Week 3: Intro to Bayes and git branches

12/03/24

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.

```
Loading required package: ggplot2
```

```
-- 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 lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
```

```
-- 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 becom

Attaching package: 'janitor'


The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

Here we know that $Y|\theta \sim Bin(n, \theta)$. Now given $Y_1 = 118$ with $n = 129$:

It then follows that $L(\theta; Y)$ is:

$$\begin{aligned} L(\theta; Y) &= \prod_{i=1}^{1} f(y_i; \theta) \\ &= \binom{n}{y_1} \theta^{y_1} (1-\theta)^{n-y_1} \end{aligned}$$

The log likelihood function then follows:

$$\begin{aligned} l(\theta; Y) &= log(L(\theta; Y)) \\ &= log\binom{n}{y_1} + y_1 log(\theta) + (n - y_1) log(1-\theta) \end{aligned}$$

MLE: take derivatives:

$$\begin{aligned} \frac{d}{d\theta} l(\theta; Y) &= 0 + \frac{y_1}{\theta} - \frac{n-y_1}{1-\theta} \\ &= \frac{y_1}{\theta} - \frac{n-y_1}{1-\theta} \end{aligned}$$

It then follows:

$$\frac{d}{d\theta} l(\hat{\theta}; Y) = 0$$

$$\frac{y_1}{\hat{\theta}} = \frac{n-y_1}{1-\hat{\theta}}$$

$$y_1 - y_1\hat{\theta} = n\hat{\theta} - y_1\hat{\theta}$$

Thus, it follows that

$$\hat{\theta} = \frac{y_1}{n}$$

95% interval:

$$\begin{aligned} I(\theta) &= -\frac{d^2}{d\theta^2} l(\theta; Y) \\ &= -\frac{d}{d\theta}\left(\frac{y_1}{\theta} - \frac{n - y_1}{1-\theta}\right) \\ &= \frac{y_1}{\theta^2} + \frac{n - y_1}{(1-\theta)^2} \end{aligned}$$

Then for the MLE $\hat{\theta} = \frac{y_1}{n}$, it follows that:

$$\begin{aligned} I(\hat{\theta}) &= \frac{y_1}{\hat{\theta}^2} + \frac{n - y_1}{(1-\hat{\theta})^2} \\ &= \frac{n^2}{y_1} + \frac{n^2}{n - y_1} \\ &= \frac{n}{\hat{\theta}(1-\hat{\theta})} \end{aligned}$$

Thus we have the se of this MLE estimate:

$$se(\hat{\theta}) = \sqrt{I(\hat{\theta})^{-1}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

```r
n <- 129
y_1 <- 118
(theta_hat <- y_1/n)
```

```
[1] 0.9147287
```

```r
se <- sqrt(theta_hat*(1-theta_hat)/n)
(ci <- c(theta_hat - 1.96*se, theta_hat + 1.96*se))
```

```
[1] 0.8665329 0.9629244
```

Thus, MLE estimate $\hat{\theta} = 118/129 = 0.9147287$ 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.

In class we learnt that Beta(1,1) is basically uniform (0,1). Thus,

$$\theta \sim U(0,1)$$

$$p(\theta) = 1$$

uniform prior, binomial model:

$$\theta|y \sim Beta(y+1, n-y+1)$$

```
y <- 118
n <- 129
posterior_mean <- (y+1)/(n+2)
posterior_mean
```

```
[1] 0.9083969
```

$$E(\theta|y) = \frac{y+1}{y+1+n-y+1} = \frac{y+1}{n+2} = \frac{119}{131} = 0.908$$

95% credible interval:

```
I <- c(qbeta(0.025,(y+1),(n-y+1)),qbeta(0.975,(y+1),(n-y+1)))
I
```

```
[1] 0.8536434 0.9513891
```

$$Credible Interval 95\% = (0.854, 0.951)$$

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?

What this means is that before we obtain the data, we already assumed that there are 10 successes and 10 failures results, it means we know 9 more success and 9 more failure compared to Beta(1,1).

Now:

$$\theta|y \sim Beta(y+10, n-y+10)$$

```r
y <- 118
n <- 129
posterior_mean <- (y+10)/(n+20)
posterior_mean
```

```
[1] 0.8590604
```

$$E(\theta|y) = \frac{y+10}{y+10+n-y+10} = \frac{y+10}{n+20} = \frac{128}{149} = 0.859$$

95% credible interval:

```r
I <- c(qbeta(0.025,(y+10),(n-y+10)),qbeta(0.975,(y+10),(n-y+10)))
I
```

```
[1] 0.7990363 0.9099708
```

$$CredibleInterval95\% = (0.799, 0.910)$$

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.

```
graph <- ggplot() +
  xlim(0, 1) +
  labs(title="Comparative Analysis: Likelihood, Prior, and Posterior Distributions",
       y="Probability Density", x="Theta (\u03B8)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

graph <- graph +
  geom_function(fun = dbinom, args = list(x = 118, size = 129),
                aes(color = "Likelihood"), linetype = "solid", size = 1.8)
```

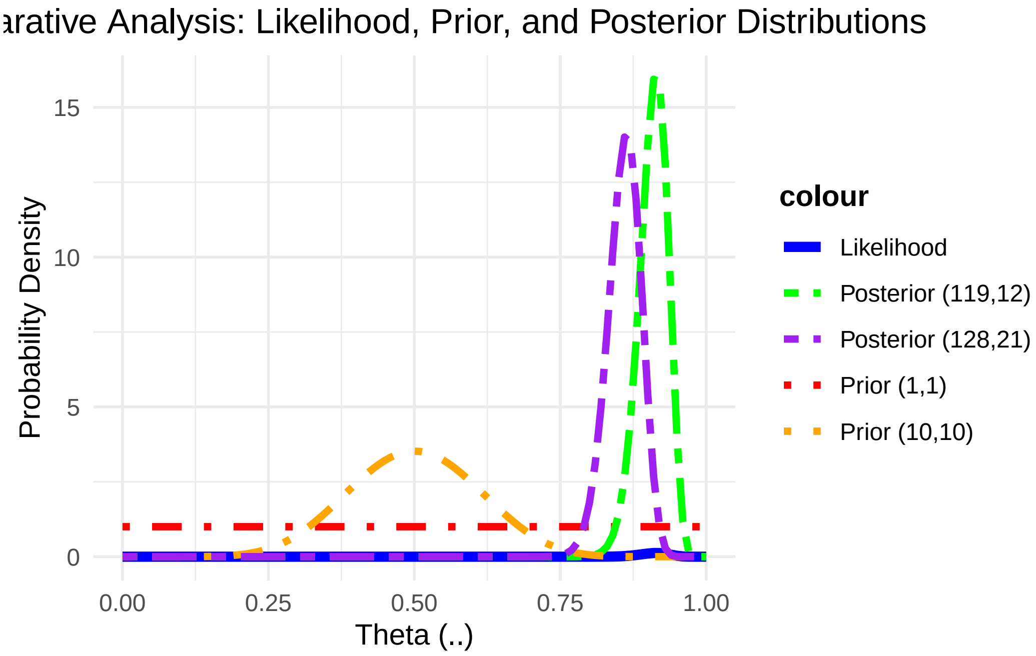
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
graph <- graph +
  geom_function(fun = dbeta, args = list(shape1 = 1, shape2 = 1),
                aes(color = "Prior (1,1)"), linetype = "dotdash", size = 1.3) +
  geom_function(fun = dbeta, args = list(shape1 = 119, shape2 = 12),
                aes(color = "Posterior (119,12)"), linetype = "twodash", size = 1.3)

graph <- graph +
  geom_function(fun = dbeta, args = list(shape1 = 10, shape2 = 10),
                aes(color = "Prior (10,10)"), linetype = "dotdash", size = 1.3) +
  geom_function(fun = dbeta, args = list(shape1 = 128, shape2 = 21),
                aes(color = "Posterior (128,21)"), linetype = "twodash", size = 1.3)

graph <- graph +
  scale_color_manual(values = c("Likelihood" = "blue", "Prior (1,1)" = "red",
                                "Posterior (119,12)" = "green", "Prior (10,10)" = "orange"
                                "Posterior (128,21)" = "purple")) +
  theme(legend.title = element_text(face = "bold"),
        legend.position = "right")

graph
```

```
Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
conversion failure on 'Theta ()' in 'mbcsToSbcs': dot substituted for <ce>
```

```
Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
conversion failure on 'Theta ()' in 'mbcsToSbcs': dot substituted for <b8>
```

```
Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
conversion failure on 'Theta ()' in 'mbcsToSbcs': dot substituted for <ce>

Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
conversion failure on 'Theta ()' in 'mbcsToSbcs': dot substituted for <b8>
```



This plot just confirmed our understanding on an informative prior's influence on a posterior distribution. As we change from a Beta(1, 1) prior to a Beta(10, 10) prior, by incorporating more information, we see that the posterior distribution shifted to closer to 0.5 which is the mean of the prior distribution of Beta(10, 10). The mean of posterior also moves closer to 0.5 compared to 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.

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

```
probability
```

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

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
- A noninformative prior: $Uniform(-\infty, \infty)$

As we learnt in class, this prior basically doesn't give any information on θ.

- A informative prior based on your best knowledge $Beta(5, 5)$

This is informative because we can assume that after one month practice, students are expected to have improvement their skills and thus we limit the value of θ to be non-negative. Beta(5,5), centered around 0.5, means that we believe that there is a moderate chance for improvement in free-throw success after practices, 5 successes and 5 failures.