Lecture 3

Principles of Bayesian inference

WILD6900 (Spring 2019)

Loading [MathJax]/jax/output/CommonHTML/jax.js

Readings

> Hobbs & Hooten

From probability to Bayes theorem

Our goal as ecologists is to understand processes that we **cannot directly observe** based on quantities that we can observe

From probability to Bayes theorem

Our goal as ecologists is to understand processes that we **cannot directly observe** based on quantities that we can observe

We will refer to the unobserved processes as θ

From probability to Bayes theorem

Our goal as ecologists is to understand processes that we **cannot directly observe** based on quantities that we can observe

We will refer to the unobserved processes as θ

θ can include parameters of our models or latent states

- population size
- occupancy status
- alive/dead state of individuals

From probability to Bayes theorem

Our goal as ecologists is to understand processes that we **cannot directly observe** based on quantities that we can observe

We will refer to the unobserved processes as θ

θ can include parameters of our models or latent states

- population size
- occupancy status
- alive/dead state of individuals

Each of these unobserved processes is governed by a probability distribution $[\theta]$

From probability to Bayes theorem

To learn about θ, we take observations y

From probability to Bayes theorem

To learn about θ, we take observations y

Before those data are collected, they are random variables - the probability of observing y conditional on θ is governed by a probability distribution $[y|\theta]$

From probability to Bayes theorem

To learn about θ, we take observations y

Before those data are collected, they are random variables - the probability of observing y conditional on θ is governed by a probability distribution $[y|\theta]$

We want to know the probability distribution of the unobserved θ conditional on the observed data y, that is $[\theta|y]$

From probability to Bayes theorem

We know from last week that: [1]

$$[\theta | y] = \frac{[\theta, y]}{[y]}$$

and [2]

$$[\theta, y] = [y | \theta][\theta]$$

Through substitution, we get **Bayes theorem** [3]:

$$[\theta | y] = \frac{[y | \theta][\theta]}{[y]}$$

From probability to Bayes theorem

To understand what Bayes theorem says and why it is such a powerful principle, let's break down each part of equation 1:

$$\underbrace{[\theta|y]}_{\textit{posterior distribution}} = \frac{\overbrace{[\theta|y]}^{\textit{likelihood}} \quad \overbrace{[\theta]}^{\textit{prior}}}{\underbrace{[y]}_{\textit{marginal distribution}}}$$

The likelihood distribution

Likelihood

The concept of likelihood may be familiar to you from previous statistics classes because it is the central principle of frequentist statistics

Likelihood

The concept of likelihood may be familiar to you from previous statistics classes because it is the central principle of frequentist statistics

The likelihood allows us to answer the question:

> what is the probability that we will observe the data if our deterministic model $g(\theta)$ is the true process that gives rise to the data?

Likelihood

The concept of likelihood may be familiar to you from previous statistics classes because it is the central principle of frequentist statistics

The likelihood allows us to answer the question:

> what is the probability that we will observe the data if our deterministic model $g(\theta)$ is the true process that gives rise to the data?

That is, in likelihood, we treat θ as fixed and known rather than a random variable

- By assuming θ is fixed and known, we can calculate the probability density of our observation y conditional on θ

Likelihood

Example

Let's say we're sampling the number of trees on small plots and that we **know** [4] the average number of trees/plot is 40

Likelihood

Example

Let's say we're sampling the number of trees on small plots and that we **know** [4] the average number of trees/plot is 40

On one plot, we observe 34 trees. What is the probability of $y = 34$?

Likelihood

Example

Let's say we're sampling the number of trees on small plots and that we **know** [4] the average number of trees/plot is 40

On one plot, we observe 34 trees. What is the probability of $y = 34$?

To answer this question, we first need to select a sensible probability distribution for the number of trees on a plot

- Because these values have to be positive integers, the Poisson distribution is an obvious choice [5]

Likelihood

Example

Let's say we're sampling the number of trees on small plots and that we **know** [4] the average number of trees/plot is 40

On one plot, we observe 34 trees. What is the probability of $y = 34$?

To answer this question, we first need to select a sensible probability distribution for the number of trees on a plot

- Because these values have to be positive integers, the Poisson distribution is an obvious choice [5]

Next, we calculate the probability $Pr(y = 34 | \lambda = 40)$:

```
dpois(x = 34, lambda = 40)
```

```
## [1] 0.04247
```

Likelihood

Example

On a second plot, we observed 42 trees

Likelihood

Example

On a second plot, we observed 42 trees

What is the probability of that observation?

- $Pr(y = 42 \,|\, \lambda = 40) =$ 0.058

Likelihood

Example

On a second plot, we observed 42 trees

What is the probability of that observation?

- $Pr(y = 42 \mid \lambda = 40) =$ 0.058

What is the joint probability of both observations?

Likelihood

Example

On a second plot, we observed 42 trees

What is the probability of that observation?

- $Pr(y = 42 \,|\, \lambda = 40) =$ 0.058

What is the joint probability of both observations?

Assuming the observations are independent, the joint probability (probability of $y = 34$ and $y = 42$) is the product of the individual probabilities:

- 0.04 x 0.06 = `0.00059`.

Likelihood

In this example, we start by assuming we know that $\lambda = 40$

Likelihood

In this example, we start by assuming we know that $\lambda = 40$

Of course that doesn't make much sense. In our research, we never know λ (or to be consistent with eq. 1, θ). We want to estimate λ using our data

Likelihood

In this example, we start by assuming we know that $\lambda = 40$

Of course that doesn't make much sense. In our research, we never know λ (or to be consistent with eq. 1, θ). We want to estimate λ using our data

We do this by using a likelihood function:

$$\underbrace{L(\theta|y)}_{likelihood\ function} = \underbrace{[y|\theta]}_{likelihood} = \prod_{i=1}^{n} [y_i|\theta]$$

Likelihood profile

An important distinction between the probability distribution $[y \mid \theta]$ and a likelihood function $L(\theta \mid y)$ is:

- In the probability distribution, we treat the parameter as fixed and the data as random
- In the likelihood function, we treat the data as fixed and the parameter as variable

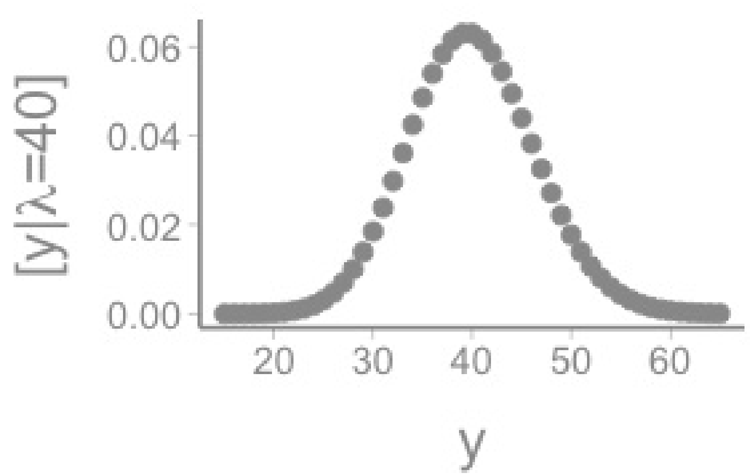
Likelihood profile

In our example, the tree counts y are random variables - they can take a range of possible values due to chance

The probability distribution $[y|\theta]$ tells us the probability of each possible value y:

```
y_probs <- data.frame(y = 15:65,
                      pr_y = dpois(15:65, lambda = 40))

ggplot(data = y_probs, aes(x = y, y = pr_y)) + geom_point() +
  scale_y_continuous(expression(paste("[y|", lambda, "=40]")))
```



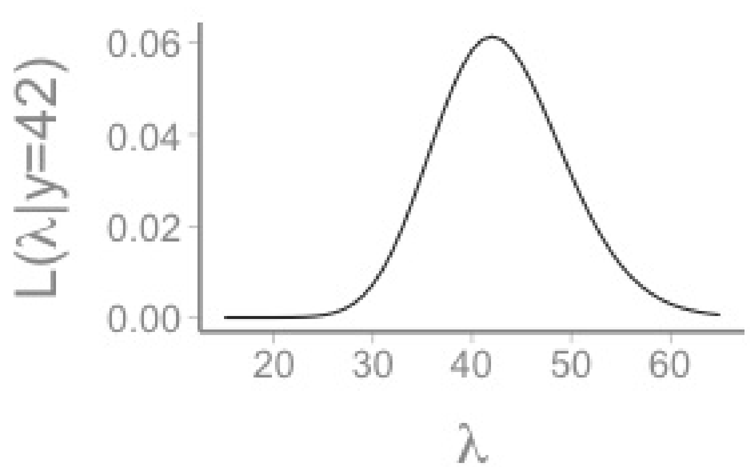
Likelihood profile

To create a likelihood profile, we flip this around

We treat our observation as fixed (for simplicity, let's use our observation $y = 42$) and estimate the probability as a function of different values of λ:

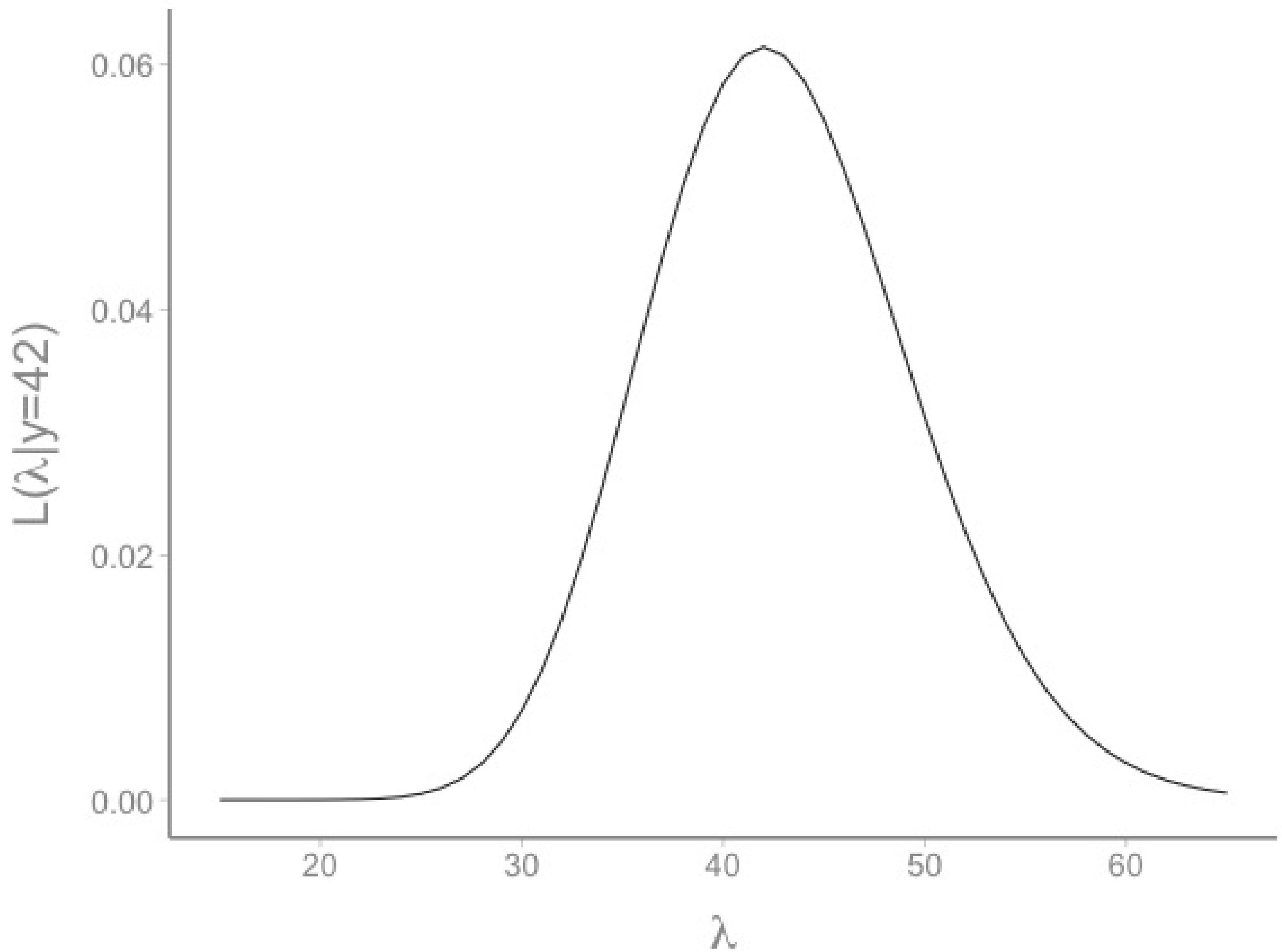
```
y <- 42
y_probs <- data.frame(lambda = 15:65,
                      pr_y = dpois(y, lambda = 15:65))

(lik_profile <- ggplot(data = y_probs, aes(x = lambda, y = pr_y)) + geom_pa
th() +
  scale_y_continuous(expression(paste("L(", lambda, "|y=42)"))) +
  scale_x_continuous(expression(lambda))
```



Likelihood profile

In this plot, the area under the curve does not equal 1 - the likelihood profile is not a probability distribution



Likelihood profile

Saying that λ is not fixed allows us to estimate the likelihood profile by varying the values of λ

But this is not the same as saying it's a random variable!

Likelihood profile

Saying that λ is not fixed allows us to estimate the likelihood profile by varying the values of λ

But this is not the same as saying it's a random variable!

For something to be a random variable, it must be defined by a probability distribution

- For the likelihood profile, we have not defined a probability distribution for λ (that is $[\lambda]$)

Likelihood profile

Saying that λ is not fixed allows us to estimate the likelihood profile by varying the values of λ

But this is not the same as saying it's a random variable!

For something to be a random variable, it must be defined by a probability distribution

- For the likelihood profile, we have not defined a probability distribution for λ (that is $[\lambda]$)

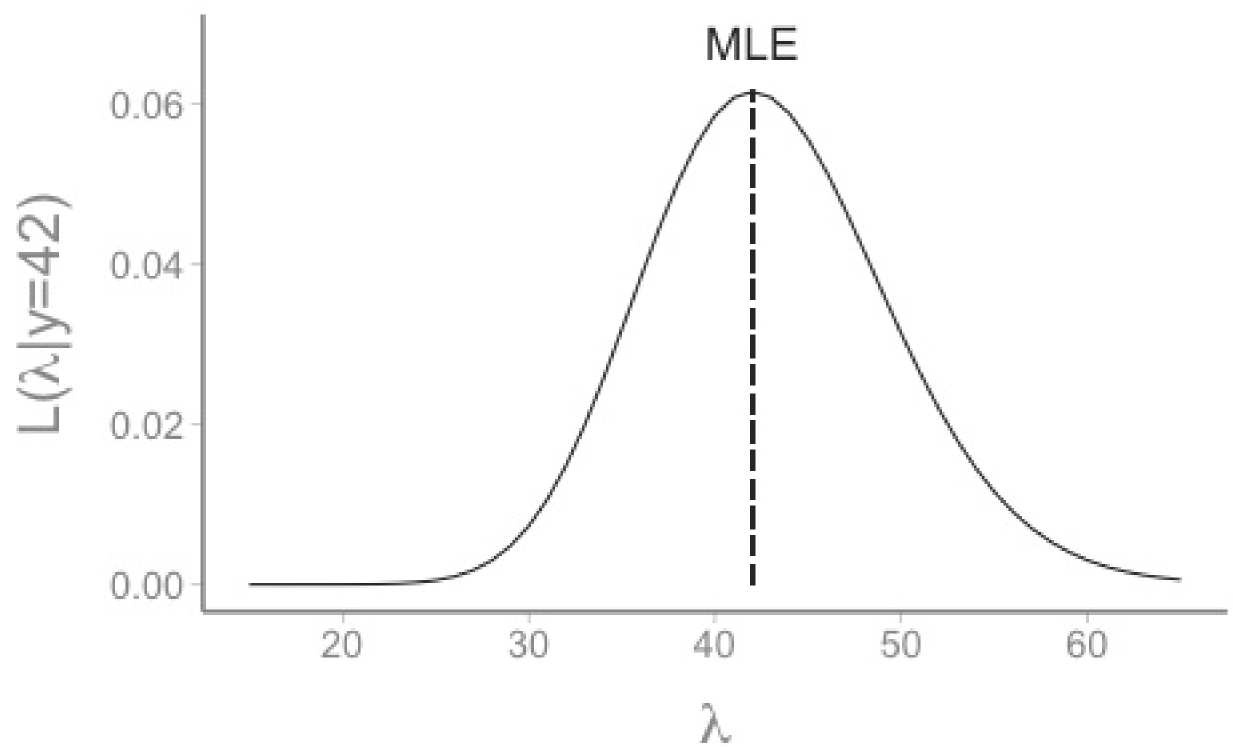
As a result, we vary λ but it is not a random variable and likelihood profiles do not define the probability or probability density of λ

Likelihood profile

This distinction between likelihood profiles and probability distributions is one of the reasons that results of likelihood-based methods can be difficult to interpret

Many of the methods familiar to ecologists use the principle of maximum likelihood to determine the value of θ that is most supported by our data

The maximum likelihood estimate is the peak of the likelihood curve [6]:



Likelihood profile

But the MLE does not tell us the probability of θ given our data!

So although MLE does tell us the value of θ that is most consistent with our data, we can not say things like:

- "There is a 90% probability that $\theta > 0$" [7]
- "There is a 96% probability that $a \geq \theta \geq b$" [7]

The prior [8]

The prior distribution

As we just learned, θ is not a random variable in the likelihood function because it is not governed by a probability distribution

The prior distribution

As we just learned, θ is not a random variable in the likelihood function because it is not governed by a probability distribution

$$\underbrace{[\theta|y]}_{posterior\ distribution} = \frac{\overbrace{[\theta|y]}^{likelihood} \quad \overbrace{[\theta]}^{prior}}{\underbrace{[y]}_{marginal\ distribution}}$$

In eq. 1, the **prior** distribution is what allows us to treat θ as a random variable

The prior distribution

As we just learned, θ is not a random variable in the likelihood function because it is not governed by a probability distribution

$$\underbrace{[\theta|y]}_{posterior\ distribution} = \frac{\overbrace{[\theta|y]}^{likelihood} \quad \overbrace{[\theta]}^{prior}}{\underbrace{[y]}_{marginal\ distribution}}$$

In eq. 1, the **prior** distribution is what allows us to treat θ as a random variable

- The prior describes what we know about the probability of θ before we collect any data

The prior distribution

As we just learned, θ is not a random variable in the likelihood function because it is not governed by a probability distribution

$$\underbrace{[\theta|y]}_{posterior\ distribution} = \frac{\overbrace{[\theta|y]}^{likelihood}\ \overbrace{[\theta]}^{prior}}{\underbrace{[y]}_{marginal\ distribution}}$$

In eq. 1, the **prior** distribution is what allows us to treat θ as a random variable

- The prior describes what we know about the probability of θ before we collect any data
- Priors can contain a lot of information about θ (informative priors) or very little (uninformative priors)

The prior distribution

As we just learned, θ is not a random variable in the likelihood function because it is not governed by a probability distribution

$$\underbrace{[\theta|y]}_{posterior\ distribution} = \frac{\overbrace{[\theta|y]}^{likelihood} \quad \overbrace{[\theta]}^{prior}}{\underbrace{[y]}_{marginal\ distribution}}$$

In eq. 1, the **prior** distribution is what allows us to treat θ as a random variable

- The prior describes what we know about the probability of θ before we collect any data
- Priors can contain a lot of information about θ (informative priors) or very little (uninformative priors)
- Well-constructed priors can also improve the behavior of our models

The prior distribution

The prior distribution provides us with a principled method of incorporating information about θ into our analysis

The prior distribution

The prior distribution provides us with a principled method of incorporating information about θ into our analysis

This information could be results from a pilot study or results from previously published studies

In some cases, the prior could simply reflect our own knowledge about how the system works

The prior distribution

The prior distribution provides us with a principled method of incorporating information about θ into our analysis

This information could be results from a pilot study or results from previously published studies

In some cases, the prior could simply reflect our own knowledge about how the system works

In this way, priors allow us to weigh conclusions drawn from our data against what we already know about our system [9]

The prior distribution

The prior distribution provides us with a principled method of incorporating information about θ into our analysis
This information could be results from a pilot study or results from previously published studies

In some cases, the prior could simply reflect our own knowledge about how the system works

In this way, priors allow us to weigh conclusions drawn from our data against what we already know about our system [9]
In the words of Mark Kéry:

> I find it hard not to be impressed by the application of Bayes rule to statistical inference since it so perfectly mimics the way of how we learn in everyday life ! In our guts, we always weigh any observation we make, or new information we get, with what we know to be the case or believe to know.

The prior distribution

Example

Say I tell you that on my way to class, I saw a 6-foot tall man

You would find this statement both believable and boring because a 6-ft tall man is consistent with what you know about the distribution of human heights

The prior distribution

Example

Say I tell you that on my way to class, I saw a 6-foot tall man

You would find this statement both believable and boring because a 6-ft tall man is consistent with what you know about the distribution of human heights

If I said I saw a 7-ft tall man, you might find this more noteworthy but believable (because your prior tells you this a possible, though rare, occurrence)

The prior distribution

Example

Say I tell you that on my way to class, I saw a 6-foot tall man

You would find this statement both believable and boring because a 6-ft tall man is consistent with what you know about the distribution of human heights

If I said I saw a 7-ft tall man, you might find this more noteworthy but believable (because your prior tells you this a possible, though rare, occurrence)

If I tell you I saw an 8-fit tall man, you'll question my credibility and require additional evidence because you know it is extremely implausible for someone to be this tall

The prior distribution

Example

In our example of tree counts, we need to define a prior for λ, the average number of trees per plot

To start, we know that λ has to be a positive real number (though not necessarily an integer)

The prior distribution

Example

In our example of tree counts, we need to define a prior for λ, the average number of trees per plot

To start, we know that λ has to be a positive real number (though not necessarily an integer)

- The gamma distribution allows for positive real values

The prior distribution

Example

In our example of tree counts, we need to define a prior for λ, the average number of trees per plot

To start, we know that λ has to be a positive real number (though not necessarily an integer)

- The gamma distribution allows for positive real values

In our discussion of likelihood functions, we assumed we know that $\lambda = 40$. Let's relax that assumption a bit

- previous research has shown that the mean number of trees per plot is 40, with a variance of 6

The prior distribution

Example

We can use moment matching to turn $\mu = 40$ and $\sigma^2 = 6$ into the two parameters that govern the gamma distribution:

$$\alpha = \frac{\mu^2}{\sigma^2}$$

$$\beta = \frac{\mu}{\sigma^2}$$

which in our sample gives $\alpha =$ 267 and $\beta =$ 7

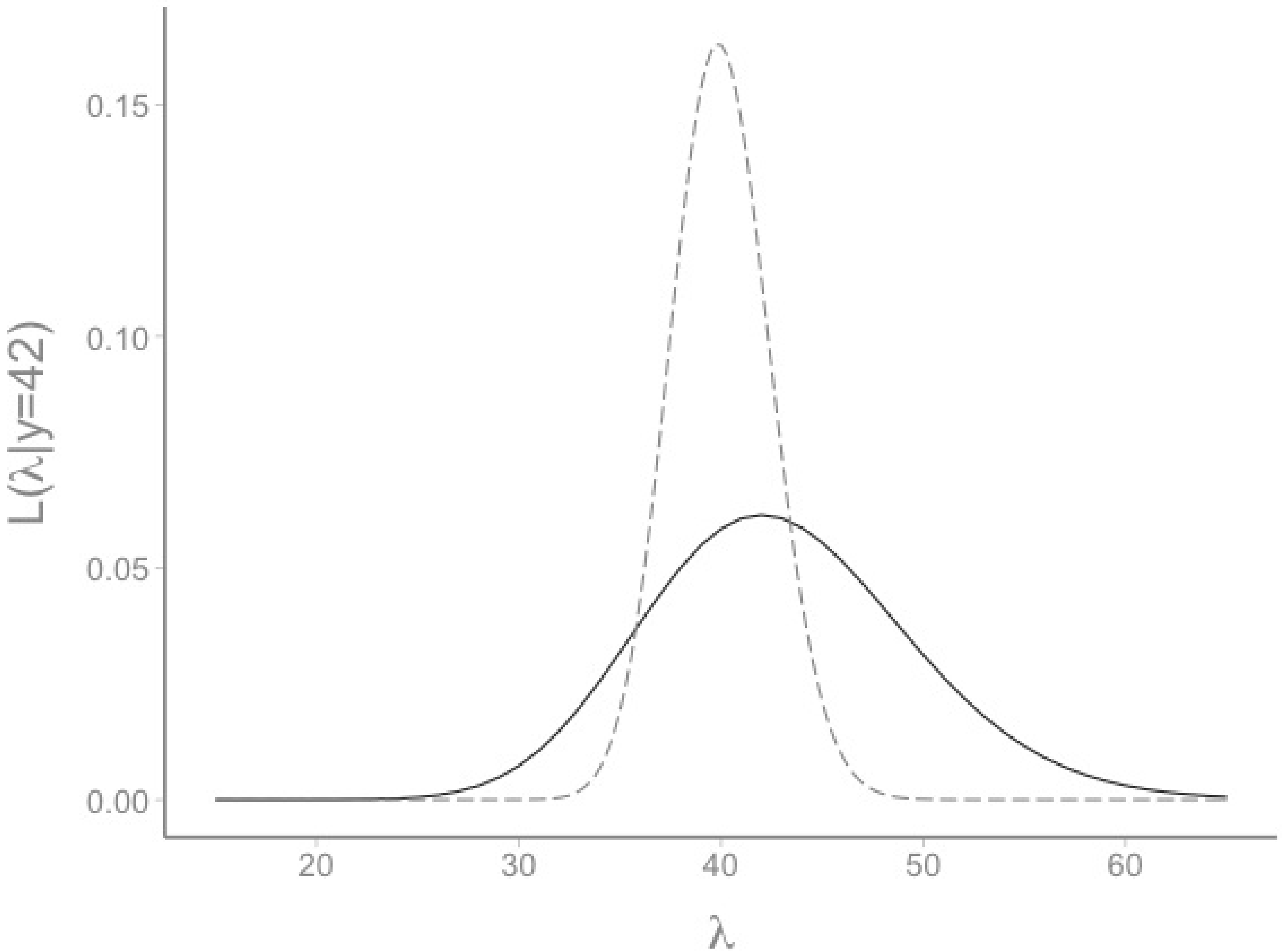
Now plot that prior alongside our previously define likelihood profile:

```
prior <- data.frame(lambda = seq(from = 15, to = 65, by = 0.25),
                    pr_lambda = dgamma(seq(from = 15, to = 65, by = 0.25),
40^2/6, 40/6))

(prior_lik <- lik_profile + geom_path(data = prior, aes(x = lambda, y = pr_
lambda), linetype = "longdash"))
```

The prior distribution

Example



The joint distribution

The joint distribution

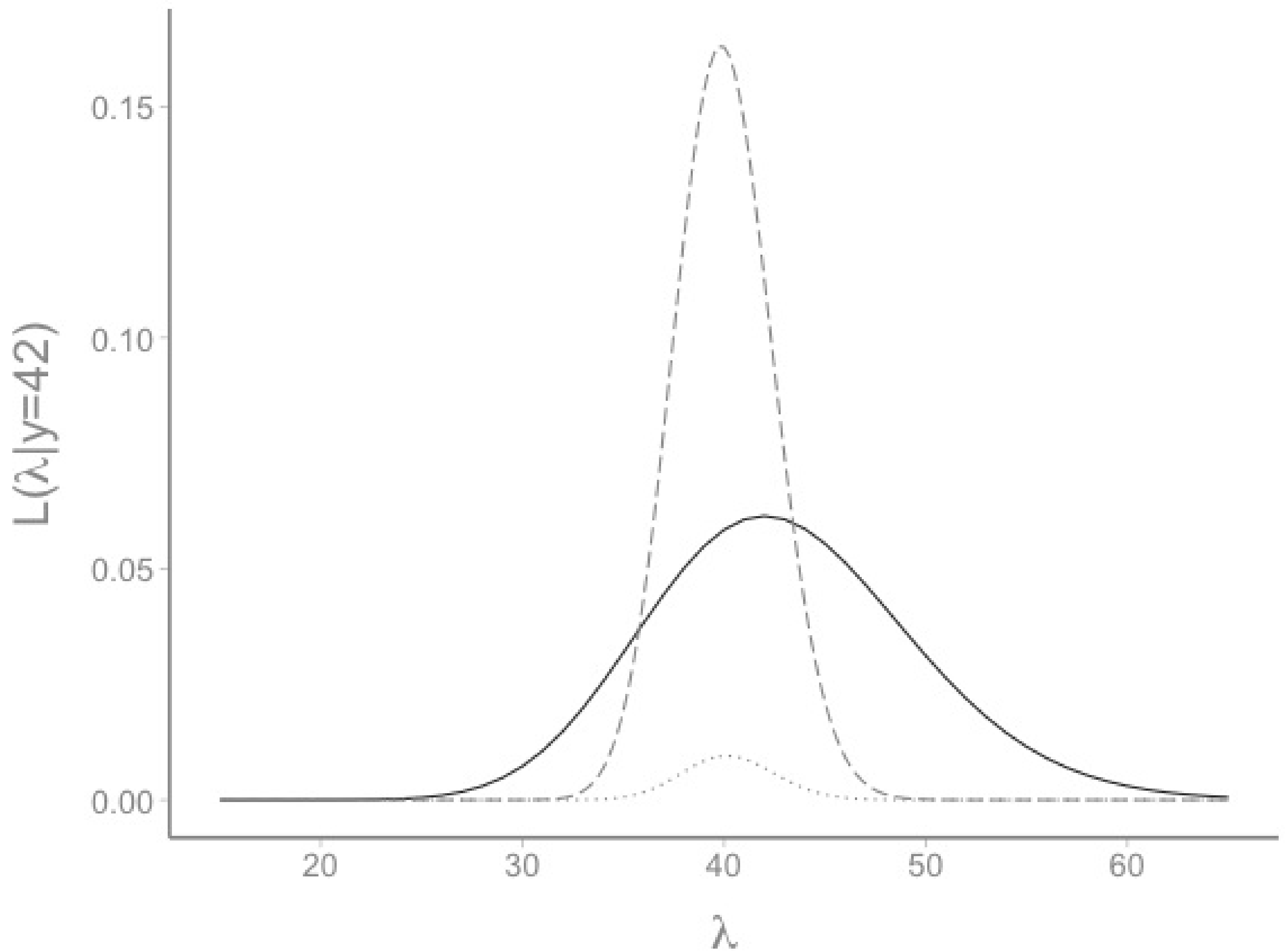
The product of the likelihood $[y \mid \theta]$ and the prior $[\theta]$ (the numerator of Bayes theorem) is called the **joint distribution**

It is important to note again that the joint distribution, like the likelihood profile, is not a probability distribution because the area under the curve does not sum to 1

```
joint <- data.frame(lambda = seq(from = 15, to = 65, by = 0.25),
                    jnt_dist = dgamma(seq(from = 15, to = 65, by = 0.25), 40
^2/6, 40/6) * dpois(42, seq(from = 15, to = 65, by = 0.25)))

(prior_lik_joint <- prior_lik + geom_path(data = joint, aes(x = lambda, y =
 jnt_dist), linetype = "dotted"))
```

The joint distribution



The marginal distribution

The marginal distribution

To convert the joint distribution into a true probability distribution, we have to divide it by the total area under the joint distribution curve

The marginal distribution

To convert the joint distribution into a true probability distribution, we have to divide it by the total area under the joint distribution curve

The denominator of eq. 1 $([y])$ is called the marginal distribution of the data - that is, the probability distribution of our data y across all possible values of θ

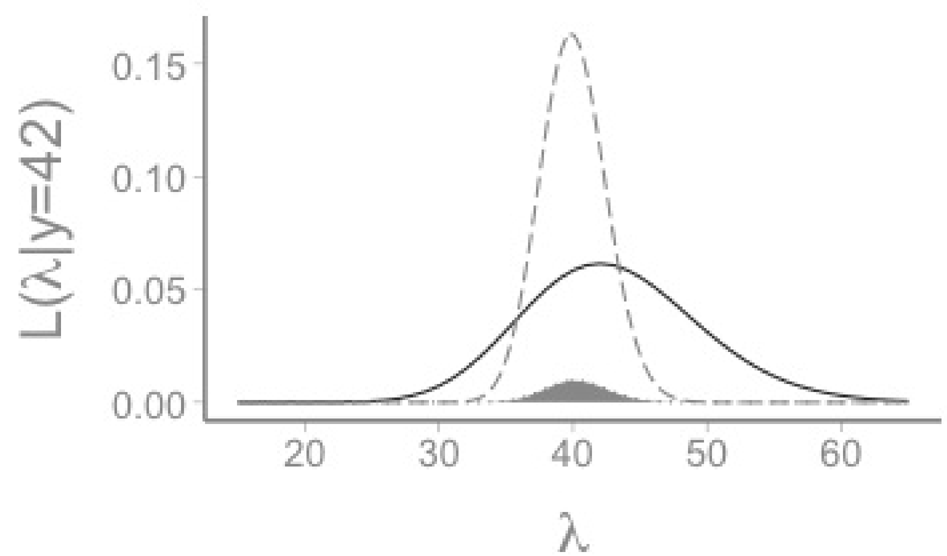
The marginal distribution

To convert the joint distribution into a true probability distribution, we have to divide it by the total area under the joint distribution curve

The denominator of eq. 1 ($[y]$) is called the marginal distribution of the data - that is, the probability distribution of our data y across all possible values of θ

Remember from our previous lecture that:

$$[y] = \int [y \mid \theta][\theta] d\theta$$



The marginal distribution

For some simple models, $[y]$ can be estimated analytically

The marginal distribution

For some simple models, $[y]$ can be estimated analytically

But in many cases, particularly in models with a moderate to large number of parameters, this is very hard to do [10]

The marginal distribution

For some simple models, $[y]$ can be estimated analytically

But in many cases, particularly in models with a moderate to large number of parameters, this is very hard to do [10]

For most of the models you will need to fit as an ecologist, estimating the marginal distribution of the data is one of a major challenges of Bayesian inference [11]

The posterior distribution

The posterior distribution

The LHS of equation, $[\theta | y]$, is known as the posterior distribution and it is what we want to know

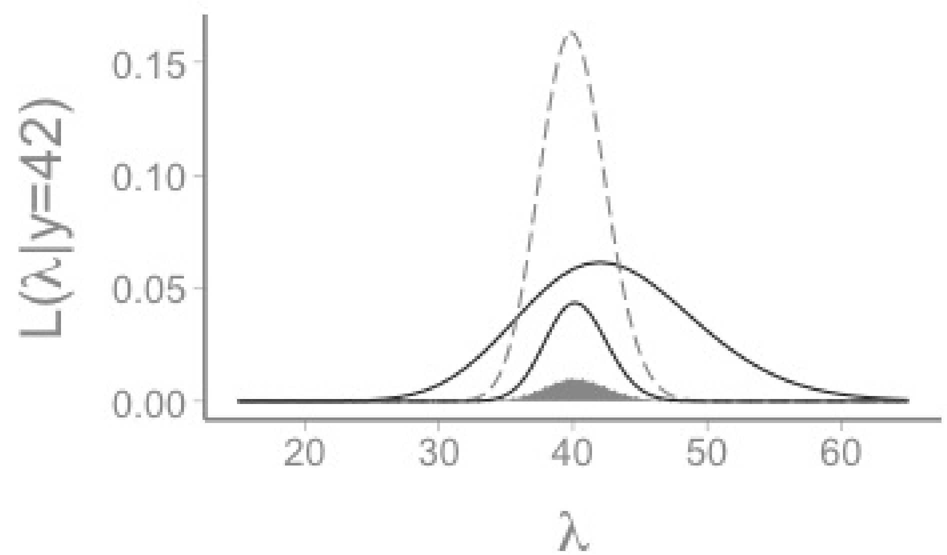
> What is the probability distribuiton of θ given our data?

The posterior distribution

The LHS of equation, $[\theta | y]$, is known as the posterior distribution and it is what we want to know

> What is the probability distribuiton of θ given our data?

The posterior distribution tells us everything we know about θ given our data (and possibly prior knowledge)



The posterior distribution

The posterior allows to make statements like:

- The most probable value of θ is x
- There is a 95% probability that $y < \theta < z$
- There is a 96% probability that $\theta < 0$

The posterior distribution

It's important to realize that before we collect data, y is a random variable governed by the marginal distribution

The posterior distribution

It's important to realize that before we collect data, y is a random variable governed by the marginal distribution

However, after we collect data, it is not longer random and $[y]$ becomes a fixed quantity

The posterior distribution

It's important to realize that before we collect data, y is a random variable governed by the marginal distribution

However, after we collect data, it is not longer random and $[y]$ becomes a fixed quantity

That means that:

$$[\theta|y] \propto [y|\theta][\theta]$$

The posterior distribution

In other words, because the denominator is a constant, the posterior distribution is proportional to the joint distribution

The posterior distribution

In other words, because the denominator is a constant, the posterior distribution is proportional to the joint distribution

This proportionality is important because it's easy to estimate the joint distribution (unlike the marginal distribution)

The posterior distribution

In other words, because the denominator is a constant, the posterior distribution is proportional to the joint distribution

This proportionality is important because it's easy to estimate the joint distribution (unlike the marginal distribution)

This means we can learn about the shape of the posterior distribution from the joint distribution even if we can't compute $[y]$

The posterior distribution

In other words, because the denominator is a constant, the posterior distribution is proportional to the joint distribution

This proportionality is important because it's easy to estimate the joint distribution (unlike the marginal distribution)

This means we can learn about the shape of the posterior distribution from the joint distribution even if we can't compute $[y]$

This is a central concept for applying modern tools for Bayesian analysis and one we will make use of shortly.

More about the posterior distribution

One of the cool things about Bayesian methods is that we don't get a point estimate of θ, we get an entire probability distribution! [12]

More about the posterior distribution

One of the cool things about Bayesian methods is that we don't get a point estimate of θ, we get an entire probability distribution! [12]

These advantages will be come clear as we move towards applications of these methods but as a quick example, let's say we are estimating the abundance of two populations (N_1 and N_2)

We want to determine whether $N_1 > N_2$

More about the posterior distribution

In a frequentist framework, it relatively straightforward to get point estimates of N_1 and N_2

More about the posterior distribution

In a frequentist framework, it relatively straightforward to get point estimates of N_1 and N_2

Saying that $N_1 > N_2$ is the same as saying $N_1 - N_2 > 0$ so to answer our question we could derive a new parameter $\Delta_N = N_1 - N_2$ and test whether $\Delta_N > 0$

More about the posterior distribution

In a frequentist framework, it relatively straightforward to get point estimates of N_1 and N_2

Saying that $N_1 > N_2$ is the same as saying $N_1 - N_2 > 0$ so to answer our question we could derive a new parameter $\Delta_N = N_1 - N_2$ and test whether $\Delta_N > 0$

Answering the question of whether $\Delta_n > 0$ requires knowing not only the magnitude of this difference but also how certain we are in the value. How do we estimate the uncertainty of our new derived variable?

More about the posterior distribution

In a frequentist framework, it relatively straightforward to get point estimates of N_1 and N_2

Saying that $N_1 > N_2$ is the same as saying $N_1 - N_2 > 0$ so to answer our question we could derive a new parameter $\Delta_N = N_1 - N_2$ and test whether $\Delta_N > 0$

Answering the question of whether $\Delta_n > 0$ requires knowing not only the magnitude of this difference but also how certain we are in the value. How do we estimate the uncertainty of our new derived variable?

That's not easy in a frequentist world and will require application of the delta method.

More about the posterior distribution

In a Bayesian world, we can actually estimate the entire posterior distribution of Δ_N!

More about the posterior distribution

In a Bayesian world, we can actually estimate the entire posterior distribution of Δ_N!

All of the uncertainty in N_1 and N_2 will propagate into our uncertainty about Δ_N

More about the posterior distribution

In a Bayesian world, we can actually estimate the entire posterior distribution of Δ_N!

All of the uncertainty in N_1 and N_2 will propagate into our uncertainty about Δ_N

It is trivially easy to estimate confidence intervals or specific probabilities for Δ_N (e.g., $Pr(\Delta_N > 0)$)

More about the posterior distribution

In a Bayesian world, we can actually estimate the entire posterior distribution of Δ_N!

All of the uncertainty in N_1 and N_2 will propagate into our uncertainty about Δ_N

It is trivially easy to estimate confidence intervals or specific probabilities for Δ_N (e.g., $Pr(\Delta_N > 0)$
)

If nothing else turns you into a Bayesian, it's probably this point.