Lecture 22

Peter Shaffery

4/8/2021

Estimating the Binomial Parameter

Say that we have a coin, and we would like to learn π, the probability that the coin will come up heads. We toss the coin n times, of which we observe k heads. What does Bayes theorem look like here?

$$P[\pi|k] \propto P[k|\pi]P[\pi]$$

Notice that I am omitting the evidence term here, $P[k]$. That is because it has no dependence on the parameter π, it's just a normalizing constant. Therefore it's common to just deal with posteriors as proportional to $P[k|\pi]P[\pi]$, rather than writing out the whole fraction.

The likelihood for this situation is familar to us, it's the Binomial distribution:

$$P[k|\pi] = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

And furthermore, observe that the $\binom{n}{k}$ term is constant in π, so we'll ignore it:

$$P[k|\pi] \propto \pi^k(1-\pi)^{n-k}$$

Leaving our prior general for now, our posterior distribution is therefore:

$$P[\pi|k] \propto \pi^k(1-\pi)^{n-k}P[\pi]$$

Now, which prior should we use? The only restriction that we have is that it must be a valid probability distribution. Otherwise, the choice of prior is up to us. Let's see a few options.

Prior 1: Uniform Prior

The most obvious (and probably the most common) prior you will ever see is the uniform prior:

$$P[\pi] = \begin{cases} 1, \pi \in [0,1] \\ 0, \text{ else} \end{cases}$$

We've actually used this prior in the previous marble example, where we justified it using an "indifference" argument.

Similarly, the idea with using a uniform prior here would be to express an initial position of *ignorance*. If you are truly starting your coin tossing experiment without any knowledge of the coin, then this is indeed an appropriate choice of prior. No value of π is initially prioritized over another.

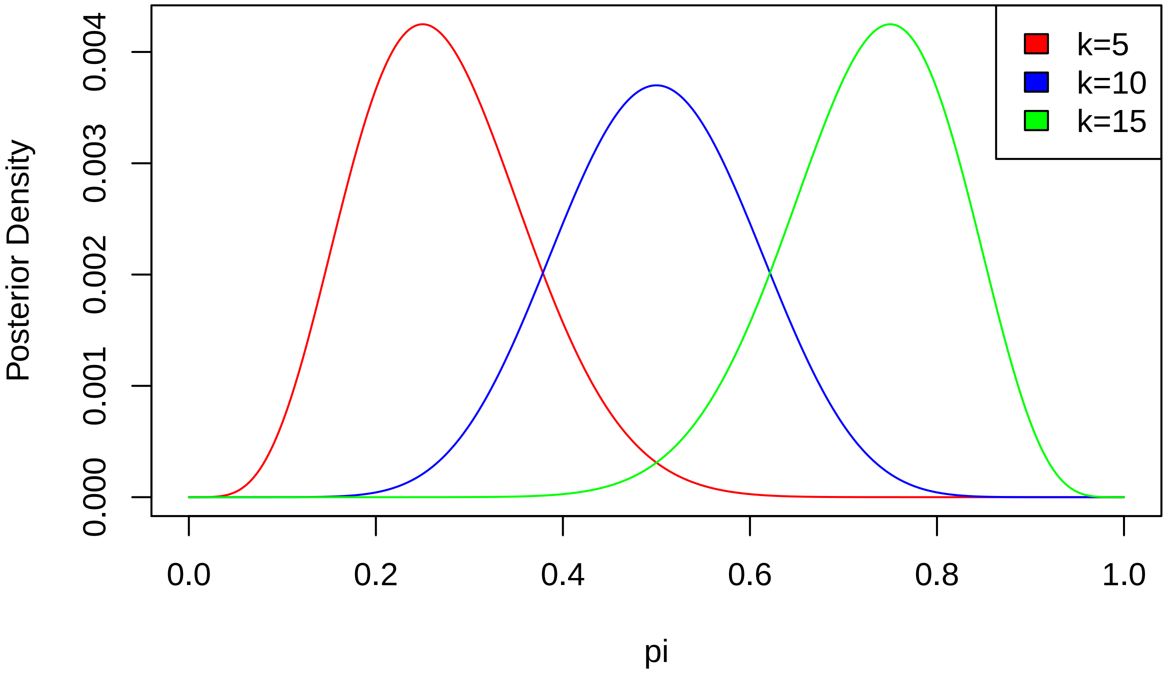
From this choice of prior we obtain our posterior distribution:

$$P[\pi|k] \propto \pi^k(1-\pi)^{n-k}\nVdash_{\pi\in[0,1]}$$

Where $\nVdash_{\theta\in[0,1]}$ is an indicator function.

Let's visualize how this posterior looks for $n = 20$, and $k = 5, 10$, and 15:

```
library(tidyverse)
n=20
pi.grid = seq(0,1,.001)

post1= (pi.grid^5)*(1-pi.grid)^(n-5)
post2 = (pi.grid^10)*(1-pi.grid)^(n-10)
post3 = (pi.grid^15)*(1-pi.grid)^(n-15)


plot(pi.grid,post1/sum(post1),type='l',col='red',xlab='pi',ylab='Posterior Density')
lines(pi.grid,post2/sum(post2),col='blue')
lines(pi.grid,post3/sum(post3),col='green')
legend(x='topright',c('k=5','k=10','k=15'),fill=c('red','blue','green'))
```

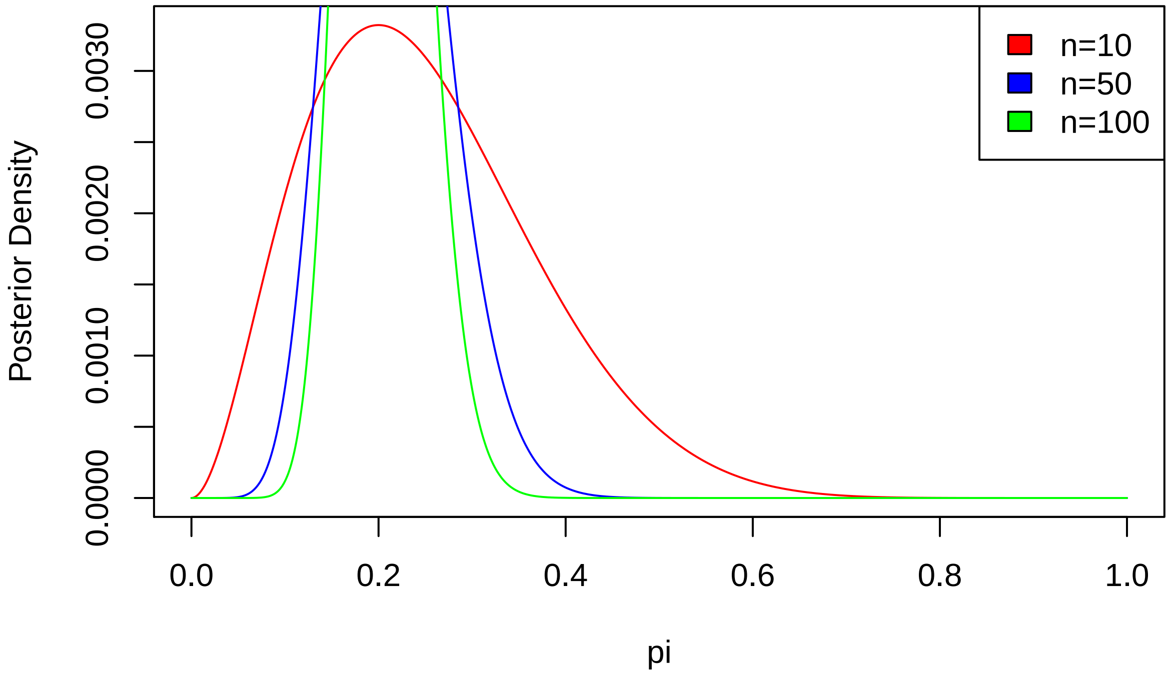


We see that, for each value of k, the mass of our posterior density covers different ranges of π. When k is large, our posterior assigns belief to higher values of π than lower ones.

Now let's fix k at 20% of n, and look at how our posterior changes as we increase $n = 10, 50, 100$:

```
library(tidyverse)
pi.grid = seq(0,1,.001)

post1= (pi.grid^(.2*10))*(1-pi.grid)^(.8*10)
post2 = (pi.grid^(.2*50))*(1-pi.grid)^(.8*50)
post3 = (pi.grid^(.2*100))*(1-pi.grid)^(.8*100)


plot(pi.grid,post1/sum(post1),type='l',col='red',xlab='pi',ylab='Posterior Density')
lines(pi.grid,post2/sum(post2),col='blue')
```

```r
lines(pi.grid,post3/sum(post3),col='green')
legend(x='topright',c('n=10','n=50','n=100'),fill=c('red','blue','green'))
```



Whereas increasing k changes the *centrality* of the posterior (eg. it's mean), increasing n changes its *spread*. This is interpreted as our model consolidating around a smaller and smaller range of plausible values of π, as more and more data become available.

This feature is common across posterior distributions. Assuming some (fairly general) conditions on your data's sampling distribution, then as you collect more and more data your posterior distribution's mass will consolidate around the true parameter values.

Both this, and the centrality feature suggest some important ways of summarizing the information in our posterior distribution: posterior mean and posterior variance:

$$E_{\theta|y}[\theta] = \int \theta P[\theta|y]d\theta$$

$$\text{Var}_{\theta|y}[\theta] = \int (\theta - E_{\theta|y}[\theta])^2 P[\theta|y]d\theta$$

These are pretty much what they sound like- the mean and variance of the posterior distribution. These quantities are fundamental to Bayesian statistics, and you can think of them as the Bayesian version of the MLE ($\hat{\theta}$) and standard error ($s.e.(\theta)$)

Introducing the Beta Distribution

A random variable x whose probability density functions is of the form:

$$P[x|a,b] \propto x^{a-1}(1-x)^{b-1}$$

Are known as a **Beta Distributed**, and their distribution is the Beta Distribution. Again note that above I am omitting a normalizing constant, which is sometimes instead written in as:

$$P[x|a,b] = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

Beta distributions are very important in Bayesian statistics, mainly because of the above example and similar. We can see that our posterior from above with a uniform prior:

$$P[\pi|k] \propto \pi^k(1-\pi)^{n-k}$$

Is a beta distribution with parameters $a = k+1$ and $b = n-k+1$. This means that, among other things, we can then use standard formulas for the mean and variance of the Beta distribution to compute our posterior mean and variance:

$$\begin{aligned} E_{\theta|y}[\theta] &= \frac{a}{a+b} = \frac{k+1}{n+2} \\ \mathrm{Var}_{\theta|y}[\theta] &= \frac{ab}{(a+b)^2(a+b+1)} = \frac{(k+1)(n-k+1)}{(n+2)^2(n+3)} \end{aligned}$$

Pay close attention to the posterior mean: $E_{\theta|y}[\theta] = \frac{k+1}{n+2}$. This formula is known as the *rule of succession*, and provides a popular alternative estimator for π besides the usual MLE $\hat{\pi} = \frac{k}{n}$. Here the additional terms 1 and 2 in the numerator and denominator reflect the influence of our *prior.* Notice that when $k = n = 0$ the MLE estimate cannot be computed, however the rule of succession says that our posterior mean is $E_{\theta|y}[\theta] = .5$. This corresponds to our preference not to favor any value of θ over another.

This formula was originally derived Pierre-Simon Laplace, who used it to calculate the probability that the Sun will rise tomorrow. His argument went that, since $k = n = d \approx 30 \text{ years } \times 365 \text{ days } = 109500$ (the number of days had had seen the Sun rise), then:

$$E[\text{Sun will rise tomorrow}|d] = \frac{k+1}{k+n+2} = \frac{d+1}{d+2} \approx 0.9999909$$

Observe that no amount of observations will ever quite get $E[\text{Sun will rise tomorrow}|d] = 1$.

Prior 2: Beta Prior

Now, Laplace realized that this was an absurd result. By bringing additional information to the table (specifically orbital mechanics) most people will realize that $P[\text{Sun will rise tomorrow}] = 1$, even without any observations.

The problem with the reasoning above then is not necessarily a flaw in Bayesian statistics, but rather in our choice of a uniform prior.

Although in most cases you don't have quite as strong prior knowledge as from orbital mechanics, in many cases we have *some* information about our parameters θ. In order to represent this, we can use what is a called an **informative prior**. Unlike a uniform prior, which looks the same everywhere, an informative prior places more probability mass over one range of θ than any other, representing that we initially believe that range of θ to be the most plausible.

For our coin-tossing example a popular choice of informative prior is the Beta distribution. This is for two reasons:

1. The Beta distribution with $a = b = 1$ is just a uniform distribution, so we can view the uniform prior as a "special case" of the beta prior
2. If we use a Beta prior, and have a Binomial likelihood, then our posterior is guaranteed to be a Beta distribution.

Let's see Property 2 in more detail. Say that our prior is $P[\pi] \propto \pi^a(1-\pi)^b$, then our posterior is:

$$
\begin{aligned}
P[\pi|k,n] &\propto \pi^k(1-\pi)^{n-k}\pi^a(1-\pi)^b \\
&\propto \pi^{k+a}(1-\pi)^{n-k+b}
\end{aligned}
$$

Thus our posterior is a Beta distribution with updated parameters $a^* = k + a + 1$ and $b^* = n - k + b + 1$.

Aside - Conjugate Priors

This type of relationship, between the Binomial and Beta distributions, has historically been very important in Bayesian statistics. We say that the Beta distribution is a **conjugate prior** to the Binomial distribution. More generally, if we have a prior of Distribution A, and a likelihood of Distribution B, then A and B are conjugate if the posterior is also of type A.

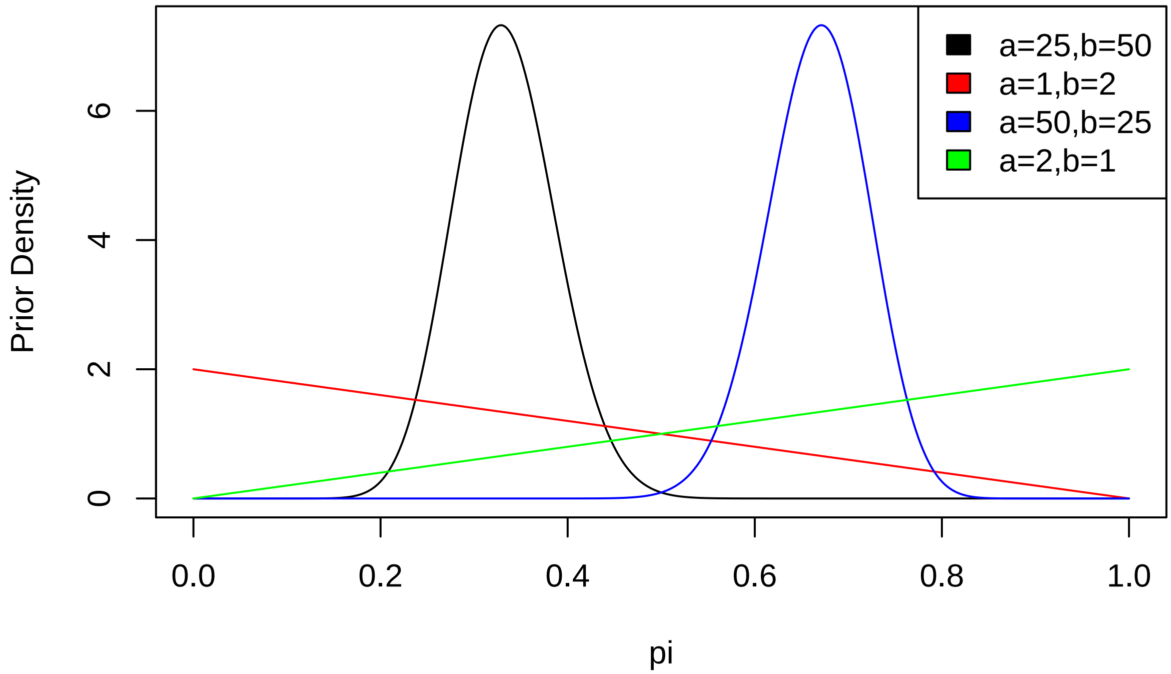
Conjugate priors used to matter quite a lot, because they made Bayesian computation possible. Now that computers exist they matter a little less, but many of the most common Bayesian models will still use conjugate priors.

</aside>

So the Beta prior in principal allows us to construct an informative prior, but how do we sensibly choose it's parameter values a and b? Well, notice that, algebraicaly, they take a similar role as the number of successes and failures in $a + b$ trials (recall that with the uniform prior our updated $a^* = k + 1$ and $b^* = n - k + 1$). We can use this interpretation to select a prior.

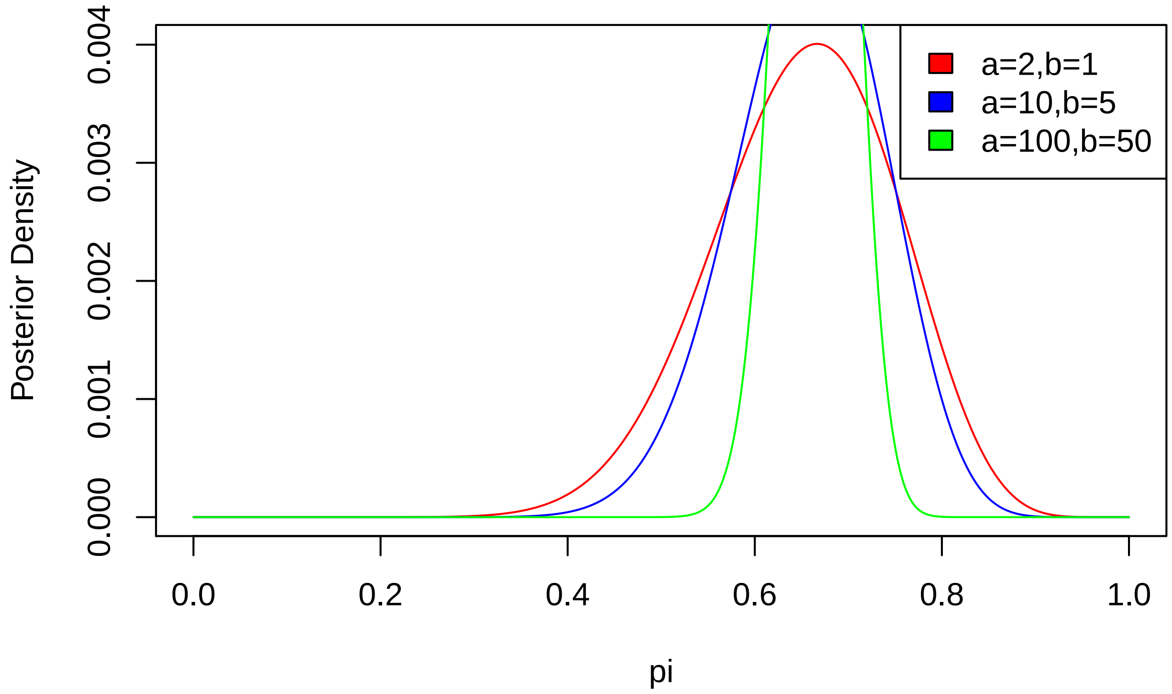
When $a + b$ (which we can think of as our "prior sample size") is large, then this will be a *highly* informative prior, and when $a > b$ then it will favor values of $\pi > .5$. Conversely if $a + b$ is small, then the prior will be closer to the uniform distribution, and if $a < b$ then our prior will favor $\pi < .5$:

```
prior1 = dbeta(pi.grid, 25, 50) # high a+b, b>a
prior2 = dbeta(pi.grid, 1, 2) # low a+b, b>a
prior3 = dbeta(pi.grid, 50, 25) # high a+b, a>b
prior4 = dbeta(pi.grid, 2, 1) # low a+b, a>b

plot(pi.grid,prior1,type='l',xlab='pi',ylab='Prior Density')
lines(pi.grid,prior2,col='red')
lines(pi.grid,prior3,col='blue')
lines(pi.grid,prior4,col='green')
legend(x='topright',c('a=25,b=50','a=1,b=2','a=50,b=25','a=2,b=1'),fill=c('black','red','blue','green'))
```



This interpretation of the prior parameters enables us to encode our prior beliefs. Say that our friend tells us that the coin is twice as likely to be heads than tails. We would therefore choose a Beta prior where $a = 2b$. However, depending on how much we trust our friend, make this prior more or less informative. Let's look at the posteriors we'd obtain with priors $a = 2, b = 1$, $a = 10, b = 5$, and $a = 100, b = 50$. Say that we have performed $n = 20$ trials, and observed $k = 13$ successes:

```
prior1 = dbeta(pi.grid,2,1)
prior2 = dbeta(pi.grid,10,5)
prior3 = dbeta(pi.grid,100,50)

k=13
n=20
post1= prior1*(pi.grid^(k))*(1-pi.grid)^(n-k)
post2= prior2*(pi.grid^(k))*(1-pi.grid)^(n-k)
post3= prior3*(pi.grid^(k))*(1-pi.grid)^(n-k)

plot(pi.grid,post1/sum(post1),type='l',col='red',xlab='pi',ylab='Posterior Density')
lines(pi.grid,post2/sum(post2),col='blue')
lines(pi.grid,post3/sum(post3),col='green')
legend(x='topright',c('a=2,b=1','a=10,b=5','a=100,b=50'),fill=c('red','blue','green'))
```



As we might expect, a more informative prior has the same effect on our posterior distribution as collecting more data. In this way, a prior literally encodes the information that you are bringing to the analysis before sampling data.

Bayesian Analysis of Normally Distributed Data

Simplest Case: Known Variance, Single Observation

Now that we've seen the basics of what Bayesian statistics looks like, let's turn to a more useful example: estimating the mean of a normal distribution. To start we'll assume that the variance of this distribution (σ^2) is known.

Say that we have a single observation y, our likelihood is then:

$$P[y|\theta] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}(y-\theta)^2\right]$$

Now, which prior should we choose for θ?

As before, one option is a uniform prior, so that the posterior is exactly equal to the likelihood:

$$P[\theta \; y] \propto \exp\left[\frac{-1}{2\sigma^2}(y-\theta)^2\right]$$

In this case, because the mean of a normal distribution is the same as its mode, our Bayesian estimate of θ, $E_{\theta|y}[\theta]$ will be exactly equal to the MLE.

However, we could also use the conjugate prior for a normal distribution to construct an informative prior. It turns out that the conjugate of a normal is also a normal distribution:

$$P[\theta] \propto \left[\frac{-1}{2\tau^2}(\theta-\mu)^2\right]$$

One extra bit of terminology here. The parameters of the prior distribution, μ and τ here and a and b in the beta prior example above, are referred to as **hypterparameters**. The idea here is that they are still, in a sense, parameters of our model (since your choice of prior directly effects your output estimator). However you can think of them more as *tuning parameters*, values which live "above" the world of standard parameters, and which are not (usually) estimated.

Given a normal prior distribution, our posterior is then:

$$P[\theta|y] \propto \left[\frac{-1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2}\right)\right]$$

It is common to then expand the exponents, collect up like terms, and complete the square in θ to obtain:

$$P[\theta|y] \propto \left[\frac{-1}{2\tau'^2}(y-\mu')^2\right]$$

Where μ' and τ'^2 are the posterior mean and variance, given:

$$\begin{aligned} E_{\theta|y}[\theta] = \mu' &= \frac{\frac{1}{\tau^2}\mu + \frac{1}{\sigma^2}y}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} \\ \text{Var}_{\theta|y}[\theta] = \tau'^2 &= \frac{1}{\tau^2} + \frac{1}{\sigma^2} \end{aligned}$$

Does the expression for the posterior mean here remind you of anything??

Notice that the posterior mean is a *compromise* between the prior mean and the observed data, weighted by the inverse variances of each. When the prior is *diffuse* (ie. has high variance, meaning weakly informative) then the model favors the data. If the prior is narrow (ie. strongly informative) then the model favors the prior information.

Less Simple Case: Known Variance, Multiple Observations

Consider the case where we have n observations y_i from the distribution $N(\theta;\sigma^2)$, where still σ is known. We will retain our normal prior distribution from above, with mean μ and variance τ^2.

It turns out that adding more observations doesn't change this very much:

$$\begin{aligned} P[\theta|y] &\propto P[y_1, ..., y_n|\theta]P[\theta] \\ &\propto P[\theta]\prod_i^n P[y_i|\theta] \\ &\propto \exp\left[\frac{-1}{2\tau^2}(\theta-\mu)^2\right]\prod_i^n \exp\left[\frac{-1}{2\sigma^2}(y_i-\theta)^2\right] \\ &\propto \exp\left[\frac{-1}{2\tau^2}(\theta-\mu)^2 + \sum_i^n \frac{-1}{2\sigma^2}(y_i-\theta)^2\right] \end{aligned}$$

As before, we can apply (a lot of) algebra to simplify this down some. Doing so gives us:

$$\begin{aligned} E_{\theta|y}[\theta] = \mu' &= \frac{\frac{1}{\tau^2}\mu + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \\ \text{Var}_{\theta|y}[\theta] = \tau'^2 &= \frac{1}{\tau^2} + \frac{n}{\sigma^2} \end{aligned}$$

Where $\bar{y}$ is the sample mean. Again, observe that the posterior mean is a *compromise* between the prior mean and the sample mean, weighted by the variances of each and by the sample size. **Again: does this formula remind you of anything?**

Bayesian "Hypothesis Testing"

The kind of statistics that we have been working with so far, frequentist statistics, uses hypothesis testing because it assumes that model parameters cannot be *random*. In the frequentist understanding, the β in $y_i = \beta_0 + \beta_1 x_i$ are fixed quantities. Only the *estimates* $\hat{\beta}$ are random, and so hypothesis testing is built around the random estimates. We *cannot* make statements of the form "There is a 99% probability that β_1 is in the range $[a, b]$", so instead we construct confidence intervals which will contain the true value 99% of the time and use those.

Because Bayesian statistics also allows *subjective* probability, it has no problem with statements of the form "There is a 99% probability that β_1 is in the range $[a, b]$", so long as we understand that "99% probability" refers to our subjective beliefs, rather than a long run stable frequency. We are saying that, to us, we would charge someone .99$ to gamble that β_1 is in the range $[a, b]$. Intervals such as $[a, b]$ are known as **credible intervals**, and they are the Bayesian version of *confidence* intervals.

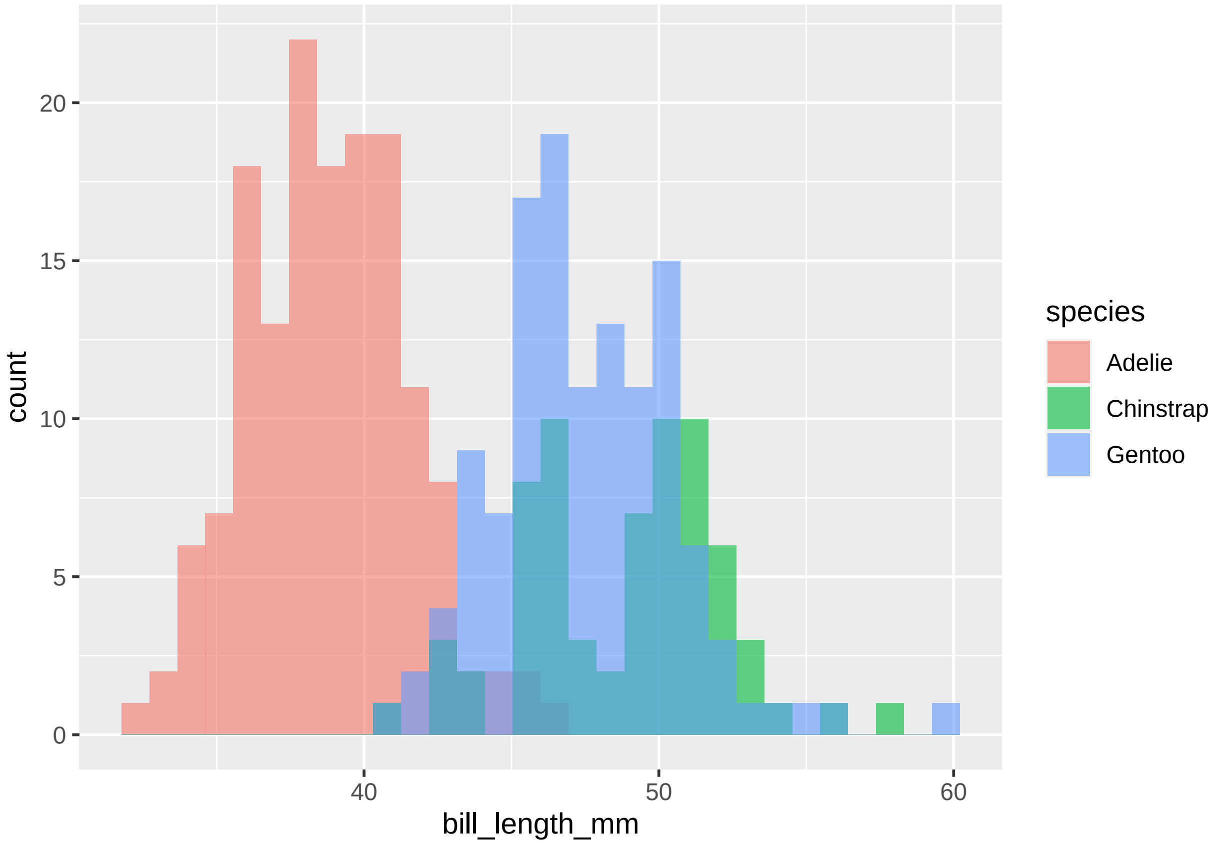
Frequentist	Bayesian
MLE	Posterior Mean
Std. Error	Posterior St. Dev.
Confidence Interval	Credible Interval

Credible Intervals Example - Penguins

Do Adelie penguins have shorter bills, on average, than other types of penguins? Let's find out using credible intervals.

```r
library(palmerpenguins)
library(magrittr)
library(tidyverse)

ggplot(penguins, aes(x=bill_length_mm,fill=species)) +
  geom_histogram(position='identity',alpha=.6)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



We'll perform our credible interval "test" in three steps:

1. Compute the mean bill length of non-Adelie penguins θ_{others}
2. Select a prior over the Adelie's mean bill length and compute our posterior distribution
3. Using the above posterior, compute th 95% credible interval, and see if it contains θ_{others}

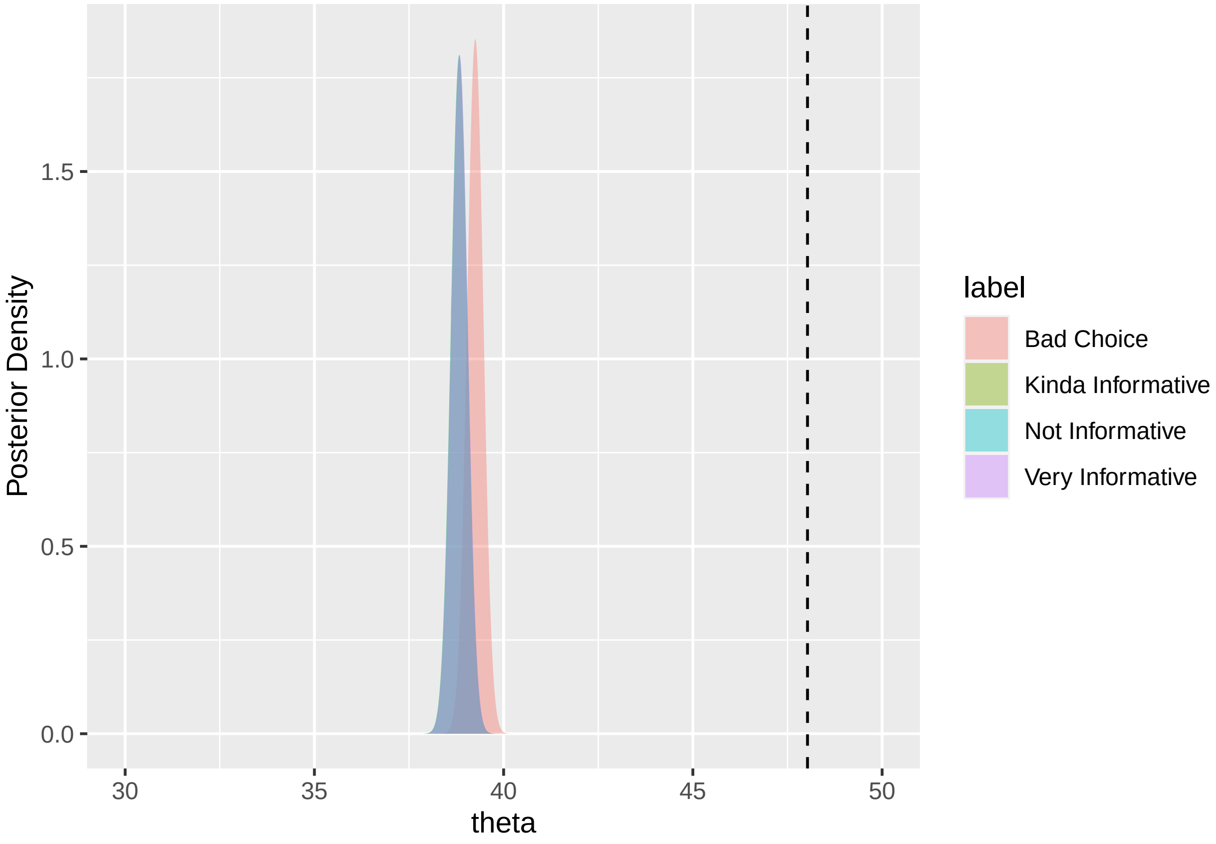
First: Step 1. Let's partition up the data, and compute mean bill length for the non-Adelies:

```
adelie = penguins %>% filter(species=='Adelie') %>% drop_na
others = penguins %>% filter(species!='Adelie') %>% drop_na

others.mn = mean(others$bill_length_mm)
```

For Step 2 we'll first need to select a prior. We'll use a normal prior with mean 45mm. Usually it's a good idea to pick a handful of priors representing different levels of informativeness, and check whether our choice has any actual effect on our results. If we need to use the most informative prior to get results, then our results aren't really caused by the data, they're just our prior beliefs.

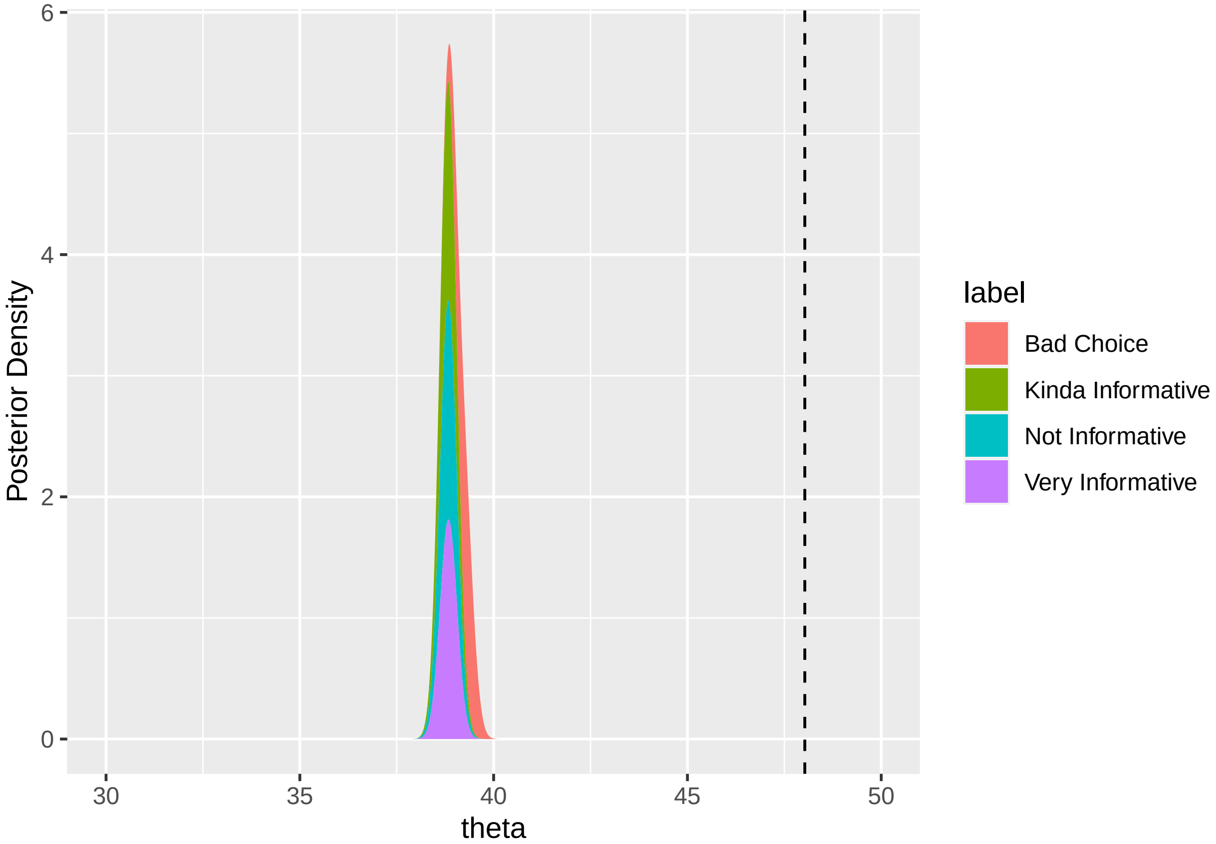
```
h=.01
theta.grid = seq(30,50,h)

log.prior1 = dnorm( theta.grid, mean=45, sd= 5, log=TRUE) # tau=25, very informative
log.prior2 = dnorm( theta.grid, mean=45, sd= 10, log=TRUE) # tau=100, somewhat informative
log.prior3 = dnorm( theta.grid, mean=45, sd= 50, log=TRUE) # tau=2500, not at all informative
log.prior.bad = dnorm( theta.grid, mean=others.mn, sd= 1, log=TRUE) # mu=others.mn tau=1, VERY informat


adelie.sd = adelie$bill_length_mm %>% sd # we'll assume that sigma is known for this example, and just
```

```
# This block will iterate through the values of theta we are considering. For each value of theta it wi
log.like = sapply(theta.grid,
                  function(theta){
                    sum(
                      dnorm(adelie$bill_length_mm, mean=theta, sd=adelie.sd,log=TRUE)
                      )
                    }
                  )

# We don't need to apply the above formula to compute the posterior, we can just multiply the gridded p
normalize = function(x){ return(x/sum(x*h)) } # here sum(x*h) is an application of rieman sums to appro
post1 = exp(log.like + log.prior1) %>% normalize
post2 = exp(log.like + log.prior2) %>% normalize
post3 = exp(log.like + log.prior2) %>% normalize
post.bad = exp(log.like + log.prior.bad) %>% normalize


plt.df = data.frame('theta'=theta.grid,
                    'post'=c(post1,post2,post3,post.bad),
                    'label'=rep(c('Very Informative','Kinda Informative','Not Informative','Bad Choice')
                                each=length(theta.grid))
                    )

ggplot(plt.df, aes(x=theta,y=post,fill=label)) +
  geom_area(position='identity',alpha=.4) +
  geom_vline(xintercept=others.mn,linetype='dashed') +
  labs(y='Posterior Density')
```



It's a little hard to see, but our choice of prior has no effect whatsoever on the posterior (except the "bad" prior). We can make this more visible by "stacking" the densities, rather than overlaying them:

```
ggplot(plt.df, aes(x=theta,y=post,fill=label)) +
  geom_area() +
  geom_vline(xintercept=others.mn,linetype='dashed') +
  labs(y='Posterior Density')
```



This makes sense, as the sample size of $n = 146$ is more than enough to accurately estimate θ.

Having computed the posterior density, we now just need to compute our credible interval. There are actually a few different ways to compute a P% confidence interval, although in practice the choice will not often matter:

1. *Mean-centered:* The interval $I = [E_{\theta|y}[\theta] - w, E_{\theta|y}[\theta] + w]$ where the width w is chosen so that $P[\theta \in I] = P$
2. *Quantile:* The interval $I = [Q_\alpha, Q_{1-\alpha}]$ where $\alpha = (1 - P)/2$ and Q_p is the p-th quantile of the posterior distribution
3. *Highest Posterior Density:* (HPD) the *smallest* interval I that contains P% of the posterior density

In the case of the normal distributed posterior, these are all the same. To make the code cleaner I'll use option 2 here:

```
P = .95
alpha = (1-P)/2
post.cdf = cumsum(h*post1)

ci.lower = theta.grid[ which(post.cdf >= alpha) %>% min ]
ci.upper = theta.grid[ which(post.cdf >= (1-alpha)) %>% min ]
```

```r
print('The 95% credible interval is:')
```

```
## [1] "The 95% credible interval is:"
```

```r
print(c(ci.lower,ci.upper))
```

```
## [1] 38.40 39.27
```

```r
print('The non-Adelie mean is:')
```

```
## [1] "The non-Adelie mean is:"
```

```r
print(others.mn)
```

```
## [1] 48.02834
```

Since this credible intervals *does not* contain the mean of the non-Adelie bill lengths, we can conclude that these species have different bills.