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.

```
n = 129
y = 118
theta_mle = y/n
theta_mle
```

```
## [1] 0.9147287
```

```
z = qnorm(0.05/2, lower.tail = FALSE)
se = sqrt( theta_mle*(1-theta_mle)/129 )
CI = c(theta_mle - z*se, theta_mle + z*se)
CI
```

```
## [1] 0.8665338 0.9629236
```

Question 2

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

Note that $\text{Beta}(1,1) \overset{d}{=} \text{Uniform}(0,1)$, and so from Lecture, we have that

$$\theta|Y \sim \text{Beta}(y+1, n-y+1)$$

It then follows that the posterior mean and 95% credible interval for θ are

```
# Posterior mean
post_mean = (y+1)/(n+2)
post_mean
```

```
## [1] 0.9083969
```

```
# Credible interval
CI = qbeta(c(0.05/2, 1-0.05/2), y+1, n-y+1)
CI
```

```
## [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 expected value of θ under each of the distributions remains the same, but we gain more information about the number of observations with the Beta(10, 10) prior, namely that there are 10 women observed to be happy in a cohort of 20, rather than of two women observed to be happy, so we are assuming more information with the new prior.

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)

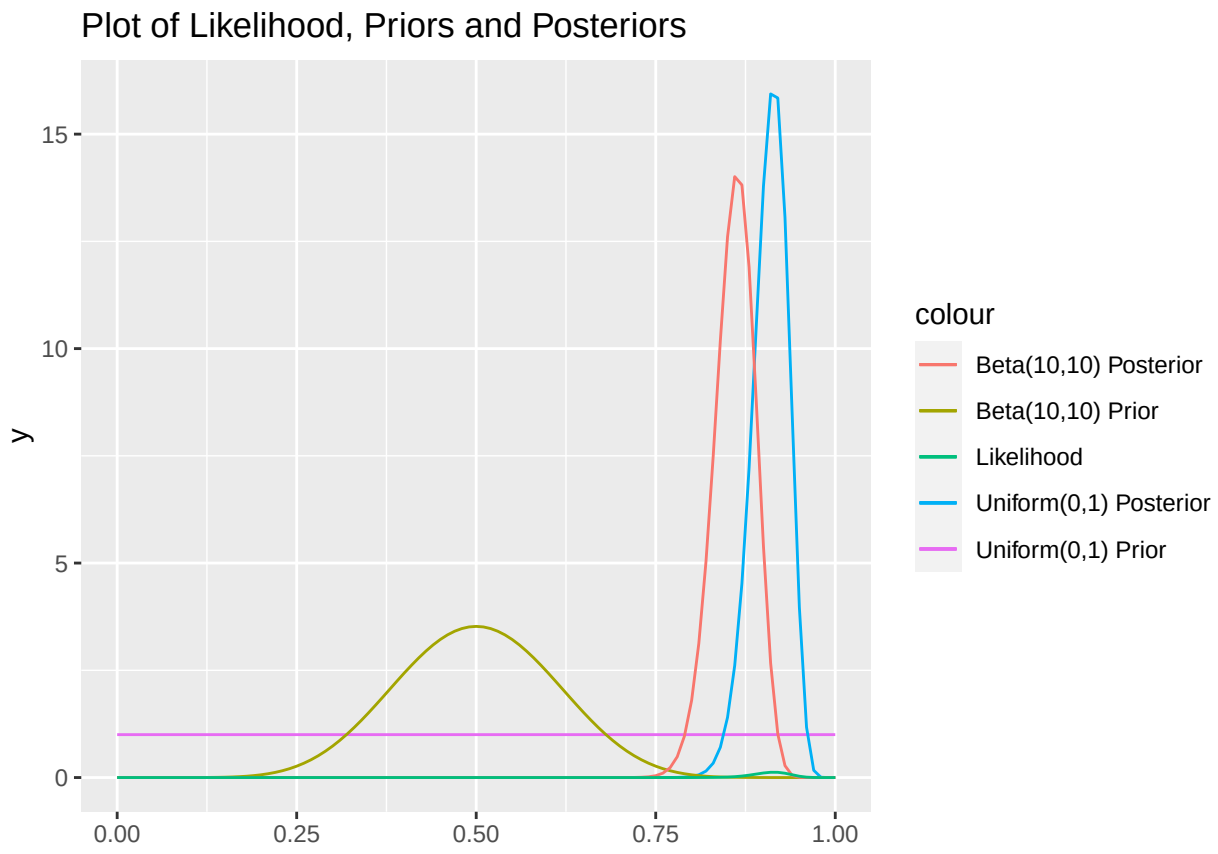
Comment on what you observe.

From Math Stats II, If $Y|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$, then the posterior distribution is $\text{Beta}(y + \alpha, n + y - \beta)$. In particular, the posterior distribution for the $\text{Beta}(10, 10)$ prior in Question 3 is $\text{Beta}(y + 10, n + y - 10)$, with $y = 118$ and $n = 129$. Plotting each together gives

```
my_data = data.frame(theta=seq(0,1,by=0.01))

my_fun = function(theta){
  choose(n,y)*(theta^y)*(1-theta)^(n-y)
}

ggplot(data=my_data) +
  stat_function(fun=dbeta, args=c(1,1),
                aes(color="Uniform(0,1) Prior")) +
  stat_function(fun=dbeta, args=c(y+1,n-y+1),
                aes(color="Uniform(0,1) Posterior")) +
  stat_function(fun=dbeta, args=c(10,10),
                aes(color="Beta(10,10) Prior")) +
  stat_function(fun=dbeta, args=c(y+10,n-y+10),
                aes(color="Beta(10,10) Posterior")) +
  stat_function(fun=my_fun, aes(color="Likelihood")) +
  labs(title="Plot of Likelihood, Priors and Posteriors")
```



In the plot, the peak of the Uniform$(0, 1)$ posterior is closer to the peak of the likelihood than the Beta$(10, 10)$ posterior, and so the posterior expected value is closer to the MLE. Comparing the posteriors directly, the Beta$(10, 10)$ prior is further left of the Uniform$(0, 1)$ prior, meaning that the true proportion of happy females is lower under the Beta$(10, 10)$ prior.

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.

Give two prior distributions for θ (explaining each in a sentence):

- A noninformative prior:

Take $\theta \sim$ Uniform$(-1, 1)$. Note that θ needs to be allowed to be negative in case a final score is worse than the initial score. Use a flat prior since we are assuming no prior information.

- A subjective/informative prior based on your best knowledge:

It's reasonable to assume that the month of practice will increase free throw shooting, so we should choose a distribution that has a positive mean, and contains negative values in its support, leaving only the Uniform and Normal distributions (among common distributions). So take $\theta \sim N(0.25, 0.1)$ so that the average person improves by 25% with a 10% margin of error