Week 3: Intro to Bayes

28/01/23

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.

$$
\begin{aligned}
L(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\
l(x|\theta) &= xln(\theta) + (n-x)ln(1-\theta) + C \\
l'(x|\theta) &= x/\theta - (n-x)/(1-\theta) \\
0 &= x/\hat{\theta} - (n-x)/(1-\hat{\theta}) \\
\hat{\theta} &= x/n = 118/129 \\
l''(x|\theta) &= -x/\theta^2 - (n-x)/(1-\theta)^2 \\
I(\theta) &= x/\theta^2 + (n-x)/(1-\theta)^2 \\
Var(\hat{\theta}) &\approx 1/I(\theta) = \frac{\theta(1-\theta)}{n} \\
\hat{Var}(\hat{\theta}) &\approx \frac{\hat{\theta}(1-\hat{\theta})}{n} = \frac{\frac{118}{129}(1-\frac{118}{129})}{129} \\
CI_{0.95} &\approx \hat{\theta} \pm 1.96\sqrt{\hat{Var}(\hat{\theta})} \\
&= \frac{118}{129} \pm 1.96\sqrt{\frac{\frac{118}{129}(1-\frac{118}{129})}{129}}
\end{aligned}
$$

Question 2

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

- The conjugate beta prior and posterior for parameter θ in binomial $Bin(n, \theta)$ is prior $Beta(\alpha, \beta)$ and posterior$Beta(x + \alpha, n - x + \beta)$.
- Therefore, the posterior for theta here is: $\pi(\theta|x) = Beta(118 + 1, 129 - 118 + 1) = Beta(119, 12)$

```
mean_q2 = 119/(119+12)
ci_q2 = c(qbeta(0.025,119,12),qbeta(0.975,119,12))

cat("The posterior mean is: 119/131 = ", mean_q2, "\n")
```

```
The posterior mean is: 119/131 =  0.9083969
```

```
cat("The (quantile-based) 95% credible interval is:", ci_q2, "\n")
```

```
The (quantile-based) 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?

- The interpretation of this prior is that: for the distribution of theta, we are at the level of certainty as if we have already observed 10 successes and 10 fails. Therefore we are assuming that we know more information then the case of question 2 (20 observations worth of info here comparing to 2 observations worth of info in question 2).
- (Since this question does not ask for the estimation and credible interval, they are omitted. The calculation should be very similar to question 2.)

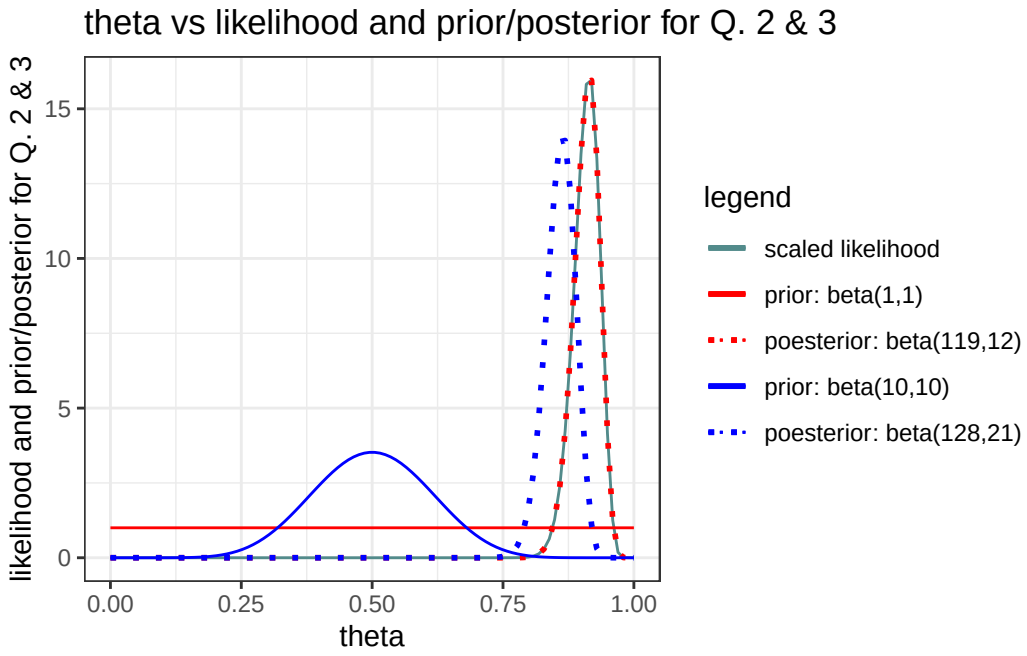
Question 4

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.

```
ggplot() +
  theme_bw() +
  labs(title = "theta vs likelihood and prior/posterior for Q. 2 & 3") +
  xlab("theta") +
  ylab("likelihood and prior/posterior for Q. 2 & 3") +
  stat_function(fun = dbeta, n = 100, args = list(shape1 = 119, shape2 = 12), aes(color =
  stat_function(fun = dbeta, n = 100, args = list(shape1 = 1, shape2 = 1), aes(color = 'pr
  stat_function(fun = dbeta, n = 100, args = list(shape1 = 119, shape2 = 12), aes(color =
  stat_function(fun = dbeta, n = 100, args = list(shape1 = 10, shape2 = 10), aes(color = '
  stat_function(fun = dbeta, n = 100, args = list(shape1 = 128, shape2 = 21), aes(color =
  scale_colour_manual(breaks = c("scaled likelihood", "prior: beta(1,1)", 'poesterior: bet
                      values = c("darkslategray4", "red", "red", "blue", "blue"),
                      name = "legend") +
  scale_linetype_manual(breaks = c("scaled likelihood", "prior: beta(1,1)", 'poesterior: b
                        values = c("solid", "solid", "dotted", "solid", "dotted"),
                        name = "legend")
```



- The q3 prior is stronger than the q2 prior, and caused the q3 posterior to be closer to q3 prior and further away from the likelihood function when comparing to q2. A stronger prior will "pull" the posterior to be more towards it and further away from the likelihood function.

Question 5

(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 here is taken to be Uniform(-1,1), because this prior does nor favor any particular improvement (worsening) value.
- The subjective/informative prior here is taken to be Uniform(0,1). To be honest, I don't know much about basketball, I did some research but was not able to find any study about basketball accuracy around 50 practice shots per day. However, I think the Uniform(0,1) makes sense as practice should improve the accuracy. I decide to leave the support of the distribution as [0,1] and not narrow it down further because I don't want to give a biased prior due to the lack of knowledge.