Lecture 4

More about prior distributions

WILD6900 (Spring 2019)

Readings

> Hobbs & Hooten 90-105

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

More about prior distributions

Selecting priors is one of the more confusing and often contentious topics in Bayesian analyses

More about prior distributions

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

More about prior distributions

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

More about prior distributions

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

More about prior distributions

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 subjectivity into the modeling workflow, which conflicts with the idea that we should only base our conclusions on what our data tell us

More about prior distributions

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 subjectivity into the modeling workflow, which conflicts with the idea that we should 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

A note on this material

Best practices 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 construct priors will greatly benefit your analyses so we need to spend some time on this topic.

Non-informative priors

Non-informative priors

In much of the ecological literature the standard method of selecting priors is to choose priors that have minimal impact on the posterior distribution, so called **non-informative priors**

Non-informative priors

In much of the ecological literature 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:

Non-informative priors

In much of the ecological literature 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:

- they let the data "do the talking"

Non-informative priors

In much of the ecological literature 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:

- they let the data "do the talking"

- they return posteriors that are generally consistent with frequentist-based analyses \(^1\)

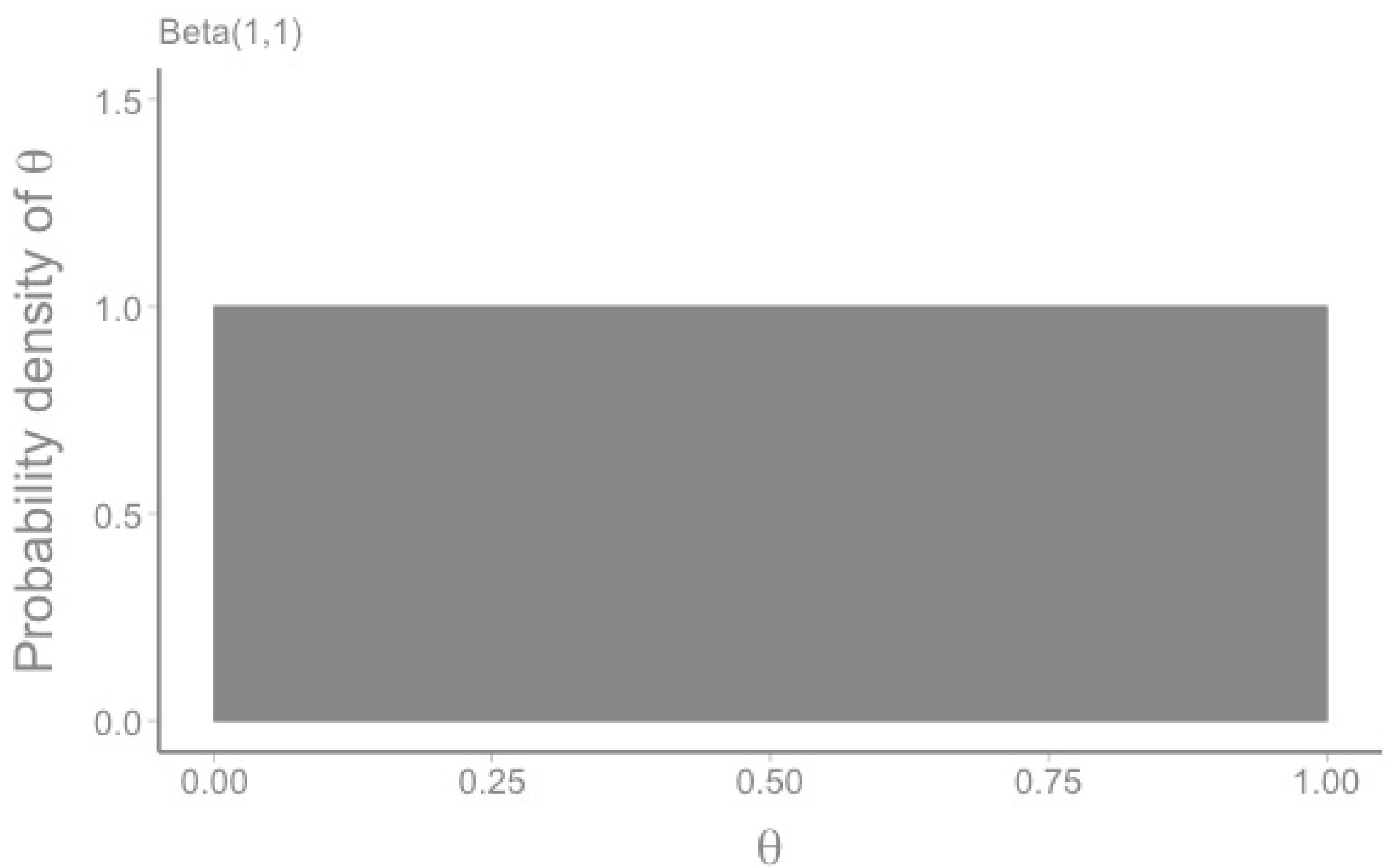
Non-informative priors

Non-informative priors generally try to be agnostic about the prior probability of \(\large \theta\)

Non-informative priors

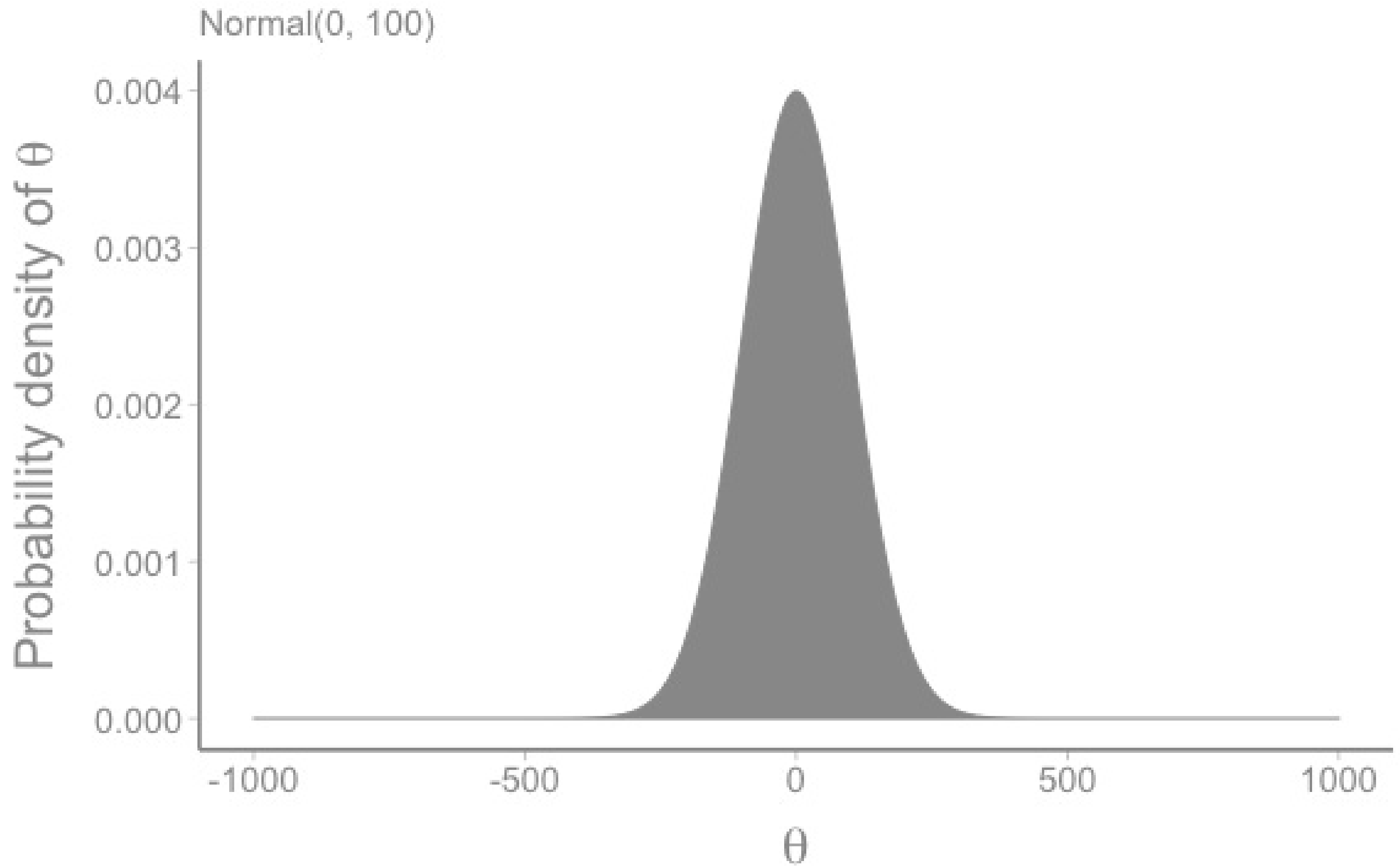
Non-informative priors generally try to be agnostic about the prior probability of \(\large \theta\)

For example, if \(\theta\) is a probability, \(Uniform(0,1)\) gives equal prior probability to all possible values \(^2\):



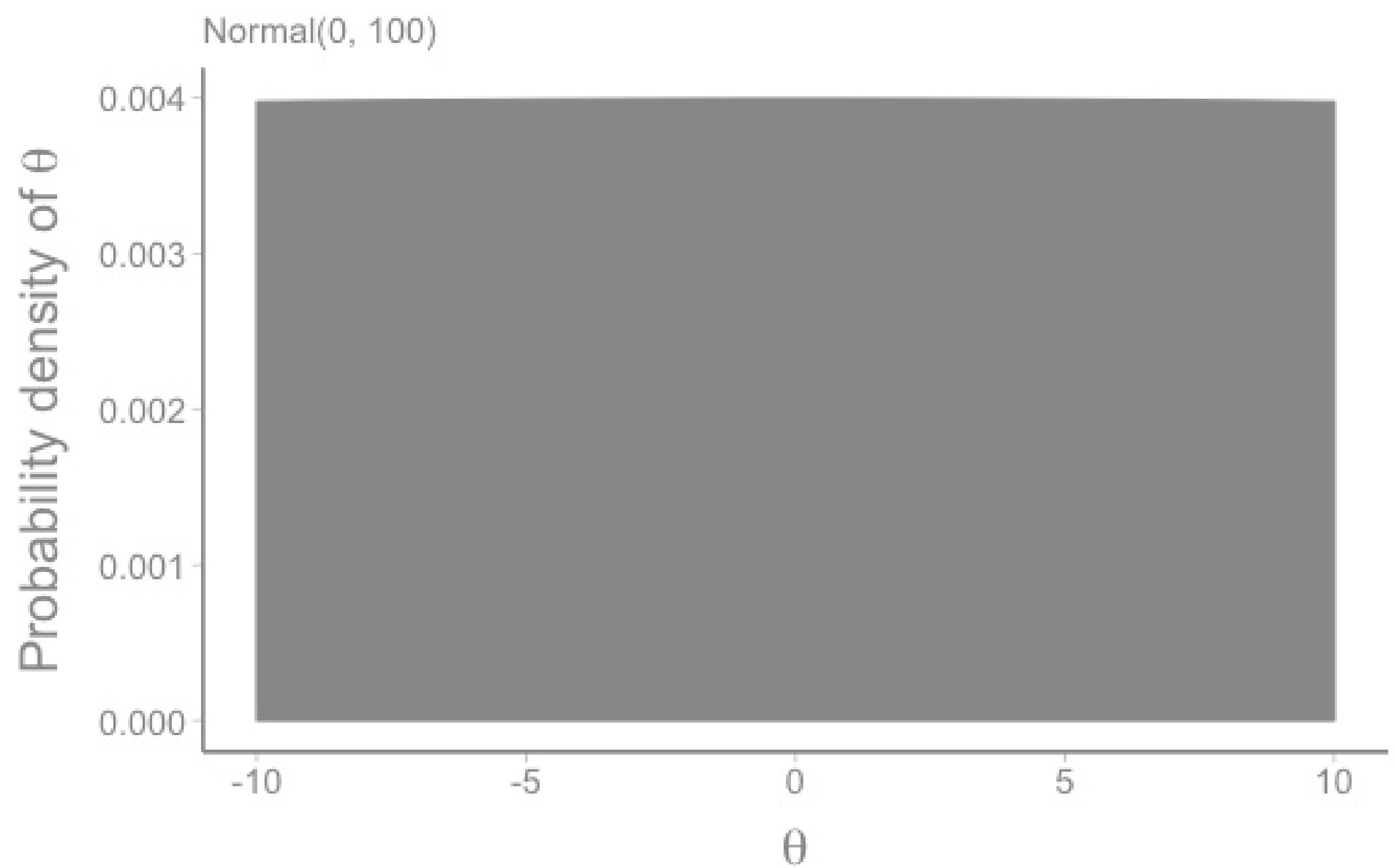
Non-informative priors

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



Non-informative priors

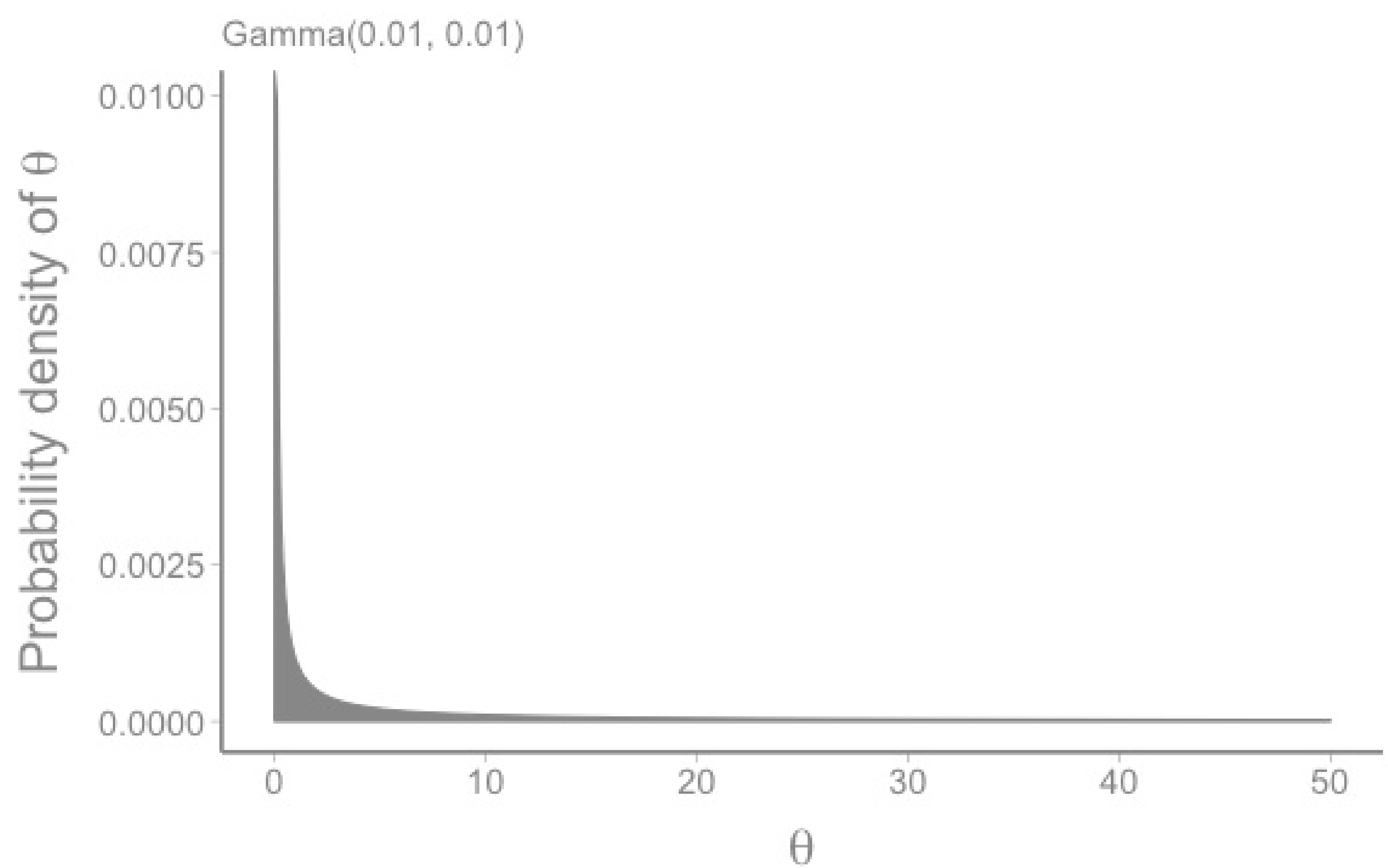
Over realistic values of \(\theta\), this distribution appears relatively flat:



Non-informative priors

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, e.g. $gamma(0.01, 0.01)$



Non-informative priors

Non-informative priors are appealing because they appear to take the subjectivity out of our analysis \(^1\)

- they "let the data do the talking"

Non-informative priors

Non-informative priors are appealing because they appear to take the subjectivity out of our analysis [1]

- they "let the data do the talking"

However, non-informative priors are often not the best choice for practical and philosophical reasons

Practical issues with non-informative priors

The prior always has some influence on the posterior and in some cases can actually have quite a bit of influence

Practical issues with non-informative priors

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 $t+1$ [1].

- assume 8 individuals died during our study (so two 1's and eight 0's)

Practical issues with non-informative priors

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 \(t+1\) \(^1\).

- assume 8 individuals died during our study (so two 1's and eight 0's)

A natural model for these data is

$$Binomial(p, N = 10)$$

Practical issues with non-informative priors

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 $t+1$ [1].

- assume 8 individuals died during our study (so two 1's and eight 0's)

A natural model for these data is

$$Binomial(p, N = 10)$$

Because p is a parameter, it needs a prior:

- $Beta(1,1)$ gives equal prior probability to all values of p between 0 and 1

Conjugate priors

It turns out when we have binomial likelihood and a beta prior, the posterior distribution will also be a beta distribution

Conjugate priors

It turns out when we have binomial 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.

Conjugate priors

It turns out when we have binomial 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.

A few conjugate distributions

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

Conjugate priors

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

Conjugate priors

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

In our case, the posterior distribution for \(p\) is:

$$\Large beta(2 + 1, 8 + 1) = beta(3, 9)$$

Practical issues with non-informative priors

No prior is truly non-informative

Practical issues with non-informative priors

No prior is truly non-informative

As we collect more data, this point will matter less and less

- if we tracked 100 individuals and 20 lived, our posterior would be $beta(21, 81)$

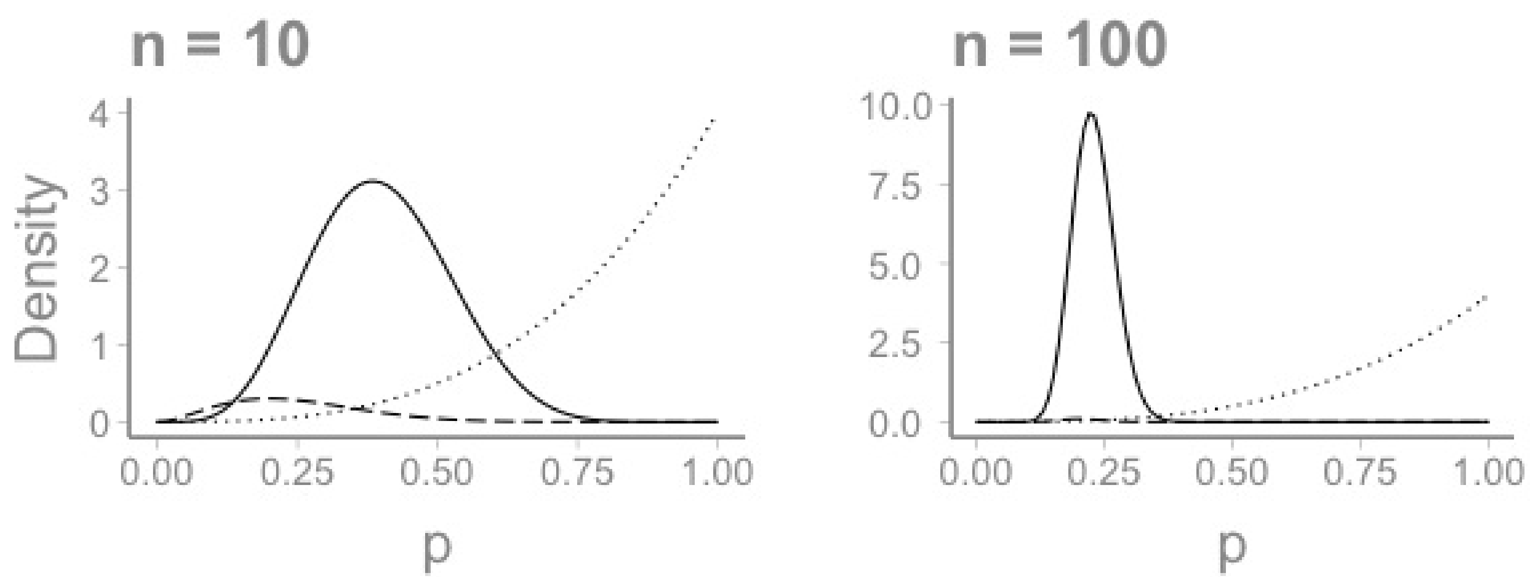
Practical issues with non-informative priors

No prior is truly non-informative

As we collect more data, this point will matter less and less

- if we tracked 100 individuals and 20 lived, our posterior would be \(beta(21, 81)\)

  With a sample size that big, we could use a pretty informative prior (e.g., \(\large beta(4,1)\)) and it doesn't matter too much:



Practical issues with non-informative priors

Another potential issue with non-informative priors is that they are not always non-informative if we have to transform them

Practical issues with non-informative priors

Another potential issue with non-informative priors is that they are not always non-informative if we have to transform them

For example, let's say instead of modeling p on the probability scale (0-1), we model it on the logit scale[1]:

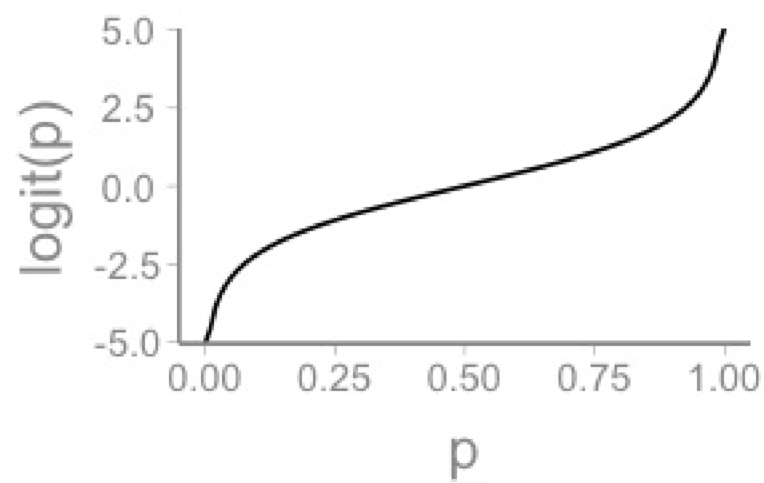
$$logit(p) \sim Normal(0, \sigma^2)$$

Practical issues with non-informative priors

Another potential issue with non-informative priors is that they are not always non-informative if we have to transform them

For example, let's say instead of modeling \(p\) on the probability scale (0-1), we model it on the logit scale \(^1\):

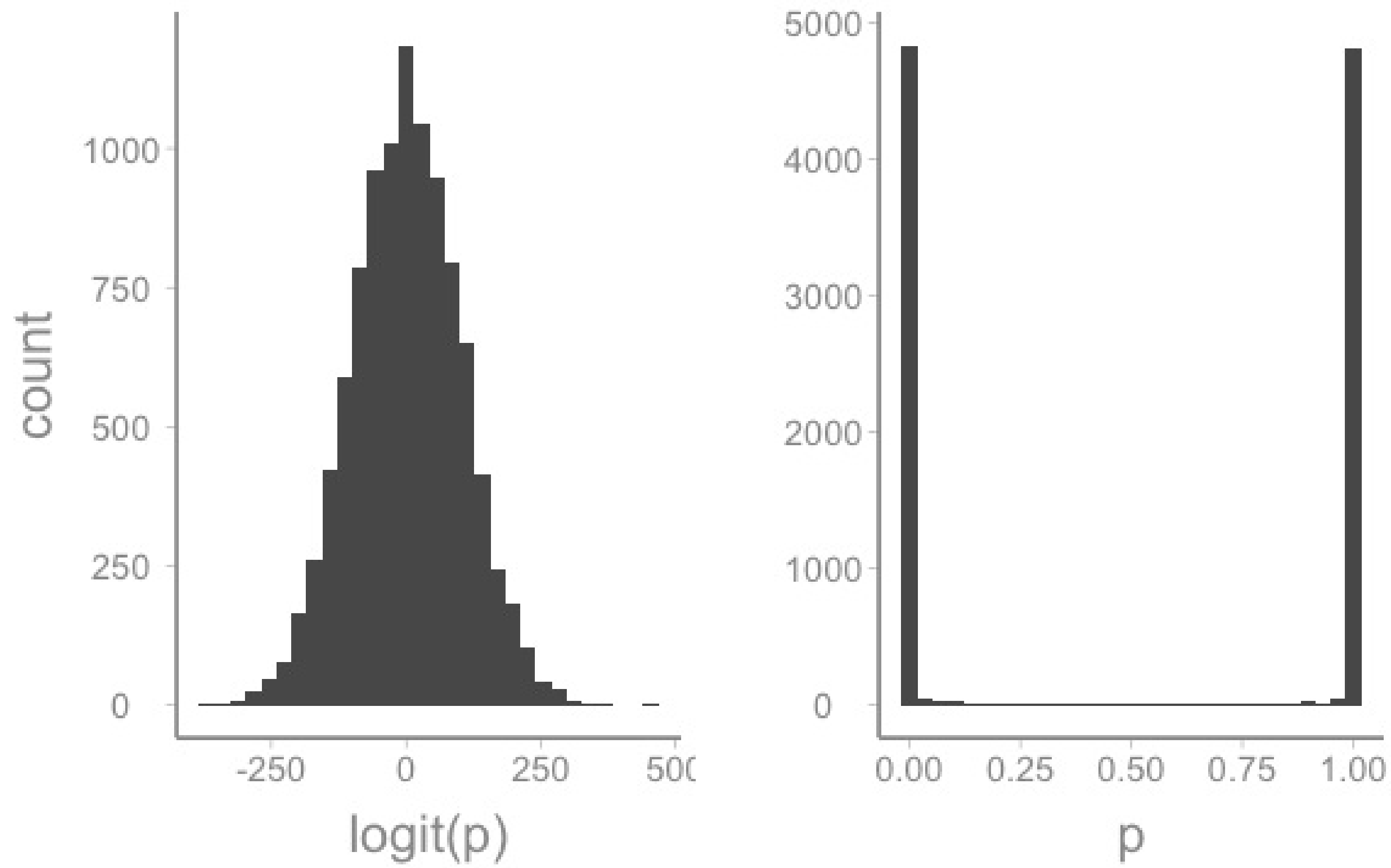
$$logit(p) \sim Normal(0, \sigma^2)$$
The logit scale transforms the probability to a real number:



Practical issues with non-informative priors

We saw that for real numbers, a "flat" normal prior is often chosen as a non-informative prior

But what happens when we transform those values back into probabilities?



Philosophical issues with non-informative priors

Aside from the practical issues, there are several philosophical issues regarding the use of non-informative priors

Philosophical issues with non-informative priors

Aside from the practical issues, there are several philosophical issues regarding the use of non-informative priors

Many of the justifications for using non-informative priors stem from the idea that we should be aim to reduce the influence of the prior so our data "do the talking""

Philosophical issues with non-informative priors

Aside from the practical issues, there are several philosophical issues regarding the use of non-informative priors

Many of the justifications for using non-informative priors stem from the idea that we should be aim to reduce the influence of the prior so our data "do the talking""

But even MLE methods are subjective \(^1\) - so avoiding Bayesian methods or choosing non-informative priors doesn't make our analyses more "objective"

Philosophical issues with non-informative priors

Aside from the practical issues, there are several philosophical issues regarding the use of non-informative priors

Many of the justifications for using non-informative priors stem from the idea that we should be aim to reduce the influence of the prior so our data "do the talking""

But even MLE methods are subjective \(^1\) - so avoiding Bayesian methods or choosing non-informative priors doesn't make our analyses more "objective"

Worse, MLE methods often hide these assumptions from view, making our subjective decisions implicit rather than explicit

Philosophical issues with non-informative priors

Science advances through the accumulation of facts

We are trained as scientists to be skeptical about results that are at odds with previous information about a system

Philosophical issues with non-informative priors

Science advances through the accumulation of facts

We are trained as scientists to be skeptical about results that are at odds with previous information about a system

If we find a unusual pattern in our data, we would be (and should be) very skeptical about these results

What's more plausible? These data really do contradict what we know about the system or they were generated by chance events related to our finite sample size?

Philosophical issues with non-informative priors

Science advances through the accumulation of facts

We are trained as scientists to be skeptical about results that are at odds with previous information about a system

If we find a unusual pattern in our data, we would be (and should be) very skeptical about these results

What's more plausible? These data really do contradict what we know about the system or they were generated by chance events related to our finite sample size?

Bayesian analysis is powerful because it provides a formal mathematical framework for combining previous knowledge with newly collected data \(^1\)

Philosophical issues with non-informative priors

Why shouldn't we use prior knowledge to improve our inferences?

> "Ignoring prior information you have is like selectively throwing away data before an analysis" (Hobbs & Hooten)

Philosophical issues with non-informative priors

Another way to think about the prior is as a hypothesis about our system

Philosophical issues with non-informative priors

Another way to think about the prior is as a hypothesis about our system

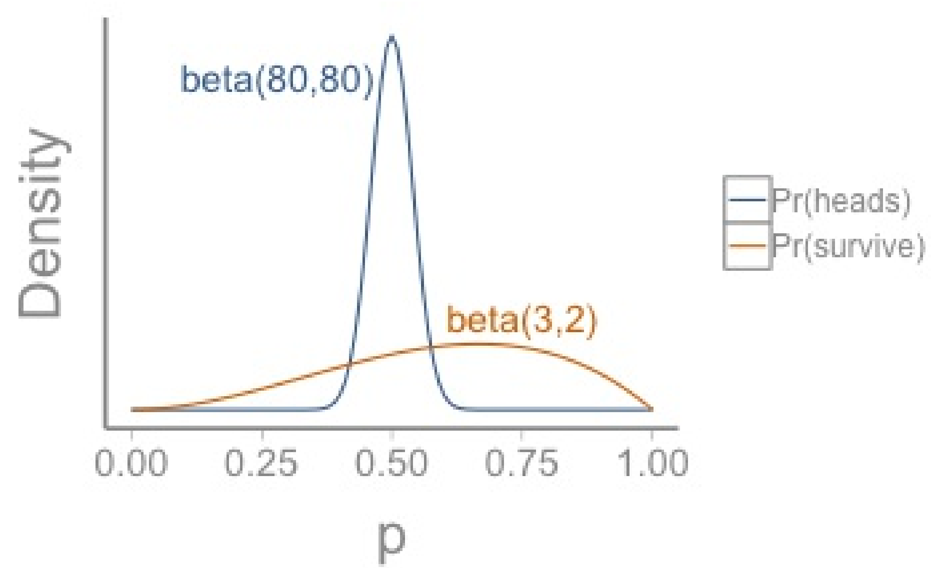
This hypothesis could be based on previously collected data or our knowledge of the system

Philosophical issues with non-informative priors

Another way to think about the prior is as a hypothesis about our system

This hypothesis could be based on previously collected data or our knowledge of the system

- If we know a lot, this prior could contain a lot of information \(^1\)
- If we don't know much, this prior could contain less information \(^2\)



Philosophical issues with non-informative priors

We allow our data to update this hypothesis

Philosophical issues with non-informative priors

We allow our data to update this hypothesis

- Strong priors will require larger sample sizes and more informative data to change our beliefs about the system

Philosophical issues with non-informative priors

We allow our data to update this hypothesis

- Strong priors will require larger sample sizes and more informative data to change our beliefs about the system
- Less informative priors will require less data

Philosophical issues with non-informative priors

We allow our data to update this hypothesis

- Strong priors will require larger sample sizes and more informative data to change our beliefs about the system
- Less informative priors will require less data

This is how it should be - if we're really confident in our knowledge of the system, we require very strong evidence to convince us otherwise

Advantages of informative priors

So there are clearly philosophical advantages to using informative priors

But there are also a number of practical advantages:

Advantages of informative priors

So there are clearly philosophical advantages to using informative priors

But there are also a number of practical advantages:

1) Improved parameter estimates when sample sizes are small

- Making inferences from small samples sizes is exactly when we expect spurious results and poorly estimated parameters
- Priors are most influential when we have few data \(^1\)
- The additional information about the parameters can reduce the chances that we base our conclusions on spurious results
- Informative priors can also reduce uncertainty in our parameter estimates and improve estimates of poorly identified parameters.

Advantages of informative priors

1) Improved parameter estimates when sample sizes are small

2) Stabilizing computational algorithms

- As models get more complex, the ratio of parameters to data becomes grows
- As a result, we often end up with pathological problems related to parameter identifiability \(^1\)
- Informative priors can improve estimation by providing some structure to the parameter space explored by the fitting algorithm

Advantages of informative priors

1) Improved parameter estimates when sample sizes are small

2) Stabilizing computational algorithms

Example

If $y \sim Binomial(N, p)$ and the only data we have in y (we don't know N), we cannot estimate N and p no matter how much data we have [1]

Including information about either N or p via an informative prior can help estimate the other other parameter [2]

Advice on priors

Advice on priors

1) Use non-informative priors as a starting point

> It's fine to use non-informative priors as you develop your model but you should always prefer to use "appropriate, well-contructed informative priors" (Hobbs & Hooten)

Advice on priors

1) Use non-informative priors as a starting point

2) Think hard

> Non-informative priors are easy to use because they are the default option. You can usually do better than non-informative priors but it requires thinking hard about the parameters in your model

Advice on priors

1) Use non-informative priors as a starting point

2) Think hard

3) Use your "domain knowledge"

> We can often come up with weakly informative priors just be knowing something about the range of plausible values of our parameters

Advice on priors

1) Use non-informative priors as a starting point

2) Think hard

3) Use your "domain knowledge"

4) Dive into the literature

> Find published estimates and use moment matching and other methods to convert published estimates into prior distributions

Advice on priors

1) Use non-informative priors as a starting point

2) Think hard

3) Use your "domain knowledge"

4) Dive into the literature

5) Visualize your prior distribution

> Be sure to look at the prior in terms of the parameters you want to make inferences about (use simulation!)

Advice on priors

1) Use non-informative priors as a starting point

2) Think hard

3) Use your "domain knowledge"

4) Dive into the literature

5) Visualize your prior distribution

6) Do a sensitivity analysis

> Does changing the prior change your posterior inference? If not, don't sweat it. If it does, you'll need to return to point 2 and justify your prior choice