Lecture 4: Principles of Bayesian inference

WILD6900

updated 2018-11-27

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. In the following section, we will refer to the observered processes as θ. θ can include parameters of our models or latent states (i.e., population size, occupancy status, alive/dead state of individuals). Each of these unobserved processes is governed by a probability distribution $[\theta]$.

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 $[\theta|y]$. We know from last week that:

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

and

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

Through substitution, we get *Bayes theorem*:

$$[\theta|y] = \frac{[y|\theta][\theta]}{[y]} \tag{1}$$

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

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

Likelihood

We'll start with the likelihood $[y|\theta]$. The concept of likelihod 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. For example, let's say we're sampling the number of on sampling plots and that we know 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. Remember

that the expected value of a Poisson distribution is governed by the parameter λ. In this case, we know that $\lambda = 40$. Next, we calculate the probability $Pr(y = 34|\lambda = 40)$. We learned last week that we can do this in `R` using `dpois(x = 34, lambda = 40)` which equals 0.042.

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* $y = 51$) is the product of the individual probabilities: 0.04 x 0.06 = `0.00059`.

In this example, we start by assuming we know that $\lambda = 40$. That, of course, 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 a probability distribution and a likelihood function is that, in a 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. This is perhaps best illustrated by example.

Let's say we want to plot the probability distribution of the number of trees on our study plots. The counts y are random variables - they can take a range of possible values due to change. We want to know the probability of each possible value y. To estimate this probability, we need to chose a probability distribution and its parameter(s). This case, we already decided to describe y as a Poisson distribution with $\lambda = 40$. To plot $Pr(y|\lambda)$, we simply estimate probabilities for different values of y and then plot them:

```
library(ggplot2)
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]"))) +
  theme_classic()
```

To reiterate, the plot above assumes we know $\lambda = 40$ and that the y's are random variables from a Poisson distribution. Because this is a probability distribution, the area under the curve is 1.

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_path() +
  scale_y_continuous(expression(paste("L(", lambda, "|y=42)"))) +
  scale_x_continuous(expression(lambda)) +
  theme_classic())
```

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

Saying that θ is not fixed allows us to estimate the likelihood profile by varying the values of θ. 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. Note that to estimate the likelihood profile, we have not defined a probability

distribution for θ (that is $[\theta]$). As a result, we vary θ but it is not a random variable and likelihood profiles do not define the probability or probability density of θ.

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:

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 $\theta > 0$" based on MLE methods" (even though that's usually what we want to know!).

The prior distribution

Bayes theorem gives us a way to move from the likelihood function to the probability distribution of θ (conditional on our data). As we just learned, θ is not a random variable in the likelihood function because it is not governed by a probability 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 (so called *informative priors*) or very little (*uninformative priors*). Well-constructed priors can also improve the behavior of our models, which we'll learn more about later. Choosing prior distributions is a complex topic that is an area of active research in the statistical community. As a result, the cultural norms for using priors in ecological modeling appears to be rapidly changing. For these reasons, we'll spend a good deal of time discussing how to choose priors in the next lecture.

One of the coolest aspects of Bayesian methods is that 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. This is a nice property because it is consistent with both the way that science advances (the accumulation of evidence in support of specific hypotheses) as well as how we naturally learn about the world around us. 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.

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, occurence). 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.

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 is a logical choice for our prior because it 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 but assuming previous research has shown that the mean number of trees per plot is 40, with a variance of 6. We can use moment matching to turn this estimate 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 $\alpha = 266.6667$ and $\beta = 6.6667$. Let's 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 joint distribution

The product of the likelihood $[y|\theta]$ 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

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

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 demoninator 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:

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

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. In fact, the difficulty of computing the marginal distribution is one of the reasons that it look a long time for Bayesian methods to be applied in practice (Bayes theorem was proven in 1763, long before the frequentist methods you are used to using were invented). 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. We will return to this topic later

The posterior distribution

The LHS of equation 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 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$$

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:

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

In other words, because the demoninator 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.

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! Although this may seem like a minor point right now, it has some really practical advantages, namely that we can easily quantify uncertainty in our parameter estimates and we can summarise the distribution is whatever way we want (mean, median, mode, 95% quantiles, 50% quantiles, etc.).

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 (we'll call them N_1 and N_2) and determining the probability that $N_1 > N_2$. In a frequentist framework, it relatively straightforward to get point estimates of N_1 and N_2 (assuming we have data that is adequate to estimate these state variables). Saying that $N_1 > N_2$ is the same as saying $N_1 - N_2 = \Delta_N > 0$ so to answer our question we could also estimate this difference as the difference between our point estimates. 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. We know there is uncertainty in N_1 and N_2 (and thus in the difference) but 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.

In a Bayesian world, we can actually estimate the entire posterior distribution of Δ_N, which makes it trivially easy to estimate confidence intervals or specific probabilities (e.g., $Pr(\Delta_N > 0)$). If nothing else turns you into a Bayesian, it's probably this point.

More about prior distributions

> As scientists, we should always prefer to use appropriate, well-constructed, informative priors on θ - Hobbs & Hooten

Selecting priors is one of the more confusing and often contentious topics in Bayesian analyses. It is often confusing because there is no direct counterpart in traditional frequentists statistical training. It is controversial because modifying the prior will potentially modify the posterior. Many scientists are uneasy with the idea that you have to *choose* the prior. This imparts some subjetivity into the modeling workflow, which conflicts with the idea that we should be "objective" and only base our conclusions on what our data tell us. But this view is both philosophically counter to the scientific method and ignores the many benefits of using priors that contain some information about the parameter(s) θ.

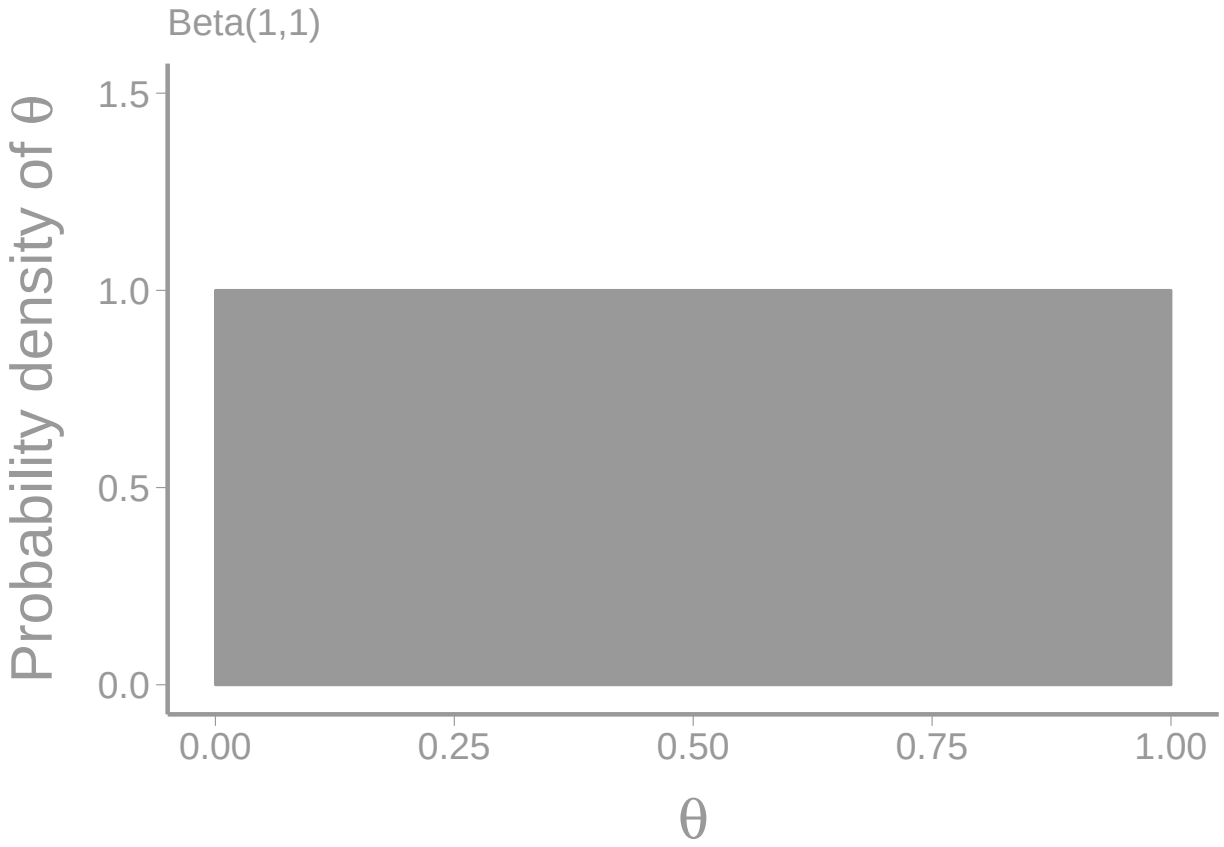
Best pratices for selecting priors is an area of active research in the statistical literature and advice in the ecological literature is changing rapidly. As a result, the following sections may be out-of-date in short order. Nonetheless, understanding how and why to contruct priors will greatly benefit your analyses so we need to spend some time on this topic.

Non-informative priors

In much of the ecological literature, including some influential textbooks, the standard method of selecting priors is to choose priors that have minimal impact on the posterior distribution, so called "non-informative priors". Using non-informative priors is intuitively appealing because, in theory, they let the data "do the talking". For simple models, non-informative priors also return posteriors that are generally consistent with frequentist-based analyses, which gives folks who are uneasy with the idea of using priors some comfort.

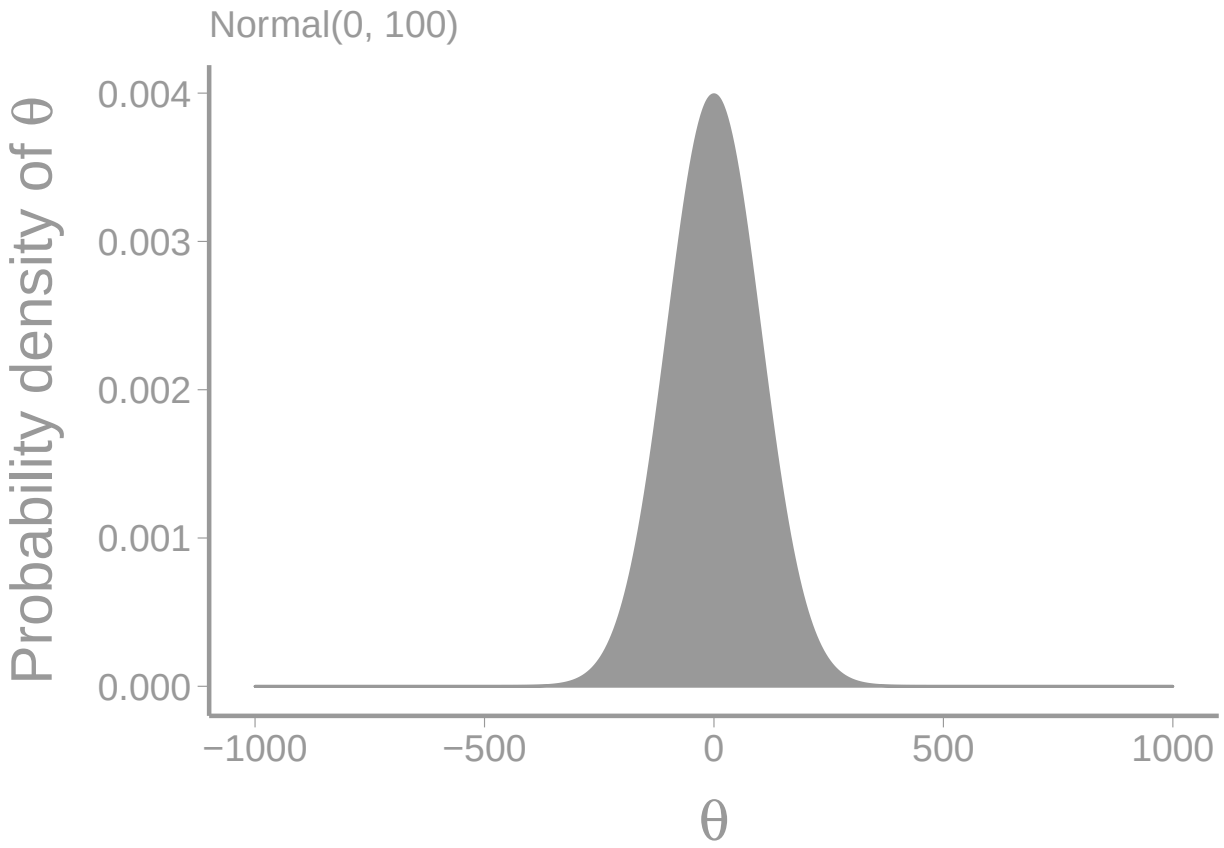
Non-informative priors generally try to be agnostic about the prior probability of θ. For example, in θ is a probability, $Uniform(0, 1)$ gives equal prior probability to all possible values. The $Uniform(0, 1)$ is a special case of the beta distribution with $\alpha = \beta = 1$:

```r
beta_df <- data.frame(x = seq(0,1,0.01),
                      y = dbeta(seq(0,1,0.01), 1, 1))

ggplot(beta_df, aes(x, y)) + geom_area(fill = WILD3810_colors$value[WILD3810_colors$name == "secondary"]
                                       color = WILD3810_colors$value[WILD3810_colors$name == "secondary"
  scale_x_continuous(expression(theta)) +
  scale_y_continuous(expression(paste("Probability density of ", theta)), limits = c(0, 1.5)) +
  labs(subtitle = "Beta(1,1)")
```



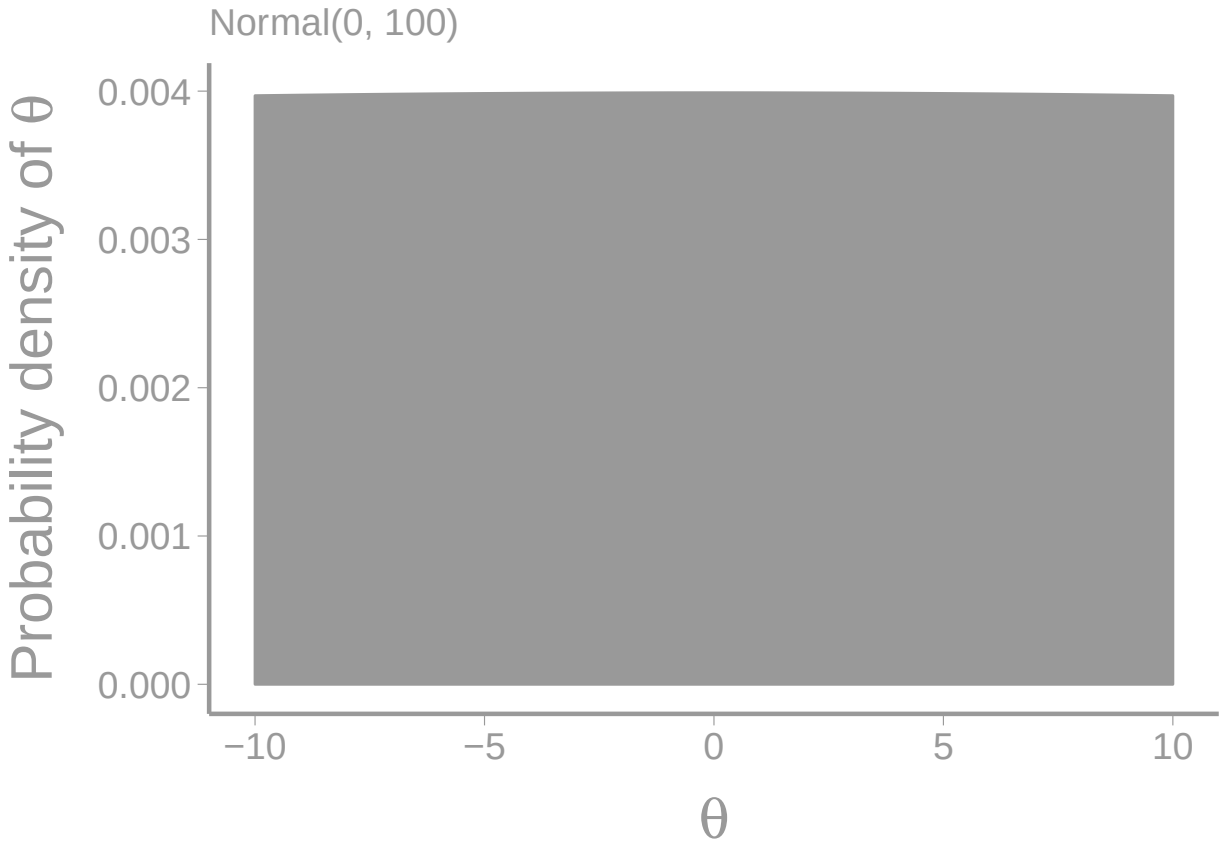
For a parameter that could be any real number, a common choice is a normal prior with very large variance:

```r
norm_df <- data.frame(x = seq(-1000,1000,1),
                      y = dnorm(seq(-1000,1000,1), 0, 100))

ggplot(norm_df, aes(x, y)) + geom_area(fill = WILD3810_colors$value[WILD3810_colors$name == "secondary"]
                                       color = WILD3810_colors$value[WILD3810_colors$name == "secondary"
  scale_x_continuous(expression(theta)) +
  scale_y_continuous(expression(paste("Probability density of ", theta))) +
  labs(subtitle = "Normal(0, 100)")
```



Over realistic values of θ, this distribution appears relatively flat:

```
norm_df <- data.frame(x = seq(-10,10,1),
                      y = dnorm(seq(-10,10,1), 0, 100))

ggplot(norm_df, aes(x, y)) + geom_area(fill = WILD3810_colors$value[WILD3810_colors$name == "secondary"]
                                       color = WILD3810_colors$value[WILD3810_colors$name == "secondary
  scale_x_continuous(expression(theta)) +
  scale_y_continuous(expression(paste("Probability density of ", theta))) +
  labs(subtitle = "Normal(0, 100)")
```



Often ecological models have parameters that can take real values > 0 (variance or standard deviation, for example). In these cases, uniform priors from 0 to some large number (e.g., 100) are often used or sometimes very diffuse gamma priors ($gamma(0.01, 0.01)$)

Non-informative priors are appealing because they appear to take the subjectivity out of our analysis (assuming we come up with the same vague priors!) and because they "let the data do the talking". However, non-informative priors are often not the best choice for pratcial and philosophical reasons.

Practical issues with non-informative priors

Non-informative priors are never "non-informative". The prior always has *some* influence on the posterior and in some cases can actually have quite a bit of influence. For example, let's say we track 10 individuals over a period of time using GPS collars and we want to know the probability that an individual survives from time t to time $t + 1$. Our data is a series of 0's (dead) and 1's (alive) and we will assume 3 individuals died during our study (so seven 1's and three 0's). A natural model for these data is the Bernoulli distribution with probability p. Because p is a parameter, it needs a prior and we can choose the $Beta(1, 1)$ distribution that we just saw gives equal prior probability to all values of p between 0 and 1. It turns out when we have Bernoulli likelihood and a beta prior, the posterior distribution will also be a beta distribution. This scenario, when the prior and the posterior have the same distribution, occurs when the likelihood and prior are *conjugate* distributions.

```
conjugates <- data.frame(Likelihood = c("$y_i \\sim binomial(n, p)$",
                                        "$y_i \\sim Bernoulli(p)$",
                                        "$y_i \\sim Poisson(\\lambda)$"),
                         Prior = c(c("$p \\sim beta(\\alpha, \\beta)$",
                                     "$p \\sim beta(\\alpha, \\beta)$",
                                     "$\\lambda \\sim gamma(\\alpha, \\beta)$")),
```

Table 1: A few conjugate distributions

Likelihood	Prior	Posterior
$y_i \sim binomial(n, p)$	$p \sim beta(\alpha, \beta)$	$p \sim beta(\sum y_i + \alpha, n - \sum y_i + \beta)$
$y_i \sim Bernoulli(p)$	$p \sim beta(\alpha, \beta)$	$p \sim beta(\sum_{i=1}^{n} y_i + \alpha, \sum_{i=1}^{n}(1 - y_i) + \beta)$
$y_i \sim Poisson(\lambda)$	$\lambda \sim gamma(\alpha, \beta)$	$\lambda \sim gamma(\alpha \sum_{i=1}^{n} y_i, \beta + n)$

```
                          Posterior = c("$p \\sim beta(\\sum y_i + \\alpha, n -\\sum y_i + \\beta)$",
                                        "$p \\sim beta(\\sum_{i=1}^n y_i + \\alpha, \\sum_{i=1}^n (1-y_
                                        "$\\lambda \\sim gamma(\\alpha \\sum_{i=1}^n y_i, \\beta + n)$"

kableExtra::kable(conjugates, "latex", align="c", booktabs=TRUE, escape = F, caption = 'A few conjugate
```

Conjugate distributions are useful because we can estimate the posterior distribution algebracically