Week 3: Intro to Bayes

29/01/23

```
library(ggplot2)
library(tidyverse)
```

```
-- 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()
```

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.

Assuming y$|\theta \sim$ Bin(n, θ), where y is y is the number of women who report to be happy out of the sample of n women. Then the MLE estimate $\hat{\theta}$ = y/n = 118/129 = 0.9147287

```
theta = 118/129
theta
```

```
[1] 0.9147287
```

```
sd = sqrt(theta*(1-theta)/129)
sd
```

```
[1] 0.02458967
```

95% Confidence interval:

```
interval = c(-1.96*sd, 1.96*sd)
interval
```

```
[1] -0.04819576  0.04819576
```

```
theta + interval
```

```
[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.

posterior follows Beta(y+1, n-y+1)

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

```
y = 118
n = 129
bay_theta =  (y+1)/(n+2)
bay_theta
```

```
[1] 0.9083969
```

95% credible interval:

```
qbeta(c(0.025, 0.975), shape1 = y+1, shape2 =  n-y+1)
```

```
[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?

I interpret the beta(10,10) prior as we are assuming 10 out of 20 women report they are happy. Additionally, the probability of women report to be happy is distributed around 0.5 since the expected value of the prior is 10/10+10= 0.5. In this case, we are assuming we know more amount of information than the non-informative prior in Q2.

Question 4

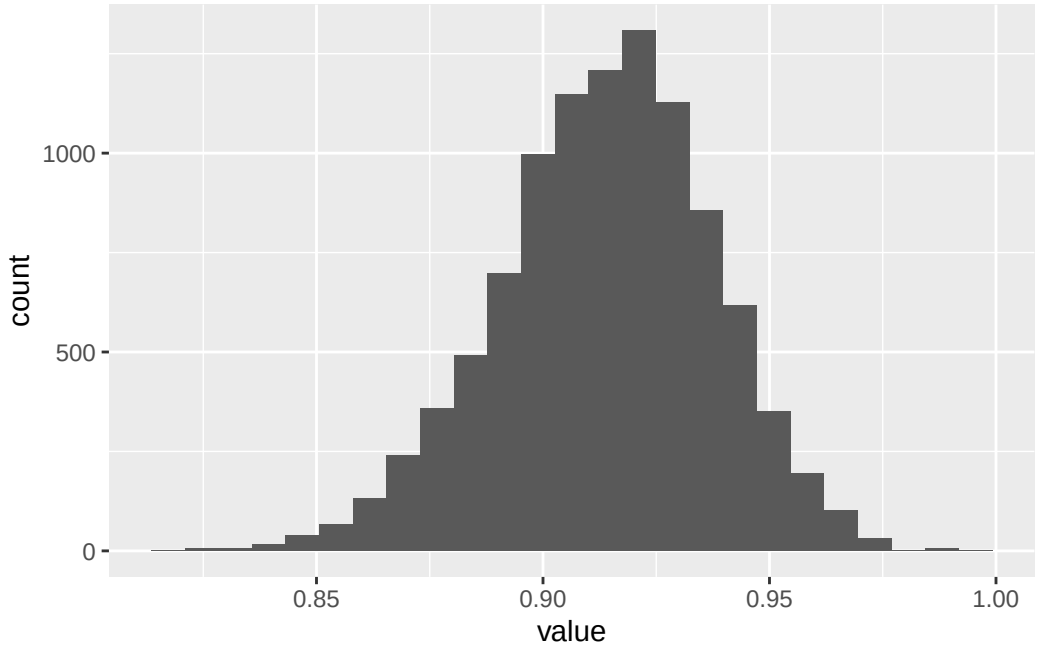
Create a graph in ggplot which illustrates

- 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.

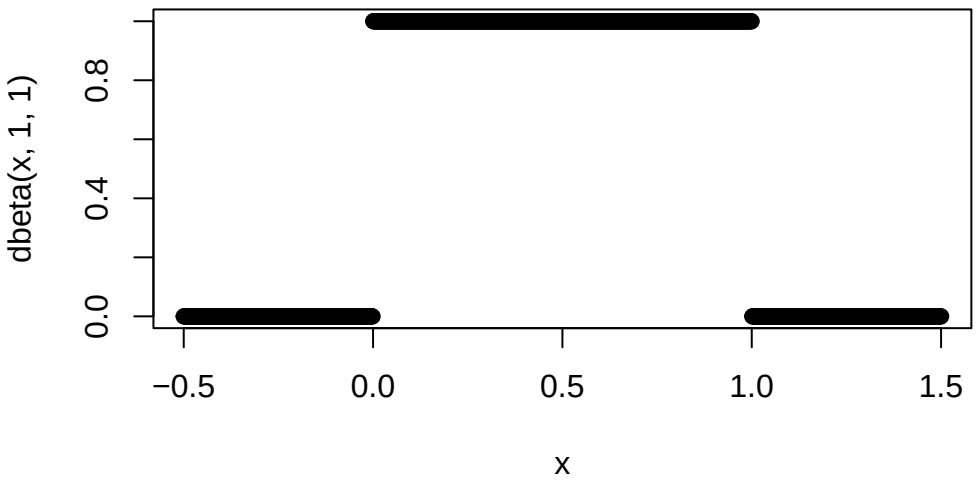
Likelihood

```
data <- rbinom(n= 10000, size = 129, prob = 118/129)
data <- data.frame (value = data/129)
data %>% ggplot(aes(x = value)) + geom_histogram(bins = 25)
```



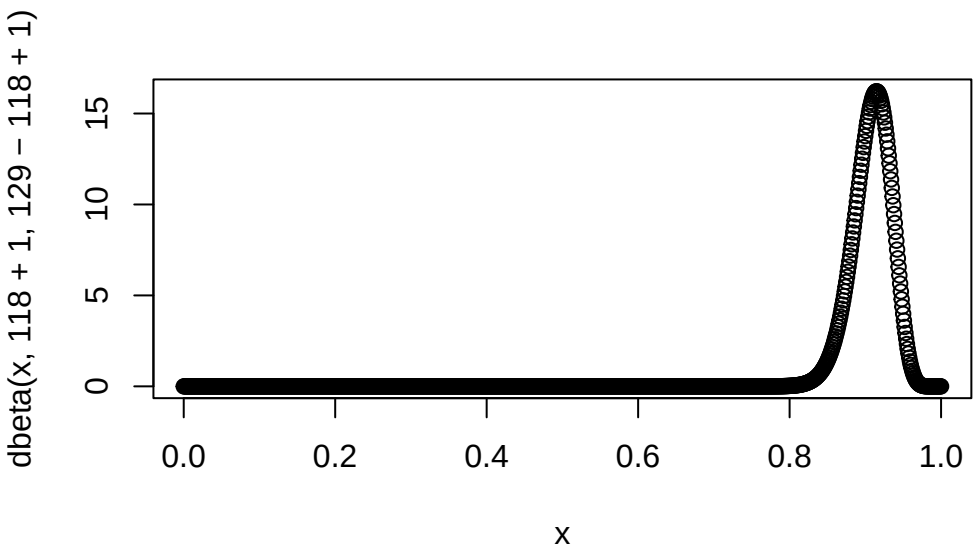
Beta(1,1) prior

```
x<-seq(from=-0.5,to=+1.5,length.out=1000)
plot(x,dbeta(x, 1, 1))
```



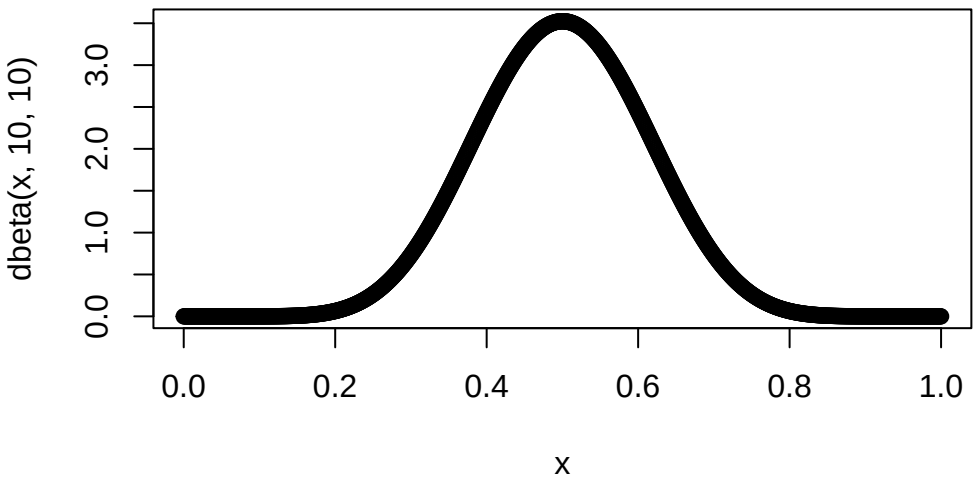
Posterior of Beta(1,1) prior

```
x<-seq(from=-0,to=+1,length.out=1000)
plot(x,dbeta(x, 118+1, 129-118+1))
```



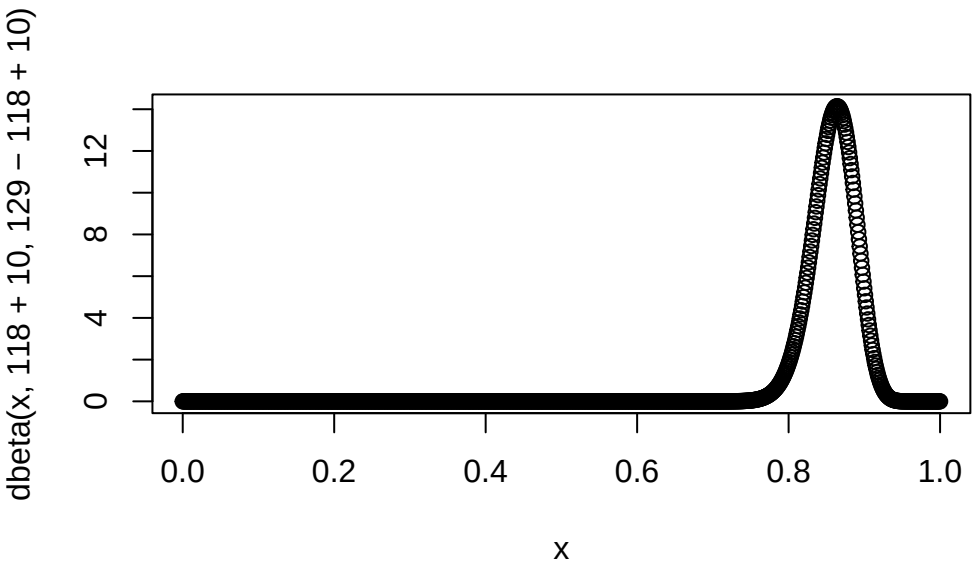
Beta(10,10) prior

```
x<-seq(from=-0,to=+1,length.out=1000)
plot(x,dbeta(x, 10, 10))
```



Posterior of Beta(10,10) prior

```
x<-seq(from=-0,to=+1,length.out=1000)
plot(x,dbeta(x, 118+10, 129-118+10))
```



It looks like different prior does have some impact on the posterior. With non-informative prior beta(1,1), our posterior is very similar to the likelihood. However, with beta(10,10) prior, the posterior is slightly shifted to the left side (towards to our prior distribution).

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

beta(1,1), This is a non-informative prior, we are assuming the improvement is uniformly distributed from 0 to 1

- A subjective/informative prior based on your best knowledge

beta(2,8). We are assuming the average improvement follows beta(2,8), which has mean of 0.2. In other words, we are assuming the improvement 20 more success in the final measurement than in the initial measurement.