Discrete_Priors

October 24, 2017

0.0.1 Bernoulli / Binomial data

Uniform Prior When we use a uniform prior for a Bernoulli likelihood, we get a beta posterior.

$$f(y|\theta) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i},\ f(\theta) = I_{\{0\leq\theta\leq 1\}}$$

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta} = \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}I_{\{0\leq\theta\leq 1\}}}{\int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}I_{\{0\leq\theta\leq 1\}}d\theta}$$

$$= \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}I_{\{0\leq\theta\leq 1\}}}{\frac{\Gamma(\sum(y_i+1))\Gamma(\sum(n-y_i+1))}{\Gamma(n+z)}\int_0^1 \frac{\Gamma(n+z)}{\Gamma(\sum(y_i+1))\Gamma(\sum(n-y_i+1))}\theta^{\sum(y_i}(1-\theta)^{(n-\sum(y_i)}d\theta}$$

$$= \frac{\sum(y_i)}{\Gamma(\sum(y_i+1))\Gamma(\sum(n-y_i+1))}\theta^{\sum(y_i}(1-\theta)^{(n-\sum(y_i)}I_{\{0\leq\theta\leq 1\}}$$

Thus we see the posterior for $\theta|y \sim Beta(\sum y_i + 1, n - \sum y_i + 1)$.

0.0.2 Conjugate priors

In fact, the uniform distribution, is a $Beta(1,1)$. And any beta distribution, is conjugate for the Bernoulli distribution. Any beta prior, will give a beta posterior.

We call the beta prior, Looks like $f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{\{0\leq\theta\leq 1\}}$

We look the posterior for θ:

$$P(\theta|y) \propto f(y|\theta)f(\theta) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{\{0\leq\theta\leq 1\}}$$

$$\propto \theta^{\alpha+\sum y_i-1}(1-\theta)^{\beta+n-\sum y_i-1}I_{\{0\leq\theta\leq 1\}}$$

The posterior distribution for θ (given that the prior is $Beta(\alpha,\beta)$) is $Beta(\alpha+\sum y_i, \beta+n-\sum y_i)$

This whole concept now of starting with the beta prior and getting a beta posterior is a really convenient one.

This whole process where we choose a particular form of prior that works with a likelihood is called using a **conjugate family**.

A family of distributions is referred to as conjugate if when you use a member of that family as a prior, you get another member of that family as your posterior.

The beta distribution is conjugate for the Bernoulli distribution. It's also conjugate for the binomial distribution. The only difference in the binomial likelihood is that there is an $\binom{n}{x}$, since that does not depend on θ then we get the same posterior.

We often use conjugate priors because they make life much more simpler.

As we're working out posteriors, if we can't recognize this form, we get some intractable integral in the denominator. And trying to work out that integral can be problematic. It can get complicated really quickly. So sticking to conjugate families allows us to get closed form solutions.

If the family is flexible enough, then you can find a member of that family that closely enough represents your beliefs. We can represent this model as a **hierarchy**.

$$Y_1 \dots Y_m \sim B(\theta)$$

$$\theta|\alpha, \beta \sim Beta(\alpha, \beta) : prior$$

$$\alpha, \beta = \alpha_o, \beta_o : hyperparameters$$

In a more complicated problem instead of setting values to α and β we want to have more flexibility by putting priors on α and/or β.

0.0.3 Posterior mean and effective sample size

Recall that the expected value of a beta:

$$\frac{\alpha}{\alpha + \beta}$$

For the case of posterior, it is:

$$\frac{\alpha + \sum y_i}{\alpha + \sum y_i + \beta + n - \sum y_i} = \frac{\alpha + \sum y_i}{\alpha + \beta + n}$$

$$= \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{\sum y_i}{n}$$

$$= prior\ weight \times prior\ mean + data\ weight \times data\ mean$$

The weight for the prior, and the weight for the data add up to one, so they are weights. And so the posterior mean is the weighted average of the prior mean and the data mean.

We can also see here, here's the **n** for our sample size, and so the corresponding term over here, tells us the effective sample size of the prior. Thus the affected sample size of the prior for beta prior on Bernoulli or a binomial likelihood is $\alpha + \beta$.

This effective sample size also gives you an idea of how much data you would need to make sure that you're prior doesn't have much influence on your posterior.

If $\alpha + \beta$ is small compared to n. And the posterior will largely just be driven by the data. If $\alpha + \beta$ is large relative to n, then your posterior will be largely driven by the prior.

** Example **

Suppose we are giving two students a multiple-choice exam with 40 questions, where each question has four choices. We don't know how much the students have studied for this exam, but we think that they will do better than just guessing randomly.

1) What are the parameters of interest?

2) What is our likelihood?

3) What prior should we use?

4) What is the prior probability P(theta>.25)? P(theta>.5)? P(theta>.8)?

5) Suppose the first student gets 33 questions right. What is the posterior distribution for theta1? P(theta1>.25)? P(theta1>.5)? P(theta1>.8)?

   What is a 95% posterior credible interval for theta1?

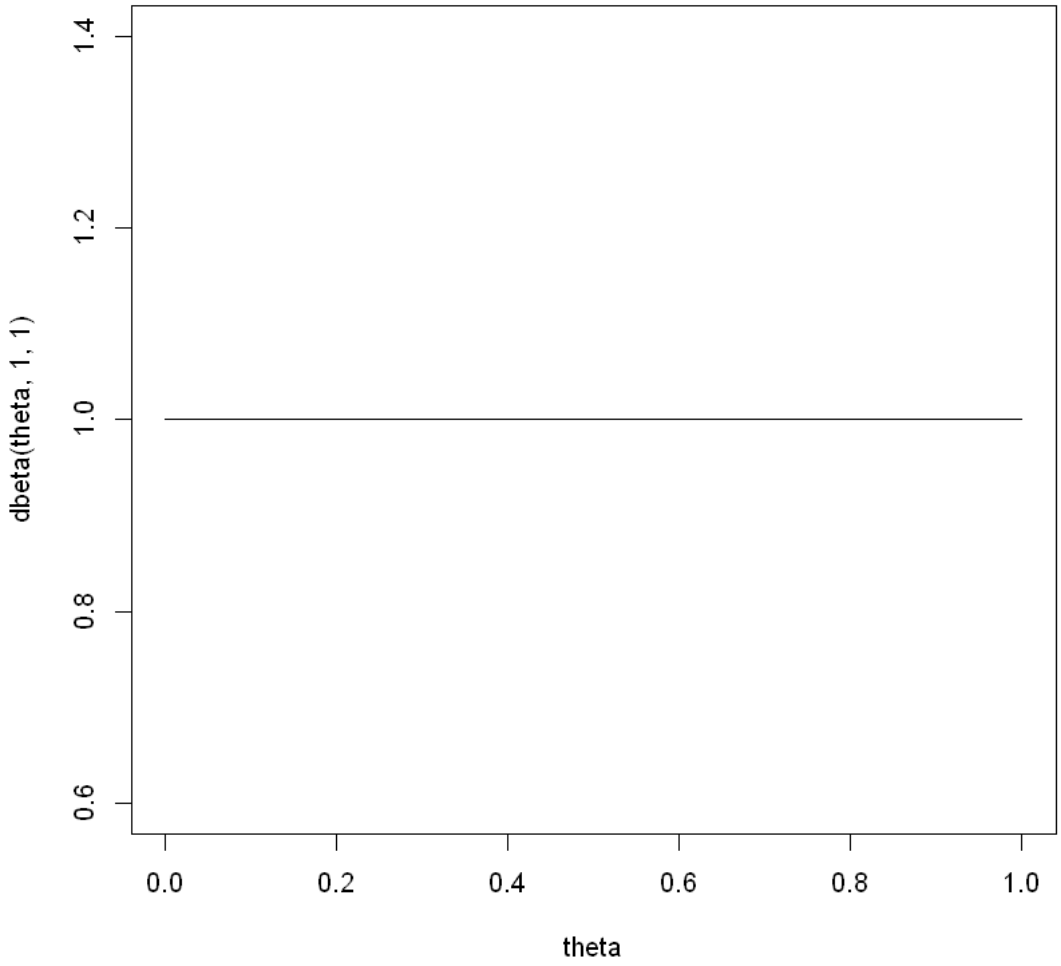
6) Suppose the second student gets 24 questions right. What is the posterior distribution for theta2? P(theta2>.25)? P(theta2>.5)? P(theta2>.8)?

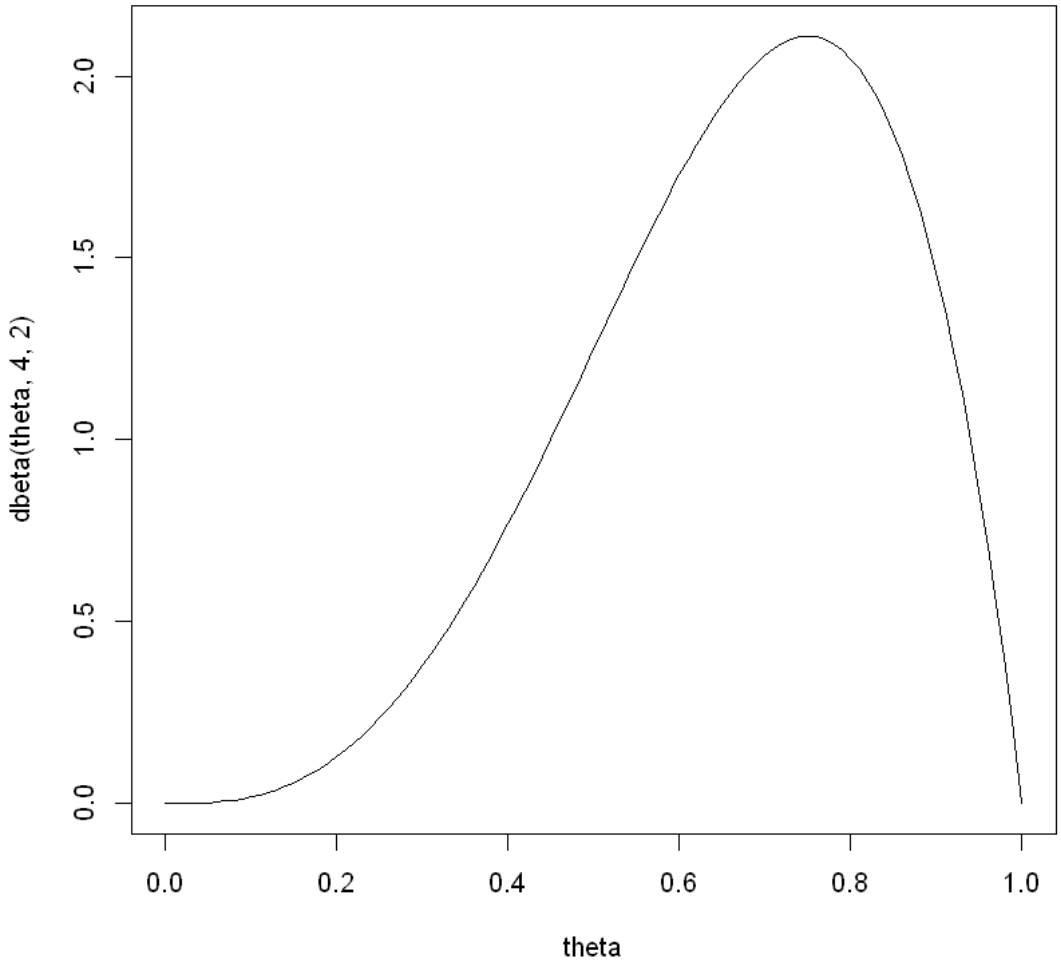
   What is a 95% posterior credible interval for theta2?

7) What is the posterior probability that theta1>theta2, i.e., that the first student has a better chance of getting a question right than the second student?

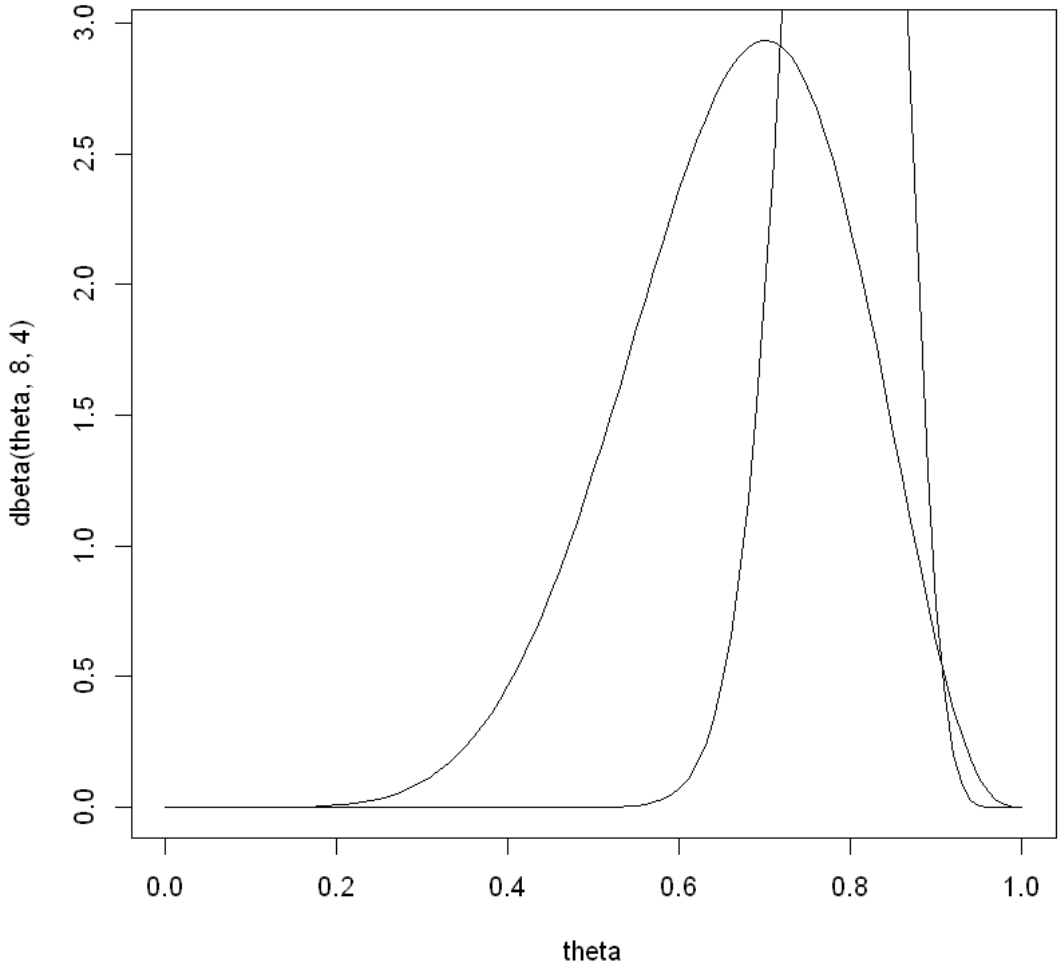
```
In [1]: # Solutions:

        # 1) Parameters of interest are theta1=true probability the first student
        #    will answer a question correctly, and theta2=true probability the seco
        #    student will answer a question correctly.

        # 2) Likelihood is Binomial(40, theta), if we assume that each question is
        #    independent and that the probability a student gets each question righ
        #    is the same for all questions for that student.

        # 3) The conjugate prior is a beta prior. Plot the density with dbeta.
        theta=seq(from=0,to=1,by=.01)
        plot(theta,dbeta(theta,1,1),type="l")
        plot(theta,dbeta(theta,4,2),type="l")
        plot(theta,dbeta(theta,8,4),type="l")

        # 4) Find probabilities using the pbeta function.
        1-pbeta(.25,8,4)
        1-pbeta(.5,8,4)
        1-pbeta(.8,8,4)

        # 5) Posterior is Beta(8+33,4+40-33) = Beta(41,11)
        41/(41+11)  # posterior mean
        33/40       # MLE

        lines(theta,dbeta(theta,41,11))

        # plot posterior first to get the right scale on the y-axis
        plot(theta,dbeta(theta,41,11),type="l")
        lines(theta,dbeta(theta,8,4),lty=2)
        # plot likelihood
        lines(theta,dbinom(33,size=40,p=theta),lty=3)
        # plot scaled likelihood
        lines(theta,44*dbinom(33,size=40,p=theta),lty=3)

        # posterior probabilities
        1-pbeta(.25,41,11)
```

```r
1-pbeta(.5,41,11)
1-pbeta(.8,41,11)

# equal-tailed 95% credible interval
qbeta(.025,41,11)
qbeta(.975,41,11)

# 6) Posterior is Beta(8+24,4+40-24) = Beta(32,20)
32/(32+20)  # posterior mean
24/40       # MLE

plot(theta,dbeta(theta,32,20),type="l")
lines(theta,dbeta(theta,8,4),lty=2)
lines(theta,44*dbinom(24,size=40,p=theta),lty=3)

1-pbeta(.25,32,20)
1-pbeta(.5,32,20)
1-pbeta(.8,32,20)

qbeta(.025,32,20)
qbeta(.975,32,20)

# 7) Estimate by simulation: draw 1,000 samples from each and see how ofte
#    we observe theta1>theta2

theta1=rbeta(1000,41,11)
theta2=rbeta(1000,32,20)
mean(theta1>theta2)


# Note for other distributions:
# dgamma,pgamma,qgamma,rgamma
# dnorm,pnorm,qnorm,rnorm
```

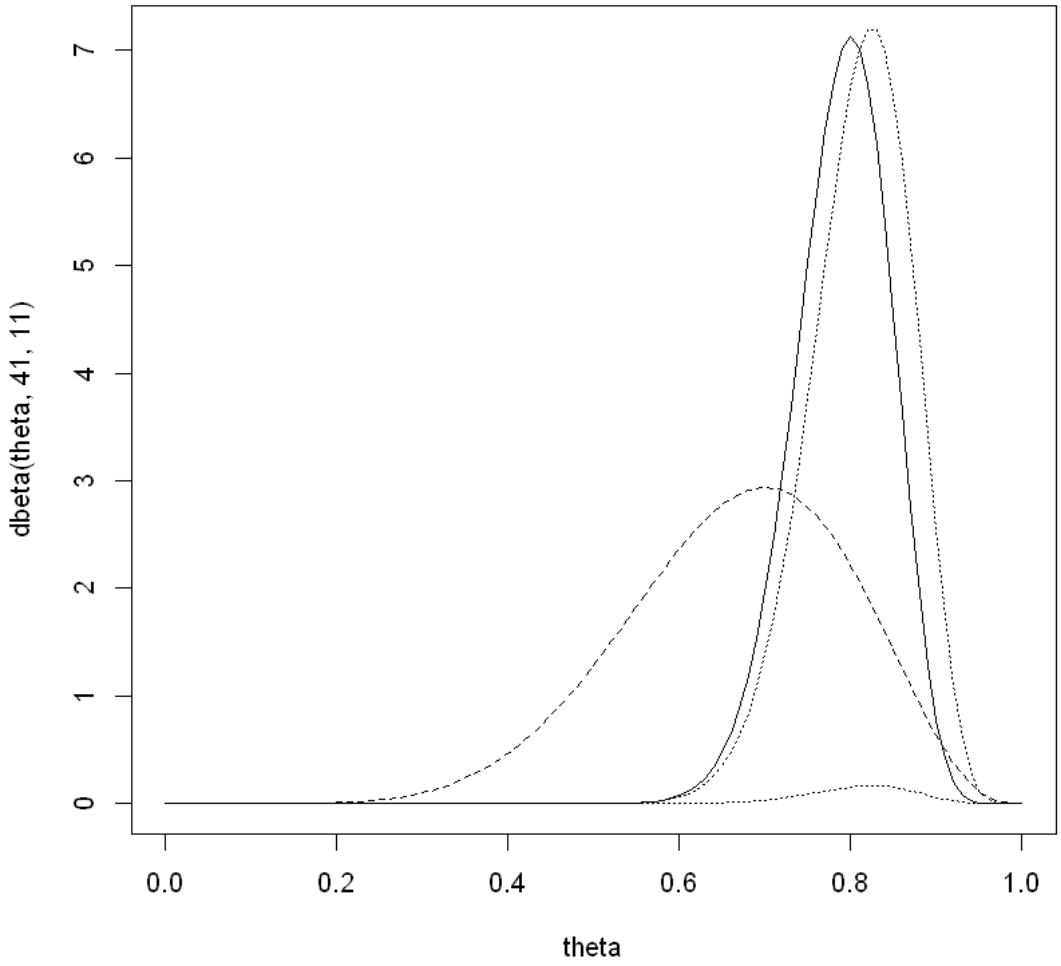

dbeta(theta, 1, 1)
1.4
1.2
1.0
0.8
0.6
0.0
0.2
0.4
0.6
0.8
1.0
theta



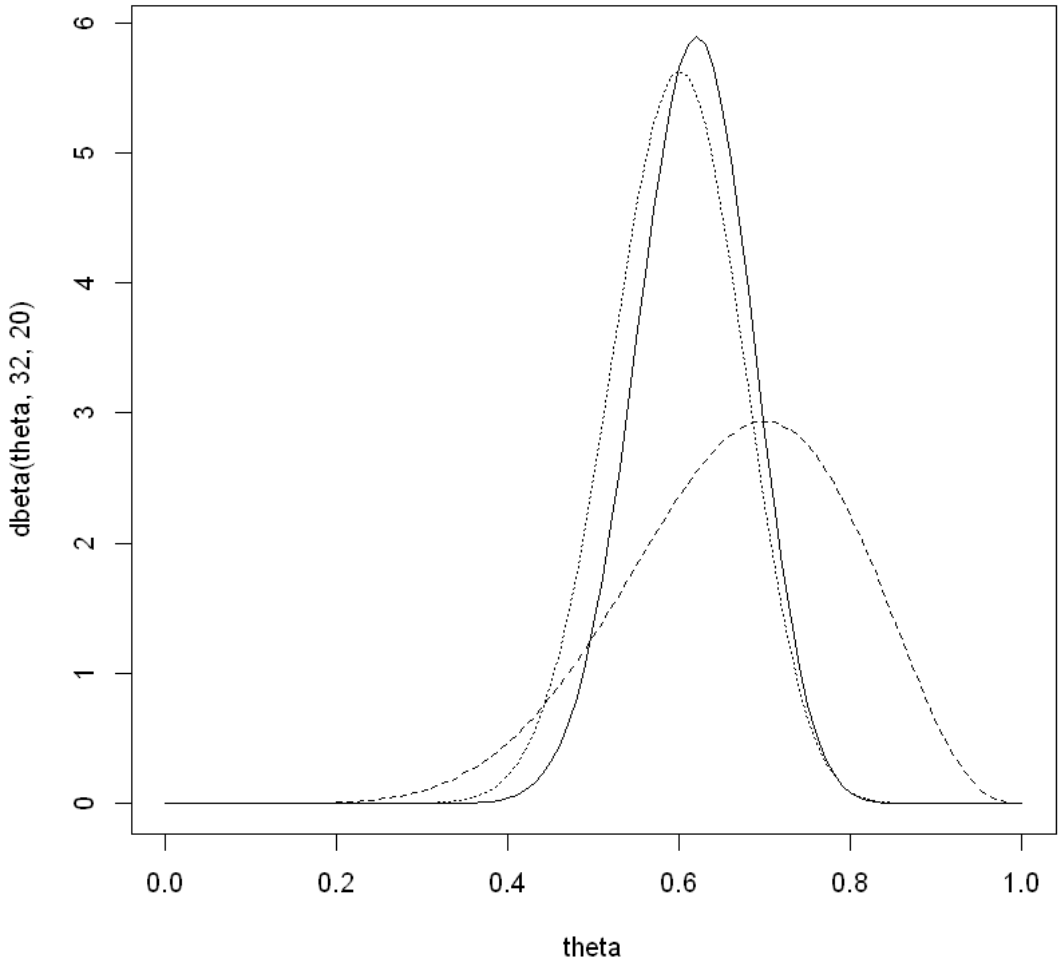
0.998811721801758
0.88671875
0.1611392
0.788461538461538
0.825



1
0.999992631142057
0.444404382568797
0.668842648747071
0.887109400250257
0.615384615384615
0.6



0.999999986313041
0.954042725092073
0.00124818985803987
0.480802158241963
0.741556364717817
0.972



** Question 1 **

Suppose we use a Bernoulli likelihood for each coin flip, i.e., f(yi|θ)=θyi(1$-\theta$)1$-$yiI{0$\leq\theta\leq$1} for yi=0 or yi=1, and a uniform prior for θ.

What is the posterior distribution for θ if we observe the following sequence: (T, T, T, T) where H denotes heads (Y=1) and T denotes tails (Y=0)?

Answer 1 $Beta(\alpha + \sum y_i, \beta + n - \sum y_i)$

Beta(1, 5)

Question 2

What is the maximum likelihood estimate (MLE) of θ if we observe the sequence (T, T, T, T)?

Answer 2

0

** Question 3**

What is the maximum likelihood estimate (MLE) of θ if we observe the sequence (T, T, T, T)?

Answer 3

```
In [2]: alpha <- 1
        beta <- 5
        alpha/(alpha + beta)
```

0.166666666666667

** Question 4**

Find the posterior probability that $\theta < 0.5$ if we observe the sequence (T,T,T,T).

Answer 4

```
In [3]: pbeta(q=0.5, 1, 5)
```

0.96875

An engineer wants to assess the reliability of a new chemical refinement process by measuring θ, the proportion of samples that fail a battery of tests. These tests are expensive, and the budget only allows 20 tests on randomly selected samples. Assuming each test is independent, she assigns a binomial likelihood where X counts the samples which fail. Historically, new processes pass about half of the time, so she assigns a Beta(2,2) prior for θ (prior mean 0.5 and prior sample size 4). The outcome of the tests is 6 fails and 14 passes.

** Question 5 **

What is the posterior distribution for θ?

Answer 5

$Beta \sim (16, 8)$

** Question 6** Calculate the upper end of an equal-tailed 95% credible interval for θ

Answer 6

```
In [4]: alpha <- 8
        beta <- 16
        qbeta(p=0.975, alpha, beta)
```

0.529191666008507

The engineer tells you that the process is considered promising and can proceed to another phase of testing if we are 90% sure that the failure rate is less than .35.

Question 7

Calculate the posterior probability P(θ<.35|x). In your role as the statistician, would you say that this new chemical should pass?

Answer 7

```
In [5]: pbeta(q=0.35, alpha, beta)
```

0.586431031445081

It is discovered that the budget will allow five more samples to be tested. These tests are conducted and none of them fail.

Question 8

Calculate the new posterior probability P(θ<.35|x1,x2). In your role as the statistician, would you say that this new chemical should pass (with the same requirement as in the previous question)?

Hint: You can use the posterior from the previous analysis as the prior for this analysis. Assuming independence of tests, this yields the same posterior as the analysis in which we begin with the Beta(2,2) prior and use all 25 tests as the data.

Answer 8

```
In [6]: pbeta(q=0.35, alpha+0, beta+5)
```

0.817906445569061

0.1 Poisson data

Let's consider Poisson data. For example, think about chocolate chip cookies. In mass produced chocolate chip cookies, they make a large amount of dough. They mix in a large number of chips, mix it up really well and then chunk out individual cookies. In this process the number of chips per cookie approximately falls a Poisson distribution.

If we were to assume that chips have high volume then this would be exactly a Poisson practice and follow exactly a Poisson distribution. But in practice chips are not that big and so they follow approximately a Poisson distribution for number of chips per cookie.

So: $Y_i \sim Pois(\lambda)$,

likelihood shall be :

$$\frac{\lambda^{\sum y_i} e^{-n\lambda}}{\prod_{i=1} y_i}$$

For this we need lambda to be positive valued.

Now our parameter has to be positive valued. What type of prior should we put on lambda? It would be convenient if we could but a conjugate prior. So we can ask, what distribution looks like λ to the something e to the minus something λ?

That distribution is the gamma distribution.

$$\lambda \sim \Gamma(\alpha, \beta)$$

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$f(\lambda|y) \propto f(y|\lambda) f(\lambda) \propto \lambda^{\sum y_i}$$
$$\propto \lambda^{\alpha+\sum y_i} e^{-(\beta+n)\lambda}$$

The posterior is:

$$\Gamma(\alpha + \sum Y_i, \beta + n)$$

The mean of gamma is:

$$\frac{\alpha}{\beta}$$

The posterior mean is:

$$\frac{\alpha + \sum y_i}{\beta + n}$$

$$= \frac{\beta}{\beta+n} \frac{\alpha}{\beta} + \frac{n}{\beta+n} \frac{\sum y_i}{n}$$

How would we chose a particular prior, a particular alpha and beta we need to specify values for?Let's take a look of two strategies. Under the first strategy, we want to include information into our prior, based on our personal knowledge. We can start by thinking about what is the prior mean? Prior mean in this case is:

$$\frac{\alpha}{\beta}$$

.

We can think, what do we think our prior mean is for the number of chips per cookie.

The strategy now involves we have to specify a second thing because there are two parameters. So if we specify two things we can solve for alpha and beta. One possible second thing we could specify is our prior standard deviation or prior variance. We can specify a prior standard deviation and that would be:

$$\frac{\sqrt{\alpha}}{\beta}$$

Another way of specifying our confidence or uncertainty in our prior is to think about the effective sample size.

As we see here, the effective sample size is β. And so we can specify how many units of information, we think we have in our prior. How sure we are in that sense versus

The second strategy shall be to represent ignorance with a vague prior.

In Bayesian statistics, a vague prior refers to one that's relatively flat across much of the space. In this case, we can think about some small ϵ.So ϵ is some small number that is strictly positive. And then we can have:

$$\Gamma(\epsilon, \epsilon)$$

As long as these are both strictly positive, this is a proper prior, it's a proper distribution.

We think about what the posterior mean looks like under this prior. The posterior mean would be:

$$\frac{\epsilon + \sum y_i}{\epsilon + n} \propto \frac{\sum y_i}{n}$$

.

If ϵ is really small then this is approximately the $\frac{\sum y_i}{n}$, just the data mean.

Review

As in the lesson, we use a Poisson likelihood to model the number of chips per cookie, and a conjugate gamma prior on λ, the expected number of chips per cookie. Suppose your prior expectation for λ is 8.

Question 1

The conjugate prior with mean 8 and effective sample size of 2 is Gamma(a,2). Find the value of a.

Answer 1

16

Question 2

The conjugate prior with mean 8 and standard deviation 1 is Gamma(a,8). Find the value of a.

Answer 2

64

Question 3

Suppose you are not very confident in your prior guess of 8, so you want to use a prior effective sample size of 1/100 cookies. Then the conjugate prior is Gamma(a,0.01). Find the value of a. Round your answer to two decimal places.

Answer 3

0.08

Question 4

Suppose you decide on the prior Gamma(8, 1), which has prior mean 8 and effective sample size of one cookie.

We collect data, sampling five cookies and counting the chips in each. We find 9, 12, 10, 15, and 13 chips.

What is the posterior distribution for λ?

```
In [7]: alpha <- 8
        beta <- 1
        cookies <- c(9, 12, 10, 15, 13)
        n <- 5
        posterior_a <- alpha + sum(cookies); posterior_a
        posterior_b <- beta + n; posterior_b
```

67

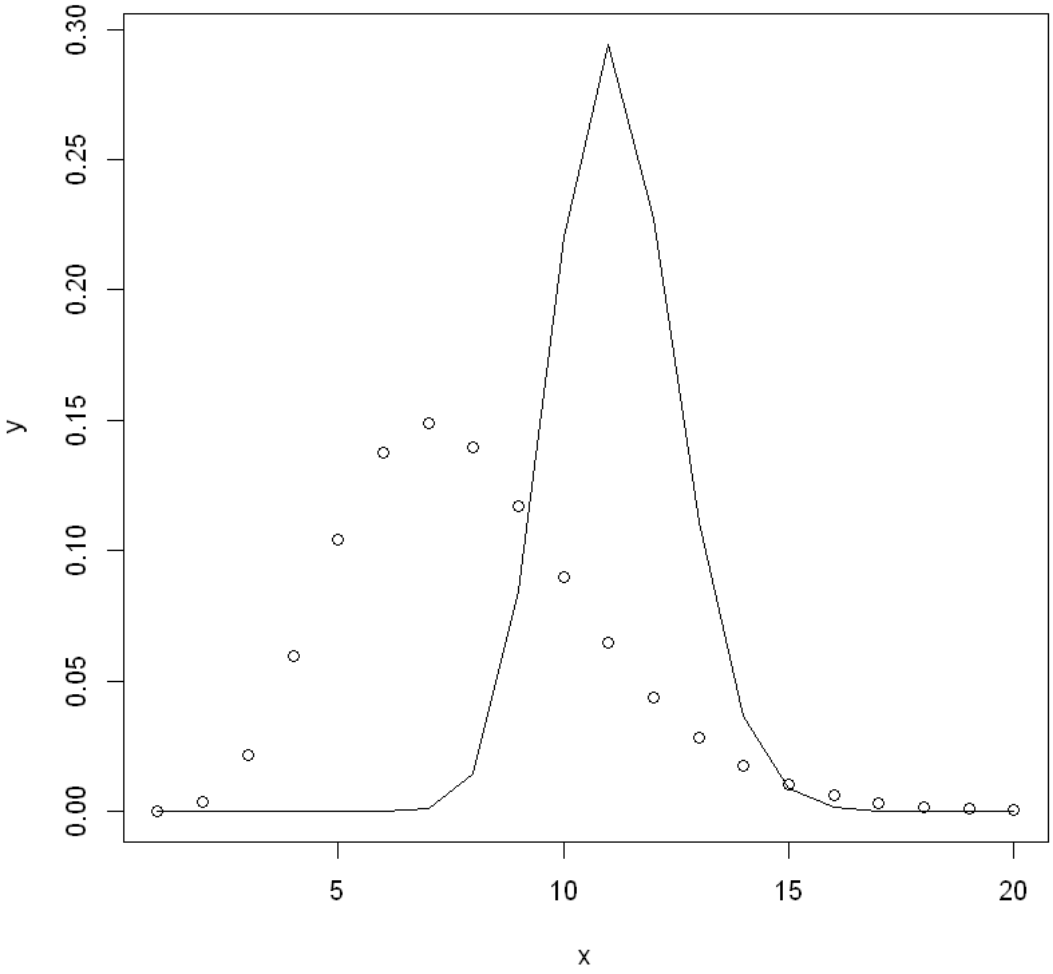
6

Question 5

Continuing the previous question, what of the following graphs shows the prior density (dotted line) and posterior density (solid line) of λ?

Answer 5

```
In [8]: x <- 1:20
        y <- dgamma(x, alpha, beta)
        y_post <- dgamma(x, posterior_a, posterior_b)
        plot(x, y, ylim=c(0,max(y_post)))
        lines(x, y_post)
```



Question 6
What is the posterior mean for λ?
Answer 6

```
In [9]: posterior_mean <- (posterior_a)/(posterior_b); posterior_mean
```

11.1666666666667
Question 7
Find the lower end of a 90% equal-tailed credible interval for λ.
Answer 7

```
In [10]: qgamma(p=0.05, posterior_a, posterior_b)
```

9.02138196638713

** Question 8**

Continuing Question 4, suppose that in addition to the five cookies reported, we observe an additional ten cookies with 109 total chips. What is the new posterior distribution for λ, the expected number of chips per cookie?

Answer 8

```
In [11]: post_a <- posterior_a + 109 ;post_a
         post_b <- posterior_b + 5 ; post_b
```

176

11

A retailer notices that a certain type of customer tends to call their customer service hotline more often than other customers, so they begin keeping track. They decide a Poisson process model is appropriate for counting calls, with calling rate θ calls per customer per day.

The model for the total number of calls is then Y$\sim$Poisson(n$\cdot$t$\cdot\theta$) where n is the number of customers in the group and t is the number of days. That is, if we observe the calls from a group with 24 customers for 5 days, the expected number of calls would be 24$\cdot$5$\cdot\theta$=120$\cdot\theta$.

Question 9

Following the same procedure outlined in the lesson, find the posterior distribution for θ.

Answer 9

$\Gamma(a+y, b+nt)$

Queston 10

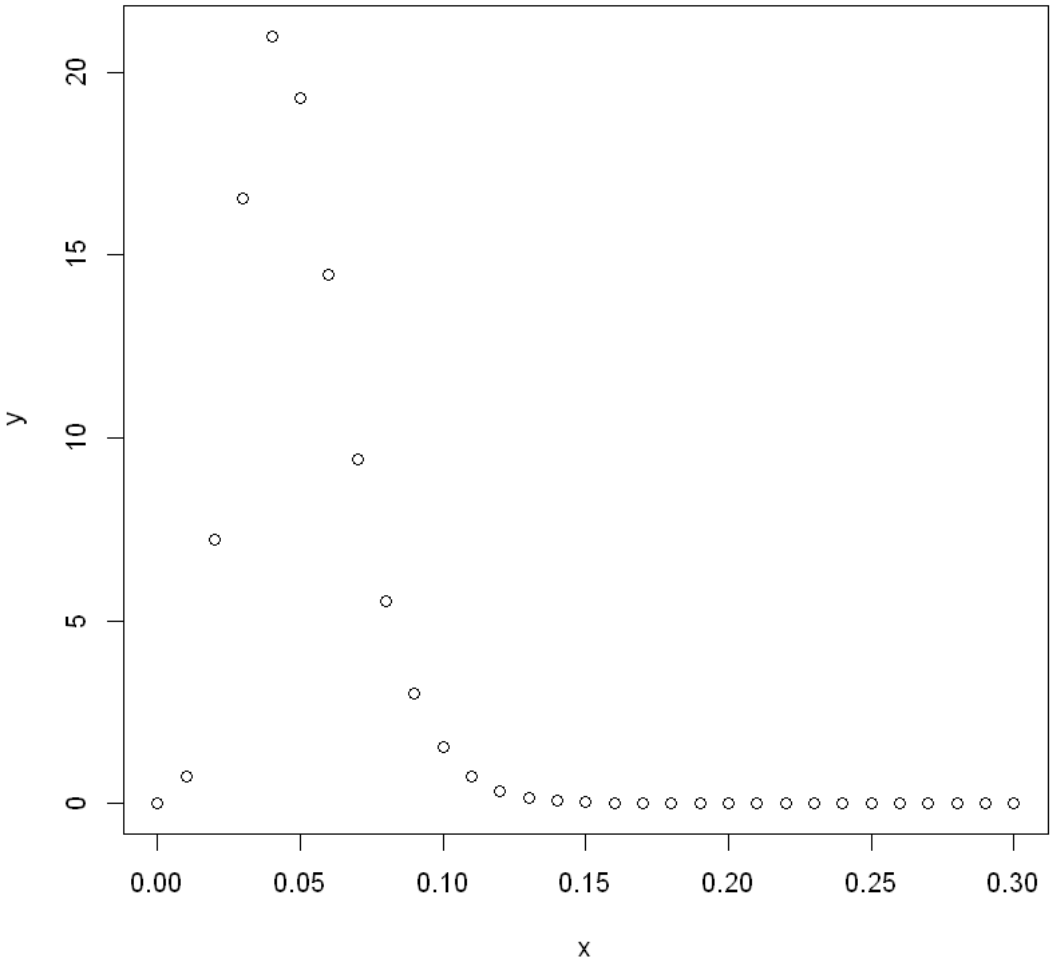
Poisson process:

On average, the retailer receives 0.01 calls per customer per day. To give this group the benefit of the doubt, they set the prior mean for θ at 0.01 with standard deviation 0.5. This yields a $\Gamma(\frac{1}{2500}, \frac{1}{25})$ prior for θ.

Suppose there are $n = 24$ customers in this particular group of interest, and the retailer monitors calls from these customers for $t = 5$ days. They observe a total of $y = 6$ calls from this group.

The following graph shows the resulting $\Gamma(6.0004, 120.04)$ posterior for θ, the calling rate for this group. The vertical dashed line shows the average calling rate of 0.01.

```
In [12]: a <- 6.0004
         b <- 120.04
         x <- seq(0,0.3, 0.01)
         y <- dgamma(x, a, b)
         plot(x, y)
```



Answer 10

Yes, most of the posterior mass (probability) is concentrated on values of θ greater than 0.01.

** Advanced questions **

Question 1

Identify which of the following conditions (possibly more than one) must be true for the sum of n Bernoulli random variables (with success probability p) to follow a binomial distribution.

a) the sum must exceed n

b) the sum must be greater than zero

c) p must be the same for each of the Bernoulli random variables

d) p must be less than .5

e) each Bernoulli random variable is independent of all others

** Answer 1**

c, e

We have found that the prior predictive distribution for a Bernoulli trial under a uniform prior on the success probability θ. We now derive the prior predictive distribution when the prior is any conjugate beta distribution.

There are two straightforward ways to to this. The first approach is the same as before. The marginal distribution of y is $f(y) = \int_0^1 f(y|\theta)f(\theta)d\theta$. Now $f(\theta)$ is a beta PDF, but the same principles apply: we can move constants out of the integral and find a new normalizing constant to make the integral evaluate to 1.

Another approach is to notice that we can write Bayes' theorem as $f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$. If we multiply both sides by $f(y)$ and divide both sides by $f(\theta|y)$, then we get $f(y) = \frac{f(y|\theta)f(\theta)}{f(\theta|y)}$ where $f(\theta)$ is the beta prior PDF and $f(\theta|y)$ is the updated beta posterior PDF.

** Question 2**

Both approaches yield the same answer. What is the prior predictive distribution $f(y)$ for this model when the prior for θ is $Beta(a, b)$?

Answer 2

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a+b+1)}\frac{\Gamma(a+y)}{\Gamma(a)}\frac{\Gamma(b+1-y)}{\Gamma(b)}, for y = 0, 1$$

```
In [13]: form <- function(a, b, y){
             return((factorial(a+b-1)/factorial(a+b))*(factorial(a+y-1)/factorial(a
         }
```

Question 3

Beta-Bernoulli predictive distribution:

Now suppose the prior for θ is Beta(2,2). What is the prior predictive probability that $y* = 1$ for a new observation $y*$? Round your answer to one decimal place.

Answer 3

```
In [14]: a <- 2
         b <- 2
         form(a, b, 1)
```

0.5

All we do, is plugging the a and b to the formula from question 2. The answer to the question is 0.5.

Question 4

Beta-Bernoulli predictive distribution:

After specifying our Beta(2,2) prior for θ, we observe 10 Bernoulli trials, 3 of which are successes.

What is the posterior predictive probability that $y* = 1$ for the next (11th) observation $y*$? Round your answer to two decimal places.

Answer 4

First, we need to calculate the posterior Beta, which is $Beta(a + \sum y_i, b + n - \sum y_i)$ 9) and we will just plug the new a and b to the formula from question 2.

```
In [15]: post_a <- a + 3;post_a
         post_b <- b + 7;post_b
         form(post_a, post_b, 1)
```

5
9
0.357142857142857

```
In [ ]:
```