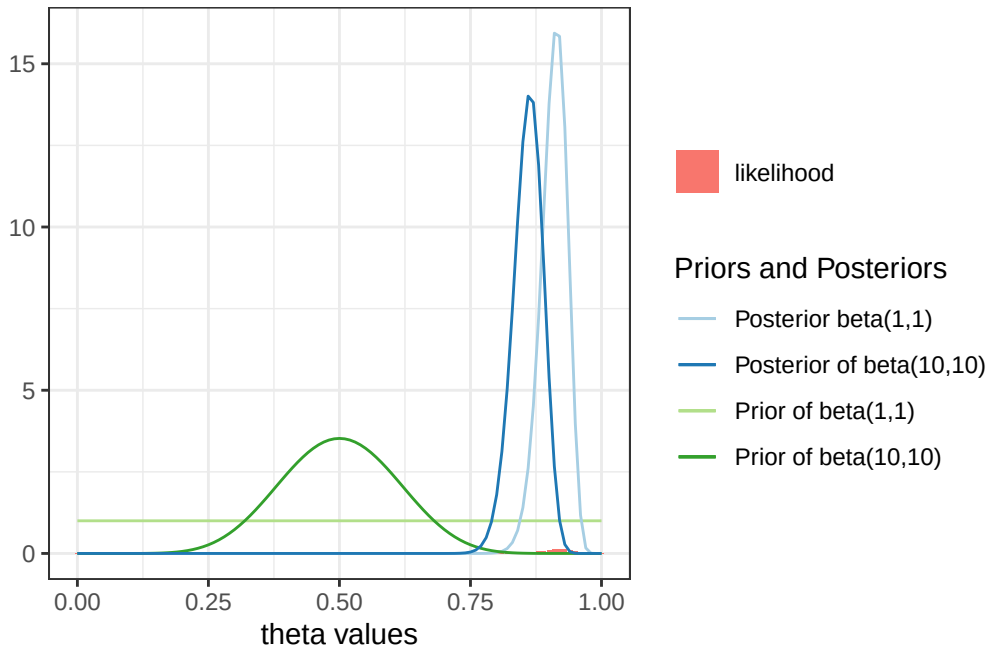
- The likelihood (easiest option is probably to use `geom_histogram` to plot the histogram of appropriate random variables)
- The priors and posteriors in question 2 and 3 (use `stat_function` to plot these distributions)

The figure below strenthen our statement in question 3. The mean of the two priors are the same, but there are some differences for our priors. In question 2, the prior beta(1,1) is equivalent to uniform(0,1), indicating all values between 0 to 1 are equal likely. The beta(10,10) is more concentrated at the mean 0.5, indicating we are have stronger belief that the mean is around 0.5.

As aresult, the posterior of beta(10,10) has a mean that is more close to 0.5 compared with the posterior of beta(1,1).

```r
#histogram
set.seed(123)
library(ggplot2)
n <- 129
y <- 118
n2 <- 100
k <- seq(0, n2)

data <- data.frame(x = seq(0,1,0.01),k = k, y = dbinom(k, n2, y / n))
ggplot(data) +
geom_bar(aes(x = x, y = y,fill = "likelihood"), stat = "identity")+
  stat_function(fun = dbeta, args = list(shape1 = 1, shape2 = 1), aes(color = "Prior of be
  stat_function(fun = dbeta, args = list(shape1 = y+1, shape2 = n-y+1), aes(color = "Poste
    stat_function(fun = dbeta, args = list(shape1 = 10, shape2 = 10), aes(color = "Prior o
  stat_function(fun = dbeta, args = list(shape1 = y+10, shape2 = 10+n-y), aes(color = "Pos
  xlab("theta values")+
  ylab("") +
  scale_color_brewer(palette = "Paired",name = "Priors and Posteriors")+
  scale_fill_discrete(name = "") + theme_bw()
```



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.

Based on the codes below, the probability that a male bith is less than 0.5 is 1.146058e-42, which is approximately 0.

```
m <- 251527
f <- 241945
probability <- pbeta(0.5, m + 1, 1 + f)
probability
```

```
[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

A noninformative prior could be uniform(-1,1), a distribution that indicates the average improvement in success probability are neural between -1 and 1. Without useful information, the performance after practice could be decreasing or increasing.

An informative prior could be beta(5,5). This is because from my prior experence, I believe constant practices contributes to the improvement of success probability. There are also some scientific papers that could side prove this.