Week 3: Intro to Bayes

31/01/23

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 $\widehat{\theta}$ and 95% confidence interval.

```
Loading required package: lattice

Loading required package: survival

Loading required package: Formula

Loading required package: ggplot2


Attaching package: 'Hmisc'

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

    format.pval, units

-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
v purrr   1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x dplyr::src()       masks Hmisc::src()
x dplyr::summarize() masks Hmisc::summarize()
```

```r
print(binconf(x=118, n=129,alpha=.05))
```

```
 PointEst    Lower     Upper
0.9147287 0.853752 0.9517195
```

```r
#code reference https://www.geeksforgeeks.org/how-to-calculate-a-binomial-confidence-inter
```

MLE estimate $\hat{\theta} = 0.9147287$

The 95% confidence interval is [0.8538, 0.9517].

Question 2

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

The idea is you find the posterior distribution for theta given a Beta(1,1) prior and then use the mean of the posterior as an estimate for theta. Given it's the same as the example in class you don't have to do the derivation; it's enough just to do the calculations in R.

```r
y= 118
n=129
a=1
b=1
mean_theta = (a+y)/(a+b+n)

mean_theta # estimate of theta
```

```
[1] 0.9083969
```

```r
print(qbeta ( c(0.025, 0.975), a+y, b+n-y))  # gives the 95% 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?

```r
#for beta(10,10)
y= 118
n=129
a=10
b=10

exp_theta = (a)/(a+b)
print(exp_theta)
```

```
[1] 0.5
```

```r
var_theta = (a*b)/((a+b+1)*(a+b)*(a+b))
print(var_theta)
```

```
[1] 0.01190476
```

```r
#for beta(1,1)
a=1
b=1
mean_theta = (a)/(a+b)
print(mean_theta)
```

```
[1] 0.5
```

```r
var_theta = (a*b)/((a+b+1)*(a+b)*(a+b))
print(var_theta)
```

```
[1] 0.08333333
```

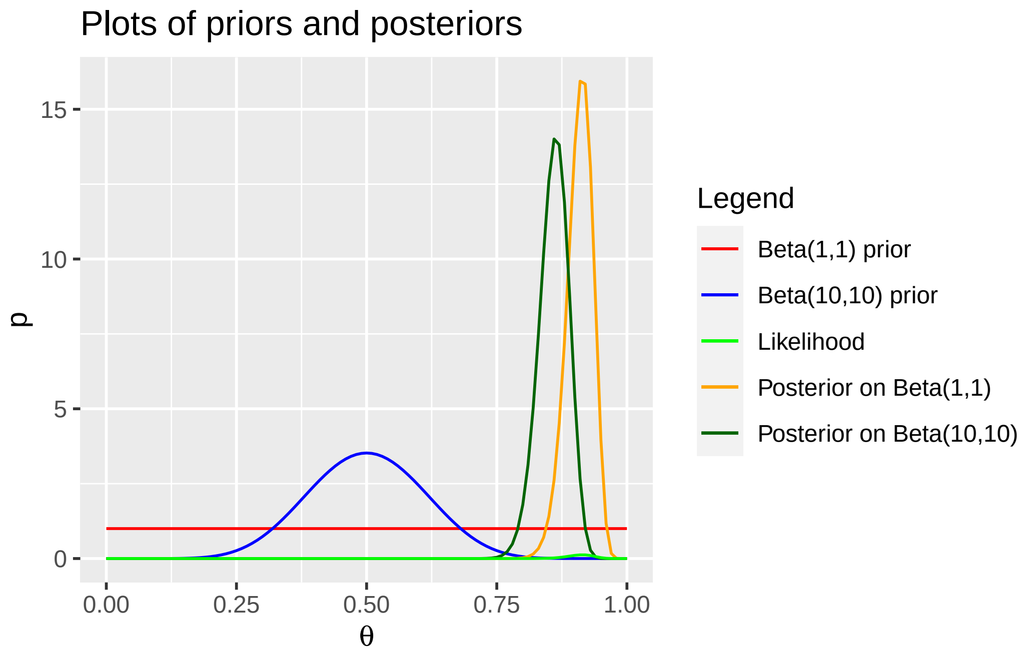
We note that the expected value of θ does not change, but the variance does. As the variance of beta(10,10) is lesser than beta(1,1), it means that the beta(10,10) curve is more peaked than beta(1,1), so it contains lesser uncertainty about our prior distribution. Hence, **we have more information through a Beta(10,10) prior.**

Question 4

Create a graph in ggplot which illustrates - The likelihood - The priors and posteriors in question 2 and 3 (use `stat_function` to plot these distributions)

```r
ggplot(data.frame(x = c(0, 1)), aes(x = x)) +
    stat_function(fun = dbeta, aes(colour = "Beta(1,1) prior"), args = list(shape1 = 1, sh
    stat_function(fun = dbeta, aes(colour = "Beta(10,10) prior"), args = list(shape1 = 10,
    stat_function(fun = dbeta, aes(colour = "Posterior on Beta(1,1)"), args = list(shape1
    stat_function(fun = dbeta, aes(colour = "Posterior on Beta(10,10)"), args = list(shape
  scale_colour_manual("Legend", values = c("red", "blue", "green", "orange", "darkgreen"))

    labs(x = expression(theta), y = 'p', title = "Plots of priors and posteriors")
```



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** A case of non-informative prior would be of uniform distribution in the theoretical range of $[-1,1]$, implying that every level of increase is equally likely.
- **A subjective/informative prior based on your best knowledge** We can model it as a beta distribution based on information about improvement in success probabilities using some past training programs.