Bayes: Midterm 1
Leslie Gains-Germain

Note to Megan: I decided to doublespace because it's easier for me to read, and it's easier to read your comments when I get it back. So, I doubled all your maximum lengths.

1. Continuing the discussion...

   Me: "What are your reasons for choosing a non-informative prior?"

   Researcher: "I don't want my prior knowledge to influence the results, I just want to let the data speak."

   Me: "I think a main strength of using Bayesian methods is the ability to incorporate prior knowledge into the analysis. Why don't you want to incorporate prior knowledge? If the reason is because you are not confident in the prior knowledge, you can absolutely incorporate your uncertainty into the distribution you choose for the prior. Additionally, it's a really good exercise to think about the prior you choose and what it means. It's easy to choose a default non-informative prior without really thinking about it. If you truly do not want to reflect any prior knowledge, I wouldn't recommend going directly to the default non-informative prior in your software package. Instead, I would go through the thought process of choosing a vague prior with natural constraints. Do you have any knowledge about the lower and upper limits of the results you expect to see?

   Researcher: "Well, we are trying to estimate the mean age of Montana residents, so I know that humans can't be more than 150 years old."

   Me: "Great! You can reflect these natural constraints in your prior. If you have absolutely no information about the true mean age of Montanans, you can use a $Unif(0, 150)$ prior. Incorporating that bit of basic knowledge into the prior can help the software package compute the posterior distribution. Some standard non-informative priors are improper probability distributions (or close to being improper), and this can result in an improper posterior distribution. We can talk more about what this means, but the overall message is that the results are not valid if the posterior distribution is improper, and it can make more work for you in the end. Overall, I recommend using an informative prior distribution that appropriately represents your prior knowledge and prior uncertainty, but if you insist on using a non-informative prior, I would choose one that is reasonable based on the natural constraints in your study.

2. As discussed in the previous problem, I think one of the reasons why people choose non-informative priors is because they don't want any prior knowledge to influence results. Instead, they want the data to drive the results. I will refer to this type of non-informative prior as an "uninformative" prior. If the choice of "uninformative" prior matters because so few data are available, then it is truly impossible to find a prior that does not influence the results. If this is the case, the researchers are better off acknowledging that the "uniformative prior" does not exist. As a result, whatever prior they choose will be informative (because it will inform the results). Essentially, they are forced to choose an informative prior, so they might as well incorporate as much knowledge as they have into that informative prior.

3. Say you are given a bucket of 12 marbles, and the probability of drawing a black marble from the bucket is θ. We are told that we can draw a sample of 6 marbles in the bucket, and we can use what we observe to estimate θ. After observing the sample of marbles, we can consider the likelihood function. This function of θ gives us a "number" for each possible value of θ. These "numbers" or "likelihoods" are completely meaningless on their own. The likelihoods only have meaning when they are compared across values of θ. The value of θ with the largest likelihood is known as the Maximum Likelihood Estimate (MLE) and is generally used as the best guess for θ. For example, if 2 black marbles are observed out of 6 draws, then we could calculate a likelihood for each possible value of $\theta \in (0, 1/12, 2/12, 3/12, ...11/12, 1)$. In this example, 4/12 would have the largest likelihood because 1/3 of the sample draws were black marbles. This means that 4/12 is the MLE for θ. Values near 4/12 will have likelihoods that are slightly less than the likelihood for 4/12, and values far from 4/12, such as 11/12 and 1 will have likelihoods that are much smaller than the likelihood for 4/12. Again, I'll say that the value of the likelihood at 4/12 is NOT important. What is important is the value of the likelihood at 4/12 compared to the values of the likelihood at all other values of θ.

   This should help explain why either the log of the likelihood function or the likelihood function itself can be used to find the MLE. Because $log(x)$ is an increasing function, it turns out that the value of θ that has the largest likelihood also has the largest log(likelihood) when compared to the log(likelihoods) for the other values of θ. Overall, the relative relationships in the log(likelihood) function are the same as the relative relationships in the likelihood function.

A posterior distribution is not the same as a likelihood function. First, it is a probability distribution. This means that the values in the posterior distribution are meaningful all on their own, rather than just in a relative sense. For example, suppose the posterior probabilities for $\theta \in (0, 1/12, 2/12, 3/12, ...11/12, 1)$ are $(0, 0.03, 0.16, 0.23, 0.25, 0.17, 0.10, 0.06, 0, 0, 0, 0, 0, 0)$. This set of probabilities represents the discrete posterior probability distribution. 0.25 can be interpreted as the probability that θ is 4/12 given the observed data. Also notice that the posterior probabilities sum to one, which is a property of a probability distribution. Second, the posterior probability distribution incorporates prior knowledge specified by the researcher. The only information going into the likelihood function, however, comes from the observed data. The MLE is often close to the posterior mean because there is generally more information about θ in the likelihood function than there is in the prior distribution for θ that is specified by the researcher. As a result, inference about θ from the likelihood function is often similar to inference about θ from the posterior probability distribution.

4. If $\theta \sim Pareto(\alpha, y_m)$ and we observe n observations from $y|\theta \sim Unif(0, \theta)$, then the posterior distribution of θ is $Pareto(\alpha + n, y_m)$. My work is shown below.

$$p(\theta|\mathbf{y}) = \left(\frac{1}{\theta}\right)^n I(\theta \geq y_{max}) \frac{\alpha y_m^\alpha}{\theta^{\alpha+1}} I(\theta \geq y_m)$$

where $y_{max} = max(y_1, ..., y_n)$. Assuming $y_m \geq y_{max}$, we have:

$$p(\theta|y) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+n+1}} I(\theta \geq y_m)$$

5. In this class, we've mostly talked about the assumption of exchangeability in the context of posterior predictions. We've talked about how posterior predicted observations are assumed to be exchangeable with observations in the original sample. We make this assumption because it allows us to write down a mathematical form for the posterior predictive data generating process. Additionally, the assumption of independence (and thus exchangeability) among all n observations in a sample allows us to write down a simple mathematical form for the joint data distribution. So far in class, we've made the assumption of exchangeability before we obtain posterior predictive distributions and posterior distributions for parameters. In this way, the exchangeability assumption "precludes formulation of

uncertainty as a numerical probability."

As far as assessing the exchangeability assumption, it's very easy to say that the assumption of exchangeability is not violated if we know very little about individual observations. If we find out more details, however, then we often find out about violations of the exchangeability assumption. The example we discussed in class illustrates this well. When we have a random sample of six students but know nothing about them, we know of no violations of exchangeability. When we find out that the students came from two different unknown classes, we are no longer comfortable with the exchangeability assumption. So, even though ignorance often makes assessing assumptions easier, I would never say that as a statistician I am *willfully* ignorant. If I am helping analyze data from a study, I want to know as much as I can about the "individual circumstances" so that I can incorporate important variables into the model. At some point we are forced to make the exchangeability assumption, but I would still say that I'd rather be informed about individual differences present in a study so that I can reasonably assess whether the assumption is reasonable. So even though exchangeability does relate to the author's statement, "It is only by ignoring certain specific considerations pertaining to the event or statement in question that a mathematical probability becomes possible," I'd rather be knowledgeable about those "specific considerations" so that I can decide if the results that rely on the exchangeability assumption can be trusted.

6. (a) The data are counts; for each island the number extinct are given. I am going to treat these as Poisson counts. They are *actually* binomial counts, but we are not given the denominator (total number of species), so we cannot model them as binomial counts. I'll hope that the denominators are large enough that the Poisson approximation is appropriate. The Poisson data generating model is:

$$y_i \sim Poi(\lambda)$$
$$y_i = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

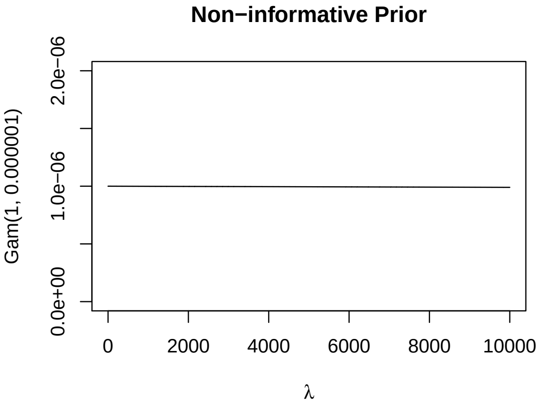
The assumptions we are making are:

- Every island has the same number of bird species at the beginning of the study in 1949. I discussed this one above. I hope the number of species on each island is large enough and similar enough across islands to make this assumption valid.

- Every island has the same extinction rate parameter, λ. I think this assumption is not reasonable. I would expect some islands have higher extinction rates than others depending on climate and other covariates.
- Independence among islands. Because the islands are separated by water, I think it would be hard for diseases, etc to travel from one island to the next. But, islands that are more populated may have more similar extinction counts than islands that are less populated. This could violate the independence assumption.

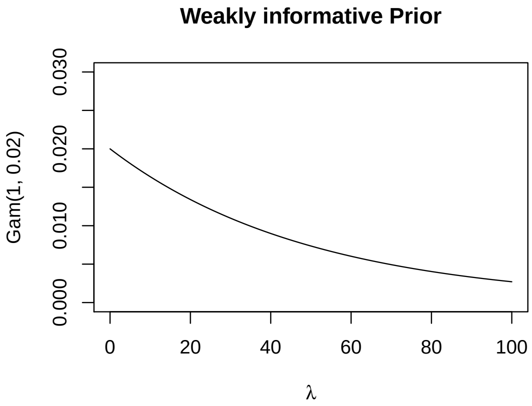
(b) i. For a non-informative prior, I would like a distribution that gives equal probability to all positive values. This problem was tricky. My goal was to find a gamma distribution that resembles a uniform distribution over a large range of positive values. I chose a $Gam(1, 0.000001)$ distribution. I plot this distribution below over 0 to 10000, and it looks very similar to a $Unif(0, 10000)$ distribution. It is not exactly a uniform distribution, larger values do have *slightly* smaller probabilities, but I figure the difference is negligible. The only knowledge I'm reflectin in this prior is that the parameter is greater than 0, and this is a consequence of using the gamma distribution as the prior.

I will point out that this is one of those "proper" prior distributions that could still result in an improper posterior.

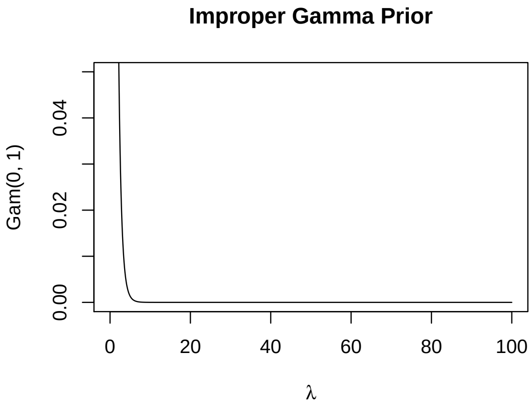


ii. Suppose I found out that the total number of bird species on each island is no more than

100, and I found out that in general lower extinction numbers are more likely than higher ones. I could incorporate this knowledge by assigning a weakly informative $Gam(1, 0.02)$ prior, with probabilities generally decreasing between 0 and 100. This prior is still vague, but I am incorporating knowledge about a general trend and a natural constraint. I wish I could make values over 100 have zero probability, but I'm not sure this can be done with the gamma distribution.



iii. I chose the improper $Gam(0, 1)$ prior. This prior gives really high probabilities to low values of λ and really low probabilities to values of λ greater than 7 or 8.



iv. Below are the questions and I asked the researcher and the answers I received:

A. **What do you expect the mean number of birds going locally extinct on the island to be?**

Not sure what island specifically you are referring to? If you're thinking about an average number over all islands, I would guess around 5, but I'm not super confident in that.

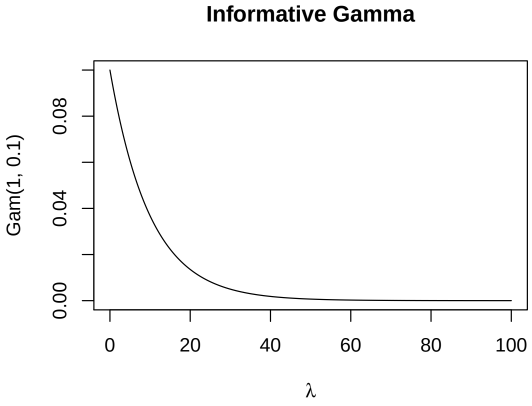
We definitely noticed in observations that we weren't seeing species we used to see, but hard to pin it down to one expected number.

B. **What is your estimated probability that the mean is between 1 and 20?** *Not quite sure how to assign a probability here – are you asking if I think it would be rare to have 0 or more than 20 birds go extinct for an island? I think it would be rare to have more than 20, but I probably wouldn't say it's impossiblejust not likely. An island could have 0 birds go extinct..that's definitely plausible.*

C. **Or, if the last question was easy for you, what is your estimated probability that the mean is between 1 and 10?**

Last question was not so easy for me :) – but I would say I think 15 would be really high for an island, and 10 would be fairly high (though would definitely expect it for some of the islands). I'm not so sure about how to think about "the mean" you're referring to, so these are answers about what I would expect for an island.

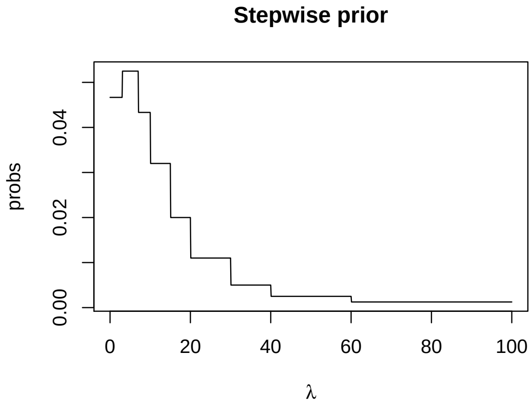
After considering the researcher's answers, and the possible shapes from the gamma distribution, I chose a $Gam(1, 0.1)$ prior. I would have preferred a shape that gave slightly lower probability to 0, but I couldn't figure out how to get this shape from the gamma distribution without giving a probability of zero to 0!. I think the $Gam(1, 0.1)$ prior is reasonable. Values of λ between 0 and 10 are most probable, and values greater than 20 are less likely but still possible.



v. I put this together by trial and error. Probabilities generally decrease for higher values of λ, but values between 0 and 3 have lower probability than values between 4 and 7. The code is shown below, along with the plot of the prior.

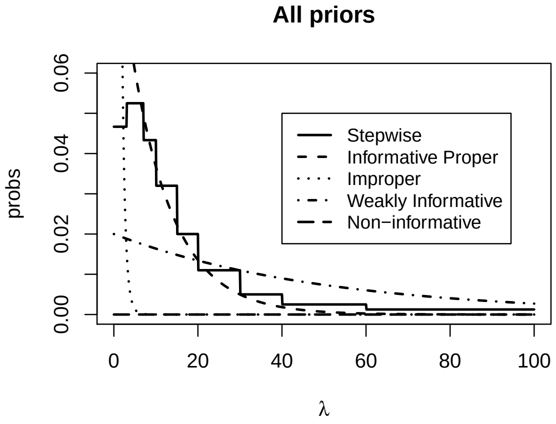
```
prior.fun <- function(theta){
  if(theta <= 3) {out <- 0.14/3}
  if(theta > 3 & theta <= 7) {out <- 0.21/4}
  if(theta > 7 & theta <= 10) {out <- 0.13/3}
  if(theta > 10 & theta <= 15) {out <- 0.16/5}
  if(theta > 15 & theta <= 20) {out <- 0.1/5}
  if(theta > 20 & theta <= 30) {out <- 0.11/10}
  if(theta > 30 & theta <= 40) {out <- 0.05/10}
  if(theta > 40 & theta <= 60) {out <- 0.05/20}
  if(theta > 60 & theta <= 100) {out <- 0.05/40}
  return(out)
}

x <- seq(0, 100, by=0.1)
input <- data.frame(x)
probs <- apply(input, 1, prior.fun)
plot(x, probs, type="l", xlab=expression(lambda), main="Stepwise prior")
```

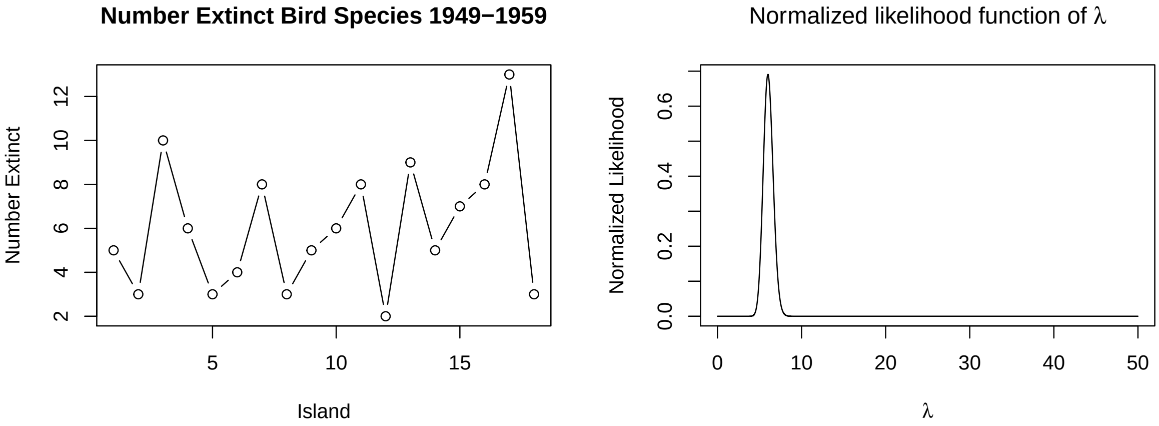


```
#check that it sums to 1
sum(0.1*probs[-c(100)])
## [1] 1
```

(c) The plot below shows all five priors together.



(d) The plots of the observed data and the likelihood function are shown below.



(e) On HW 3 I proved mathematically that for a Poisson likelihood and a $Gam(a, b)$ prior, the posterior distribution is $Gam(\sum_{i=1}^{n} y_i + a, b + n)$. I use this result to find the posterior distributions for priors (i) through (iv).

i. $Gam(1, 0.000001)$ prior → $Gam(108 + 1, 18 + 0.000001)$ posterior

ii. $Gam(1, 0.02)$ prior → $Gam(108 + 1, 18 + 0.02)$ posterior

iii. $Gam(0, 1)$ prior → $Gam(108 + 0, 18 + 1)$ posterior

iv. $Gam(1, 0.1)$ prior → $Gam(108 + 1, 18 + 0.1)$ posterior

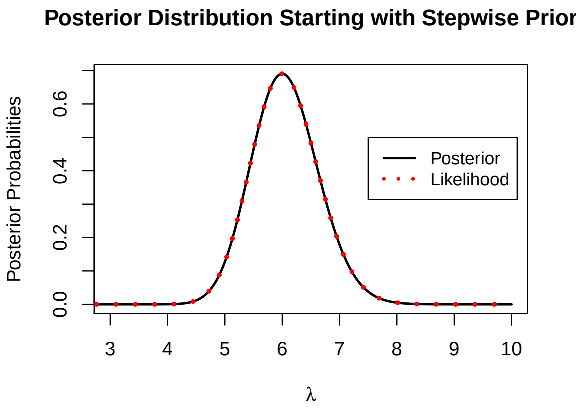
v. The function used to approximate the posterior probability distribution for λ is shown below for the stepwise prior. I found that the posterior distribution for this prior is almost identical to the normalized likelihood function.

```
#likelihood function
like.pois <- function(lambda){
  lambda^{sum(extinctions$extinct)}*exp(-18*lambda)/prod(factorial(extinctions$extinct))
}

#Write a function to calculate the posterior probability for each value of lambda
post.fun <- function(lambda) {
  prior.fun(lambda)*like.pois(lambda)
}

#Look at lambdas between 0 and 10
lambda <- seq(0, 10, by=0.001)

#Normalize the posterior distribution
numint.post <- sum(0.001*post.fun(lambda))
post.probs <- post.fun(lambda)/numint.post
```



(f) The posterior probabilities and intervals are shown for each posterior in the table below. My code for approximating these quantities for the posterior distribution found computationally are shown below.

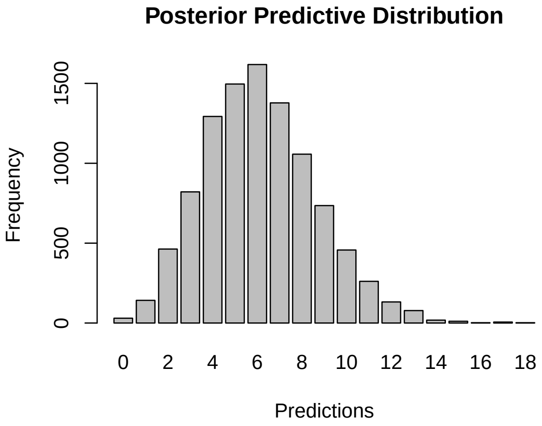
```
require(mosaic, quietly=TRUE)
set.seed(49)
post.samp <- sample(lambda, 10000, prob=post.probs, replace=TRUE)
quantile(post.samp, 0.005)
quantile(post.samp, 0.995)
1-pdata(5, post.samp)
```

Prior	Posterior	99% Post Interval	Post Prob > 5
Gam(1, 0.000001)	Gam(109, 18.000001)	(4.666, 7.654)	0.972
Gam(1, 0.02)	Gam(109, 18.02)	(4.661, 7.645)	0.971
Gam(0, 1)	Gam(108, 19)	(4.374, 7.192)	0.898
Gam(1, 0.1)	Gam(109, 18.1)	(4.640, 7.611)	0.968
Stepwise	shown above	(4.672, 7.655)	0.970

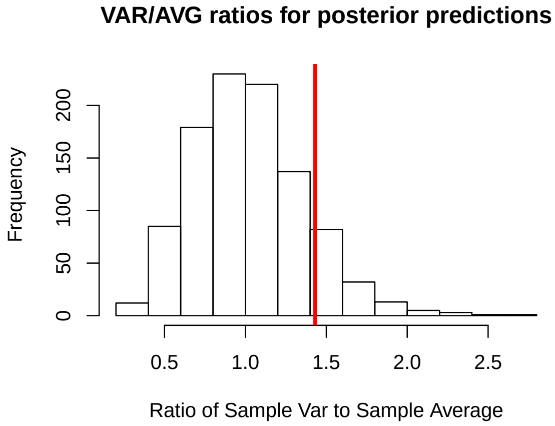
(g) The results are similar across the non-informative, weakly informative, informative ($Gam(1, 0.1)$), and stepwise priors. I think that the posterior distributions for the non-informative and weakly informative priors are similar to the stepwise and $Gam(1, 0.1)$ informative priors because these informative priors turned out to be consistent with the data observed. The one prior that gave different results was the improper $Gam(0, 1)$ prior. The improper $Gam(0, 1)$ prior is highly informative, giving very high weight to small values of λ. This prior clearly influenced the posterior distribution in this example by drawing more density towards lower values of λ. In this case, I believe that the information provided by the improper Gamma prior was incorrect, and I don't trust the results from the Gamma prior as much as I trust the results from the other priors. I think this shows that it is better to incorporate more uncertainty into a prior rather than choosing a highly informative prior that may be incorrect.

(h) The approximate posterior predictive distribution found starting from the stepwise prior is shown below. Here are the steps I went through to construct the approximation:

- Draw a random λ from the approximate posterior distribution found in part (e). To do this, I use the sample function to sample from the λ vector with probabilities equal to the posterior probabilities found in part (e).
- Draw a random observation from a $Poi(\lambda)$ distribution where λ is the one you drew above.
- Repeat steps 1 and 2 10000 times. Note: Computationally, I did it with vectors, but for understanding I think it makes more sense to explain it this way.
- Ignore λ for each pair and plot the observations. This is the approximate posterior predictive distribution.



(i) This ratio is of particular interest because the mean of the observations is assumed to be the same as the variance of the observations when assuming a Poisson model for the data generating process. In the posterior predictive samples, the predictions were generated under the assumption of variance equal to the mean, so the ratios of the sample variances to the sample averages for the posterior predictions should be centered around 1. If they weren't centered around 1, I would suspect that I did something wrong when generating the posterior predictions. In the observed data, however, I identified multiple violations of assumptions in the first part of this problem that could manifest as overdispersion. As a result, I am not surprised that the sample variance in the observed data is larger than the sample average. This does not provide evidence of overdispersion, however, because the original ratio of 1.43 or more extreme was observed 16.8% of the time when data were generated from a true Poisson process. This is a useful modeling check because if the observed ratio was way out in the tails of the below distribution, it would suggest that we should go back and rethink the model used to describe the data generating process.



7. (a) I will assume a binomial data generating process where each response is a binomial count of the number of days where an egg was laid in 14 days. The parameter, θ, is the probability of laying an egg in a day. I'd like to say that I think it is more appropriate to use a poisson process here, but because they want to estimate the probability of laying an egg in a day, I am going to use

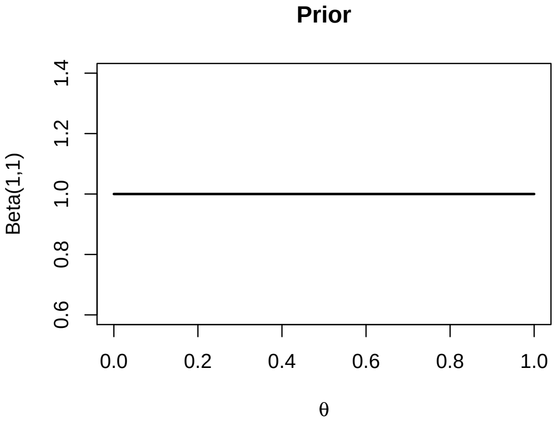
a binomial process. Most of the assumptions are not met.

$$y_i \sim Binom(\theta, 14)$$
$$f(y_i) = \binom{14}{y_i}\theta^{y_i}(1-\theta)^{14-y_i}$$

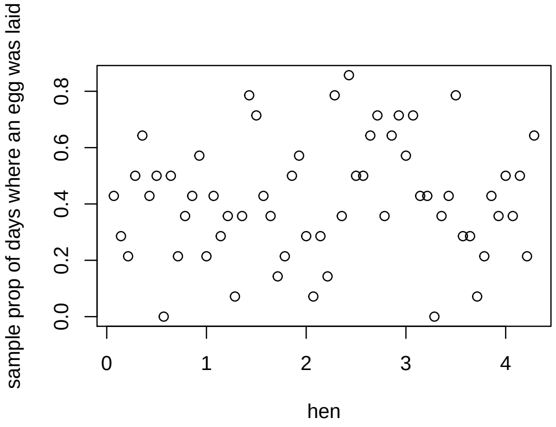
Assumptions:

- A hen will only lay 0 or 1 eggs in a day. She will never lay more than 1 egg in a day. This makes it so that each day is a binary response going into the binomial count. If she could lay two or more eggs in a day, it wouldn't be appropriate to use a binomial process.
- Every hen has the same probability of laying an egg in a day. The description acknowledges that this assumption is not met, but they want to estimate an overall probability.
- There is independence among days for each hen, so that if a hen lays an egg on day one it does not affect the probability of her laying an egg on subsequent days. I don't think this one is met. Biologically, if a hen lays an egg on day one, I expect she would be less likely to lay an egg on day two.
- There is independence among hens so that if hen 1 lays an egg, this does not affect the probability of the other hens laying eggs. I think we'd need more information to assess this assumption. There could be some spatial dependence among the hens if their roosts are in close proximity.

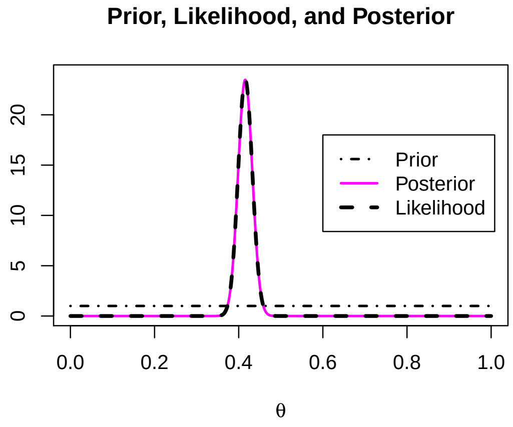
(b) I think a $Beta(1, 1)$ distribution is most appropriate. At first, I wanted to choose a $Beta(2, 2)$ distribution so that the most probable prior value of θ would be 0.5, but I couldn't figure out how to get this same shape out of a Beta distribution without completely ruling out the possibility of θ being 0 or 1. In the last problem, I saw that it was better to choose a weakly informative prior rather than a more informative one that might be incorrect. I think biologically it *is* possible that a hen could lay 0 or 14 eggs in the two week period. The $Beta(1, 1)$ prior allows for this possibility and is a conjugate prior for the binomial likelihood. With a sample size of 60, I do not think that the resulting posterior distribution will be affected much by the fact that we did not incorporate the researcher's prior knowledge that hens lay an egg every other day on average.



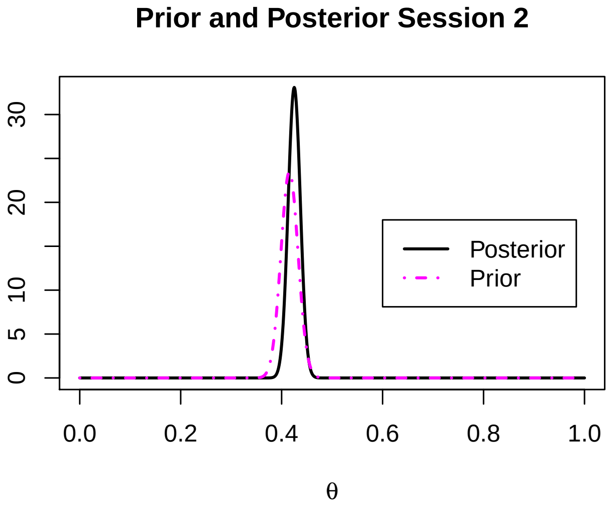
(c) The data are plotted below. You can see that the sample proportion of days where an egg was laid ranges between 0 and 0.86 for different hens.



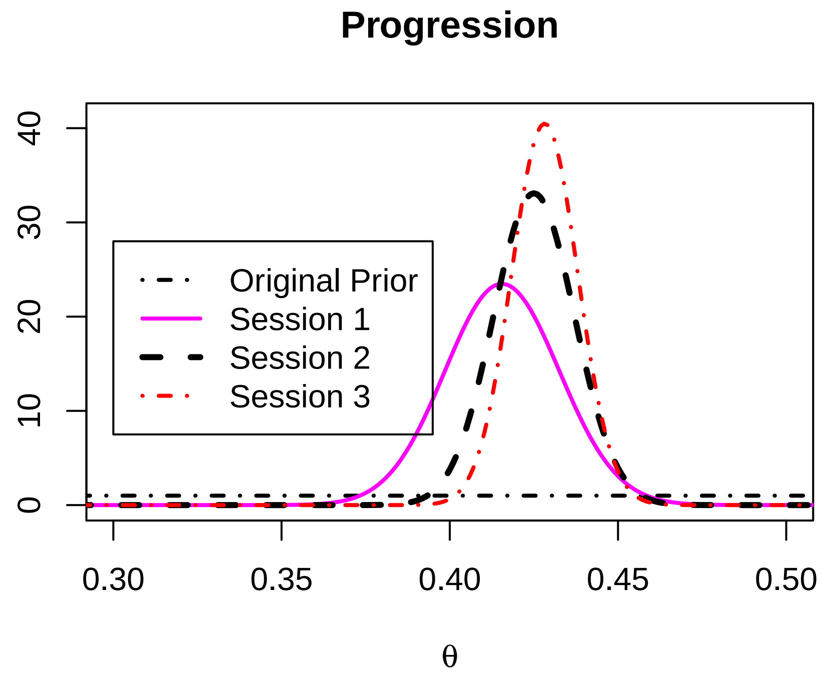
The total number of days where eggs were laid, over all hens, is 349. The total number of days where eggs were not laid, over all hens, is 491. In homework 2 we showed that for a $Beta(a, b)$ prior and a $Binom(n, \theta)$ likelihood, the posterior distribution is $Beta(\sum_{i=1}^{n} y_i + a, b + \sum_{i=1}^{n}(n - y_i))$. So, with a $Beta(1, 1)$ prior, the posterior distribution is $Beta(349 + 1, 491 + 1)$. The normalized likelihood, prior, and posterior are shown on the plot below.



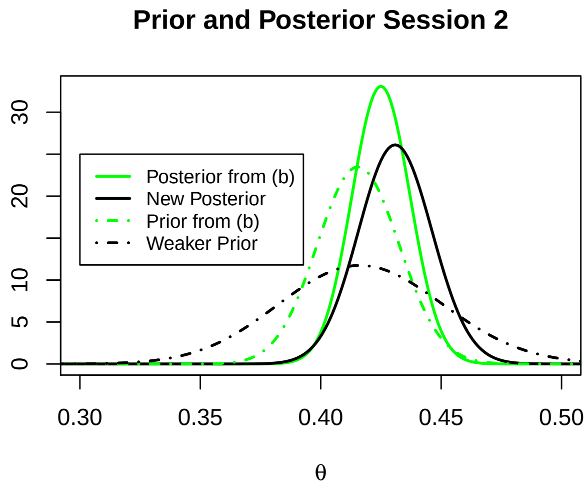
(d) If we use a $Beta(350, 492)$ prior for analyzing the second two week session of data collection, then the posterior for this two week session would be $Beta(365 + 350, 475 + 492)$. This prior and posterior are shown on the plot below. I understand why this prior could be justified, because knowledge from a previous study should be used to reflect prior knowledge in subsequent studies. But, I think the posterior from session 1 is too narrow to use as a prior. I think it puts a lot of weight on the data from session 1 and almost completely rules out values of θ less than 0.35 and greater than 0.5.



(e) Now I wil use a $Beta(715, 967)$ prior for analyzing the third session of data collection. The posterior for the third session will be $Beta(365 + 715, 475 + 967)$. The progression of knowledge is shown below.

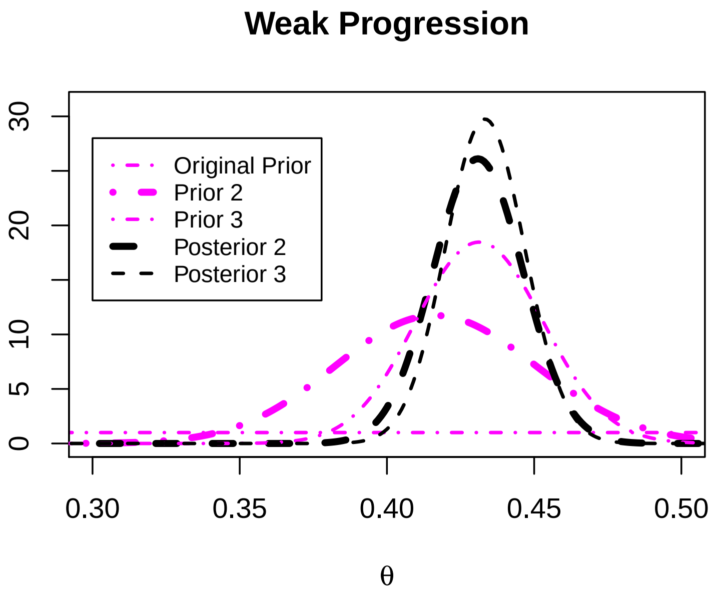


(f) i. To make the prior weaker, I simply divide the number of (previous) posterior successes and failures by four. This gives me a prior distribution of $Beta(88, 123)$ for Session 2. Dividing the posterior parameters from Session 1 by four keeps the distribution centered at the same value, but it spreads the distribution out more, resulting in a weaker prior. I think the prior used here is much more reasonable than using the exact posterior from Session 1. By incorporating more uncertainty into the previously found posterior, we don't rule out so many values for θ, but we still incorporate knowledge from the previous session. Under this prior, the posterior distribution for Session 2 is $Beta(453, 498)$.



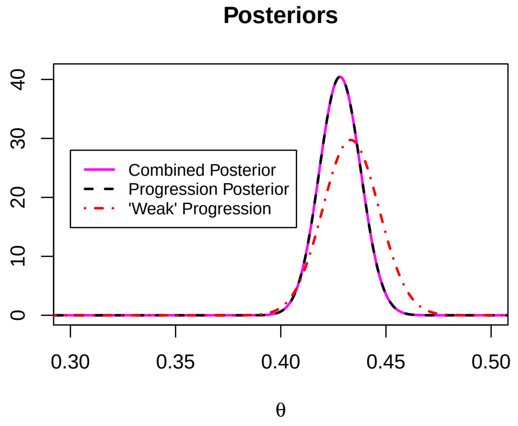
ii. For session 3, I take the posterior from session 2 and divide the parameters by two. I only divide in half this time because I trust the posterior from the second session more than I trust the posterior from the first session; it now reflects knowledge from two studies rather than one. The prior distribution for Session 3 is then $Beta(227, 299)$ and the posterior distribution for Session 3 is $Beta(592, 774)$.

(g) The progression is shown below. I zoomed in a bit so you could make out the details.



(h) If we combine all the data over three time periods, there were 1079 days on which eggs were laid, and 1441 days on which eggs were not laid (over all hens). If we start with a $Beta(1,1)$ prior, then the posterior distribution is $Beta(1079+1, 1441+1)$. The posteriors found in the previous parts are shown on the following plot, along with the posterior found by combining the three time periods into a single dataset.

As you can see, the posterior distribution is the same whether we combine the three sessions into a single dataset or whether we find the posterior from each session and use it as the prior for the next session. Wow! The posterior is different, however, if we weaken the posterior from previous sessions before using them as priors for the next session. I still stand by what I said before. The first two posteriors put a lot of weight on the observed data. I think we should take advantage of the fact that the study was repeated three times, and use this opportunity to incorporate uncertainty at each stage. I think the posterior from the weak progression should be used for inference. It is also more conservative, and I generally choose to report the more conservative results.



(i) I would account for covariates in the estimation of θ by using a regression equation to predict values of θ for given values of the covariates. My proposed model is below. We can assign priors and find posteriors for the regression parameters β_0, β_1, and β_2.

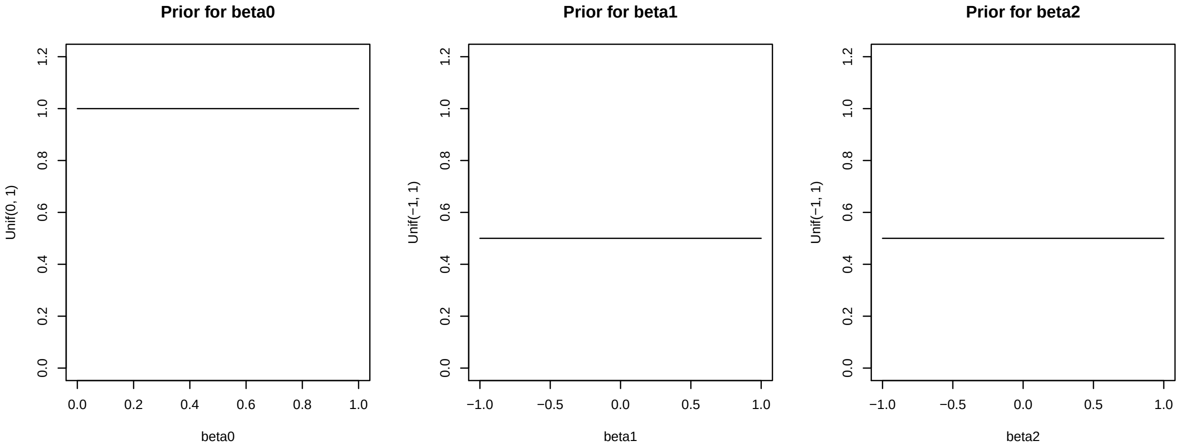
$$\begin{aligned}
y &\sim Binom(\hat{\theta}, 14) \\
\hat{\theta} &= \beta_0 + \beta_1 I_{fat} + \beta_2 I_M \\
\beta_0 &\sim Unif(0, 1) \\
\beta_1 &\sim Unif(-1, 1) \\
\beta_2 &\sim Unif(-1, 1)
\end{aligned}$$

where I_M is 1 if the hen is molting and 0 if not, and I_{fat} is 1 if eating more than 80 grams of food per day, and 0 if otherwise.

The first covariate is a measure of the the amount of food the birds eat each day. I made this an indicator variable so that it would be easy to put a prior on the coefficient associated with it, β_1. β_1 represents the difference in probabilities of laying an egg between hens that eat more than 80 grams of food per day and hens that eat less than or equal to 80 grams. I know nothing about hens and how food consumption is related to egg laying, so I will put a $Unif(-1, 1)$ prior on β_1 to allow for the possibility that an increase in food intake is associated with an increase or a decrease in the probability of laying an egg.

The second covariate is an indicator variable that is 1 if they are molting and 0 if they are not. β_2 is then the difference in the probability of laying an egg in a day between molting hens and non-molting hens. I have no idea whether molting hens are more likely to lay eggs than non-molting hens, so I will put a $Unif(-1, 1)$ prior on β_2.

β_0 is the value of θ when they are not molting and when they are eating 0 pounds of food per day. Because I don't know anything about the effect of food consumption and molting on the probability of laying an egg, I'll use a non-informative $Unif(0, 1)$ prior for β_0 as well.


Prior for beta0
Unif(0, 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
beta0
Prior for beta1
Unif(−1, 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−1.0
−0.5
0.0
0.5
1.0
beta1
Prior for beta2
Unif(−1, 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−1.0
−0.5
0.0
0.5
1.0
beta2