9 Bayesian analysis I: Solutions to Practical

Intended learning objectives: The aims of this practical are for you to: * Consolidate understanding of the likelihood; * Use Bayes' theorem to combine the prior distribution and likelihood to obtain the posterior distribution; * Plot prior and posterior distributions in R and compare them; * Compute credible intervals in R; * Find the posterior predictive distribution; * Interpret results from a Bayesian analysis of proportions, and comment on the effect of the prior.

Question 1

Suppose that we collect data y which has been generated according to a distribution with parameter θ. For example, θ could be the population prevalence of disease and y could be the number of people with the disease among a sample of size n.

1. Write down the definition of the likelihood in words.
2. Write down the definition of a prior probability of the disease prevalence in the population from a Bayesian perspective.
3. Write down the definition of a (prior) probability of the disease prevalence in the population from a frequentist perspective.

ANSWERS

1. For discrete parameters, the likelihood is the probability of the (observed) data, conditional on the value of the parameter. For continuous parameters the likelihood is the probability density of the (observed) data, conditional on the value of the parameter.
2. From a Bayesian perspective it is permissible to express our beliefs about different values of the unknown prevalence using probability distributions. The prior distribution represents our current beliefs about the values that the population prevalence takes.
3. From a frequentist perspective the population prevalence is fixed and so we cannot make statements about the probabilities of different values that the prevalence might take. Frequentists are concerned with the estimated prevalence from a sample and how likely that is for different hypothesized values of the population prevalence.

Question 2

Recall the patient who consulted her GP with concerns that she had cancer. Suppose that the GP thought that one of four outcomes were possible: cancer, food poisoning, ulcer or infection. Based on past experience of similar patients the prevalence of these outcomes are $7\%, 80\%, 8\%$ and 5% respectively.

One option is to request a test of a biomarker. Before the GP does this she reviews to results from 1000 previous patients.

Outcome	cancer	food poisoning	ulcer	infection	Total
Test positive	56	400	4	15	475
Test negative	14	400	76	35	525
Total	70	800	80	50	1000

1. Calculate the sensitivity of the biomarker test for each outcome.
2. Calculate the marginal probability of a positive biomarker test by hand and interpret this in words.
3. Calculate the posterior probability of each outcome conditional on a positive biomarker test. Compare prior and posterior probabilities of each outcome.
4. Repeat the above calculations for a negative test - how does the posterior change in this case?
5. Should the GP subject this patient to the biomarker test?

ANSWERS

1. The sensitivity of the biomarker test $T+$ for outcome O_i is $p(T+ \mid O_i)$. We have that the sensitivities are 56/70=0.8 for cancer, 400/800=0.5 for food poisoning, 4/80=0.05 for ulcer, and 14/50=0.3 for infection.
2. The marginal probability of a positive biomarker test is:

$$p(T+) = \sum_j p(T+ \mid O_i)p(O_i) = 0.8 * 0.07 + 0.8 * 0.5 + 0.05 * 0.08 + 0.3 * 0.05 = 0.475$$

1. The posterior probability of each outcome conditional on a positive biomarker test is calculated as:

$$p(O_i \mid T+) = \frac{p(T+ \mid O_i)p(O_i)}{p(T+)}.$$

The resulting posterior probabilities are: $(0.1179, 0.84219, 0.0085, 0.0316)$. Comparing these to the prior, food poisoning remains the most probable outcome, but the probability of cancer has increased to over 10%.

1. Repeating for a negative test, $p(T-) = 1 - p(T+) = 0.525$.

$$p(O_i \mid T-) = \frac{p(T- \mid O_i)p(O_i)}{p(T-)}.$$

```
results in posterior probabilities: $ (0.0267, 0.7619, 0.1448, 0.0667) $
```

Although food poisoning remains the most likely outcome, the probability of ulcer has increased to just under 15%. Cancer is an unlikely outcome.

1. Because cancer is a serious condition a post-test probability of $> 10\%$ is high enough to warrant referral for further tests. Therefore the test is worthwhile.

Question 3 (Optional)

We will use the following R function to create a leaf plot.

In [1]:

```
leafplot <- function(sensi, speci){

# define possible pre-test probabilities
pretest <- seq(0, 1, 0.01)


#calculate probability of having Covid-19 after a positive test result
pos.test <- sensi*pretest/(sensi*pretest+(1-speci)*(1-pretest))

#calculate probability of having Covid-19 after a negative test result
neg.test <- ((1-sensi)*(pretest))/((1-sensi)*pretest+speci*(1-pretest))


#plot the leaves
plot(pretest, pos.test, type="l", col="darkgreen",
xlab="Pre-test Probability", ylab="Post-test Probability")
points(pretest, neg.test, type="l", col="darkgreen")
abline(a=0, b=1, col="darkgreen")

}
```

1. Use the command below to to draw a leaf plot for a test with 80% sensitivity and 98% specificity. If your prior (pre-test) probability is 0.5 of disease, what is the posterior (post-test) probability of disease if the test was positive? and if the test was negative?

In [ ]:

```
leafplot(sensi=0.8, speci=0.98)
```

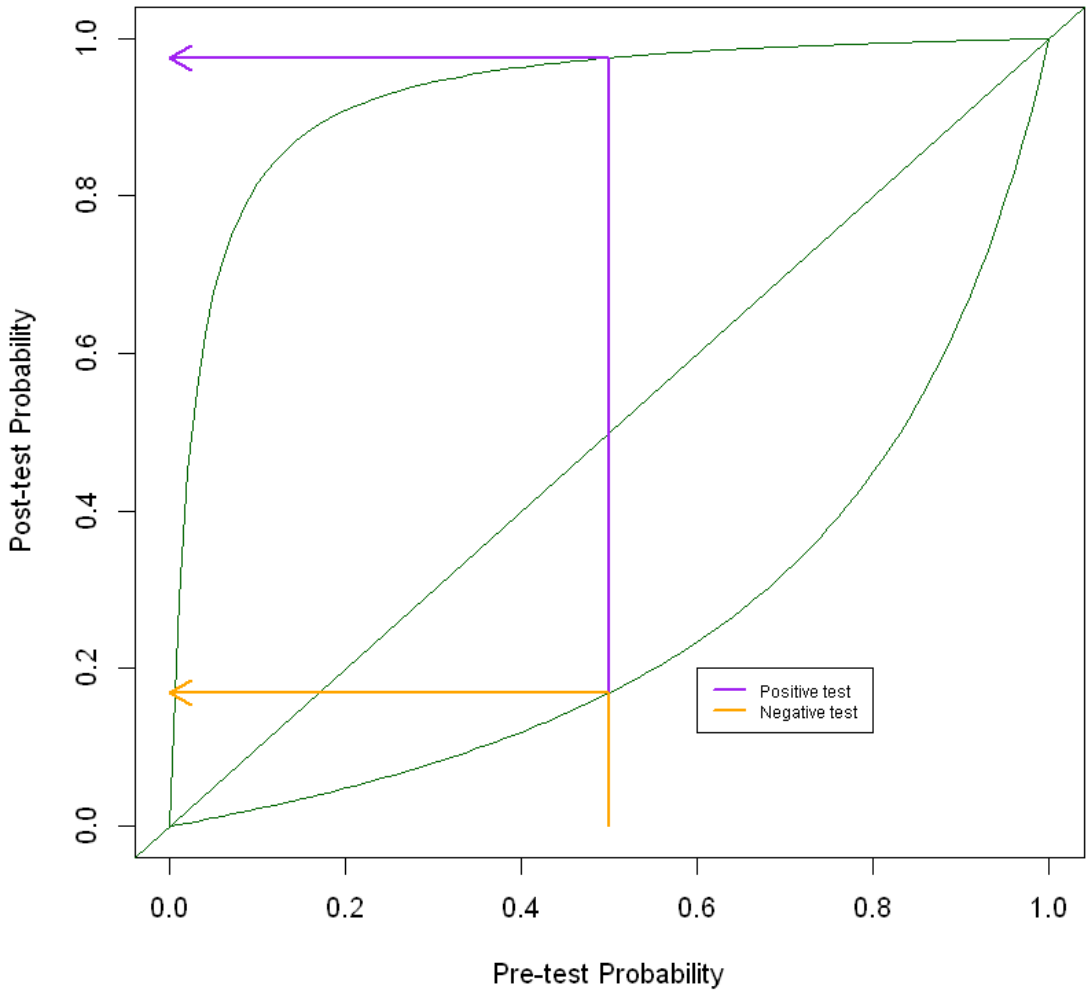
1. Repeat the above for a test with 80% sensitivity and 70% specificity.

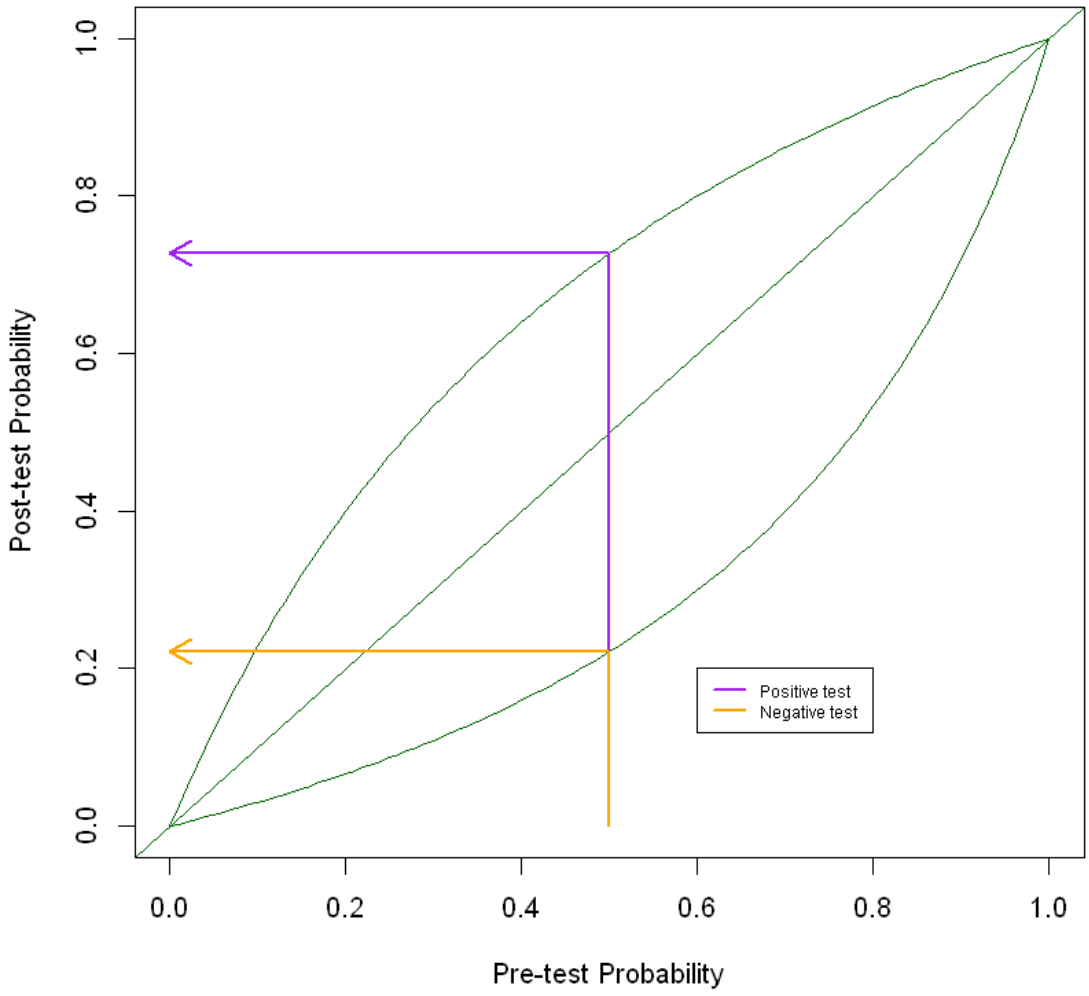
ANSWERS

The function is adapted below to add arrows on the plot to indicate the post-test probability if the pre-test probability is 0.5.

In [3]:

```
leafplot <- function(sensi, speci){

  pretest <- seq(0, 1, 0.01) #possible pre-test probabilities

  #probability of having Covid-19 after a positive test result
  pos.test <- sensi*pretest/(sensi*pretest+(1-speci)*(1-pretest))

  #probability of having Covid-19 after a negative test result
  neg.test <- ((1-sensi)*(pretest))/((1-sensi)*pretest+speci*(1-pretest))

  #plot leaves
  plot(pretest, pos.test, type="l", col="darkgreen",
       xlab="Pre-test Probability", ylab="Post-test Probability")
  points(pretest, neg.test, type="l", col="darkgreen")
  abline(a=0, b=1, col="darkgreen")



  arrows(pretest[51],0,pretest[51],pos.test[51],length=0.0,angle=30,lwd=2,col="purple")
  arrows(pretest[51],pos.test[51],0,pos.test[51],length=0.15,angle=30,lwd=2,col="purpl
e")

  arrows(pretest[51],0,pretest[51],neg.test[51],length=0.0,angle=30,lwd=2,col="orange")
  arrows(pretest[51],neg.test[51],0,neg.test[51],length=0.15,angle=30,lwd=2,col="orang
e")

  legend(0.6,0.2,c("Positive test","Negative test"),lty=1, lwd=c(2:2), col=c("purple",
"orange"), cex=0.6)
}

#For sensitivity 0.8 and specificity 0.98:
leafplot(sensi=0.8, speci=0.98)

#For sensitivity 0.8 and specificity 0.7:
leafplot(0.8, 0.7)
```


Post-test Probability
0.0
0.2
0.4
0.6
0.8
1.0
Pre-test Probability
0.0
0.2
0.4
0.6
0.8
1.0
Positive test
Negative test



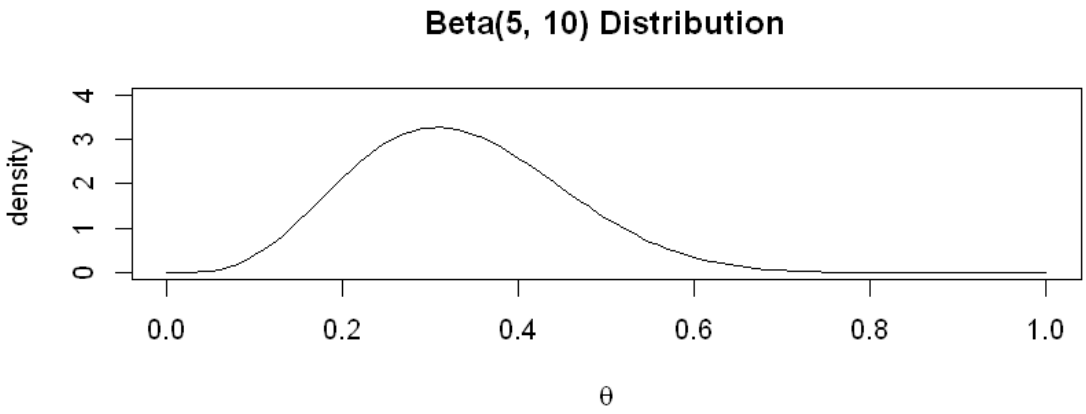
Question 4

For this question, recall the Phase I single-arm trial of a drug for relief of chronic pain. We denote the number of patients trialled by n, and the random variable representing the number of patients who experience pain relief by y. We assume that $y \sim Binomial(n, \theta)$, where θ is the probability of success and we wish to make inference on θ.

1. Suppose we have a Beta$(5, 10)$ prior distribution. We use the following code to plot this prior:

In [1]:

```
options(repr.plot.width=7, repr.plot.height=3)
theta <- seq(0, 1, 0.01)
plot(theta, dbeta(theta, 5, 10), type="l", main="Beta(5, 10) Distribution",
     xlab=expression(theta), ylab="density", ylim=c(0,4))
```



We now observe that 3 patients out of a total of 5 experience pain relief in a trial. Obtain the posterior distribution and edit the code above to add a plot of it to the graph.

1. Find the 95% posterior credible interval for θ, given the data (the credible interval that goes from the 2.5th to the 97.5th percentile). You can use the command $qbeta(p, a, b)$ to get the lower and upper bounds of this interval, where p indicates percentile of the distribution, and a and b are the parameter values of the posterior distribution.

In [ ]:

```
#Lower bound
qbeta(0.025, a, b)

#upper bound
qbeta(0.0975, a, b)
```

1. To obtain the 95% highest posterior density interval for θ given the data, we need to install a package called *HDInterval*. Use the *hdi()* function below to obtain the HPDI. How does this compare to the 95% credible interval for θ above?

In [ ]:

```
install.packages("HDInterval")
library("HDInterval")
#change a and b to the parameters of the posterior distribution that you obtained in pa
rt (1).
hdi(qbeta, 0.95, shape1=a, shape2=b)
```

1. Write down the interpretation of the HPDI. How does this differ from the interpretation of the frequentist confidence interval based on the trial data likelihood?

ANSWERS

1. We saw in the lectures that, if we have a $Beta(a, b)$ prior for θ, and you observe y successes out of n trials, the posterior is a $Beta(a + y, b + n - y)$ distribution. In our case, we have that $a = 5, b = 10, y = 3, n = 5$, so we have a $Beta(8, 12)$ distribution.
2. Using the two lines of code below, we find that our 95% credible interval is $(0.203, 0.616)$.

In [4]:

```
qbeta(0.025, 8, 12)
qbeta(0.975, 8, 12)

# width of CrI
qbeta(0.975, 8, 12) - qbeta(0.025, 8, 12)
```

0.202521438977163

0.616422076685595

0.413900637708432

1. We use the install the *HDInterval* package, load the library and use the given command to find the HPDI to be $(0.197, 0.609)$.

In [ ]:

```
install.packages("HDInterval")
library("HDInterval")
hdi(qbeta, 0.95, shape1=8, shape2=12)
# width of HPDi
0.6091670 - 0.1961515
```

The 95% Credible Interval is slightly wider than the HPDi (0.4139006 vs 0.4130155). The difference is small because the posterior distribution is close to symmetrical.

1. The 95% credible interval is an interval within which theta lies with probability=0.95.The 95% HPDi is the narrowest credible interval for theta. The 95% confidence interval means that, if the study was completed many times, 95\% of the resulting confidence intervals would contain theta.

Question 5

In a different study, there was 1 patient who experienced pain relief out of a total of 5.

1. Use this information for a Beta prior for the probability of success θ: you will need to come up with values for a and b so you have the desired expectation.
   (Note: there is more than one option!).
2. Now suppose that, in a larger study, we observe data are $y = 7$ successes out of $n = 50$. Calculate the posterior, its mean and variance.
3. Below, we have started to create an R function which finds the parameters of the posterior distribution, given the parameters of the prior Beta distribution and an observation from a binomial distribution (the data). This function should take $a.prior, b.prior, y$ and n as arguments. It should print the posterior parameter values. Fill in the parameters of the posterior distribution in lines 5 and 6:

In [9]:

```
binbayes <- function(a.prior, b.prior, y, n){


 #Fill in the parameters of the posterior distribution
 a.posterior <- ##fill this in!!!!!!!
 b.posterior <- ##fill this in!!!!!!!

 #Plot the posterior distribution
 p = seq(0,1, length=100)
 plot(p, dbeta(p, a.posterior, b.posterior), ylab="density", type="l")

 return(c(a.post=a.posterior, b.post=b.posterior))

}
```

1. Derive a Beta prior corresponding to an event probability of 15% observed in 20 patients. Using your function from Question 2.3, calculate the posterior when 3 successes are obtained in 15 patients.
2. Calculate the posterior probability that θ lies between $(0.1, 0.25)$ for Question 2.4 above. Use following command to obtain the cumulative posterior probability up to the lower bound. What is the effect of the prior?

In [ ]:

```
pbeta(0.1, a.post, b.post)
```

ANSWERS

1. The prior study suggests that an initial guess for the probability of success is $\frac{1}{5}$. Since the Beta distribution has mean given by $E(\theta|a,b) = \frac{a}{a+b}$, setting $a = 1, b = 4$ is one option (not the only option!).
2. We have that $a = 1, b = 4, y = 7, n = 50$. The posterior is a $Beta(a + y, b + n - y) = Beta(8, 47)$ distribution.

We calculate the mean and variance:

$$E(\theta|a = 8, b = 47) = \frac{8}{8 + 47} = 0.145$$

$$Var(\theta|a = 8, b = 47) = \frac{8 \times 47}{(8 + 47)^2(8 + 47 + 1)} = 0.002213$$

1. The function is completed below:

In [10]:

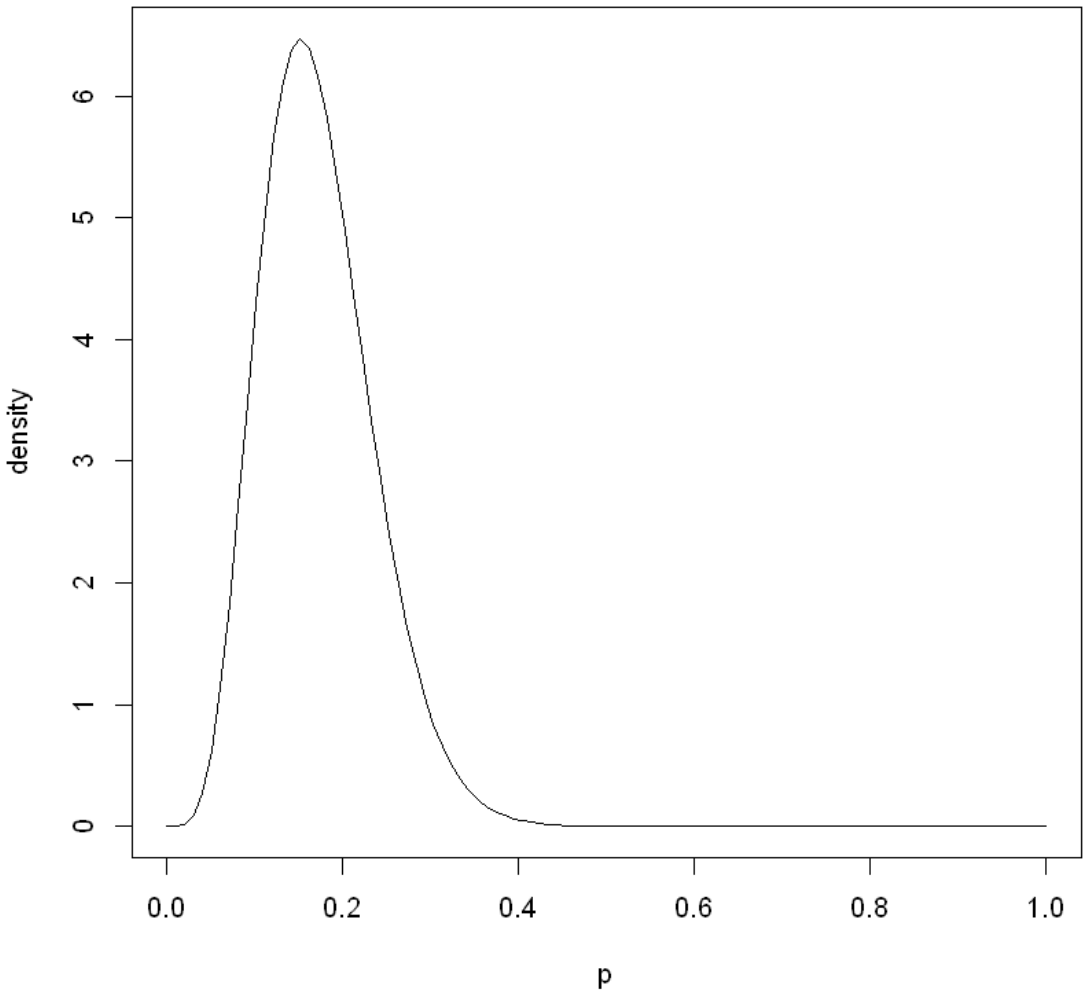
```
binbayes <- function(a.prior, b.prior, y, n){


 #Fill in the parameters of the posterior distribution
 a.posterior <- a.prior + y
 b.posterior <- b.prior + n - y

 #Plot the posterior distribution
 p = seq(0,1, length=100)
 plot(p, dbeta(p, a.posterior, b.posterior), ylab="density", type="l")

 return(c(a.post=a.posterior, b.post=b.posterior))

}
```

1. A $Beta(a = 3, b = 17)$ distribution would be appropriate as it would give us a mean of 0.15. We use the function we previously wrote:

In [11]:

```
binbayes(3, 17, 3, 15)
```

a.post
6
b.post
29



1. Using the code below, we have the probability that θ lies between (0.1, 0.25) is 0.767.

In [12]:

```
l <- pbeta(0.1, 6, 29)
u <- pbeta(0.25, 6, 29)
u-l
```

0.767674584595448

We have that approximately a 75\% probability that θ lies in the interval (0.1, 0.25). Suppose we had a uniform prior, i.e. that $a = b = 1$. Then, the probability that θ lies between (0.1, 0.25) would be 0.5266, as calculated below. The probability that θ lies in the given interval is higher due to our prior.

In [13]:

```
binbayes(1, 1, 3, 15)
pbeta(0.25, 4, 13) - pbeta(0.1, 4, 13)
```

a.post
4
b.post
13

0.526606716024733