Module 2: Introduction to Bayesian Statistics

Andrew Parnell, School of Mathematics and Statistics,
University College Dublin

Learning outcomes

- Know the difference between Frequentist and Bayesian statistics
- Understand the terms posterior, likelihood and prior. Be able to suggest suitable probability distributions for these terms
- Be able to interpret the posterior distribution through plots, summaries, and credible intervals

Who was Bayes?

An essay towards solving a problem on the doctrine of chances (1763)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

What is Bayesian statistics?

- Bayesian statistics is based on an interpretation of Bayes' theorem
- All quantities are divided up into *data* (i.e. things which have been observed) and *parameters* (i.e. things which haven't been observed)
- We use Bayes' interpretation of the theorem to get the *posterior probability distribution*, the probability of the unobserved given the observed
- Used now in almost all areas of statistical application (finance, medicine, environmetrics, gambling, etc, etc)

Why Bayes?

The Bayesian approach has numerous advantages:

- It's easier to build complex models and to analyse the parameters you want directly
- We automatically obtain the best parameter estimates and their uncertainty from the posterior samples
- It allows us to get away from (terrible) null hypothesis testing and p-values

Bayes theorem in english

Bayes' theorem can be written in words as:

$$\text{posterior is proportional to likelihood times prior}$$

. . . or . . .

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Each of the three terms *posterior*, *likelihood*, and *prior* are *probability distributions* (pdfs).

In a Bayesian model, every item of interest is either data (which we will write as x) or parameters (which we will write as θ). Often the parameters are divided up into those of interest, and other *nuisance parameters*

Bayes theorem in maths

Bayes' equation is usually written mathematically as:

$$p(\theta|x) \propto p(x|\theta) \times p(\theta)$$

or, more fully:

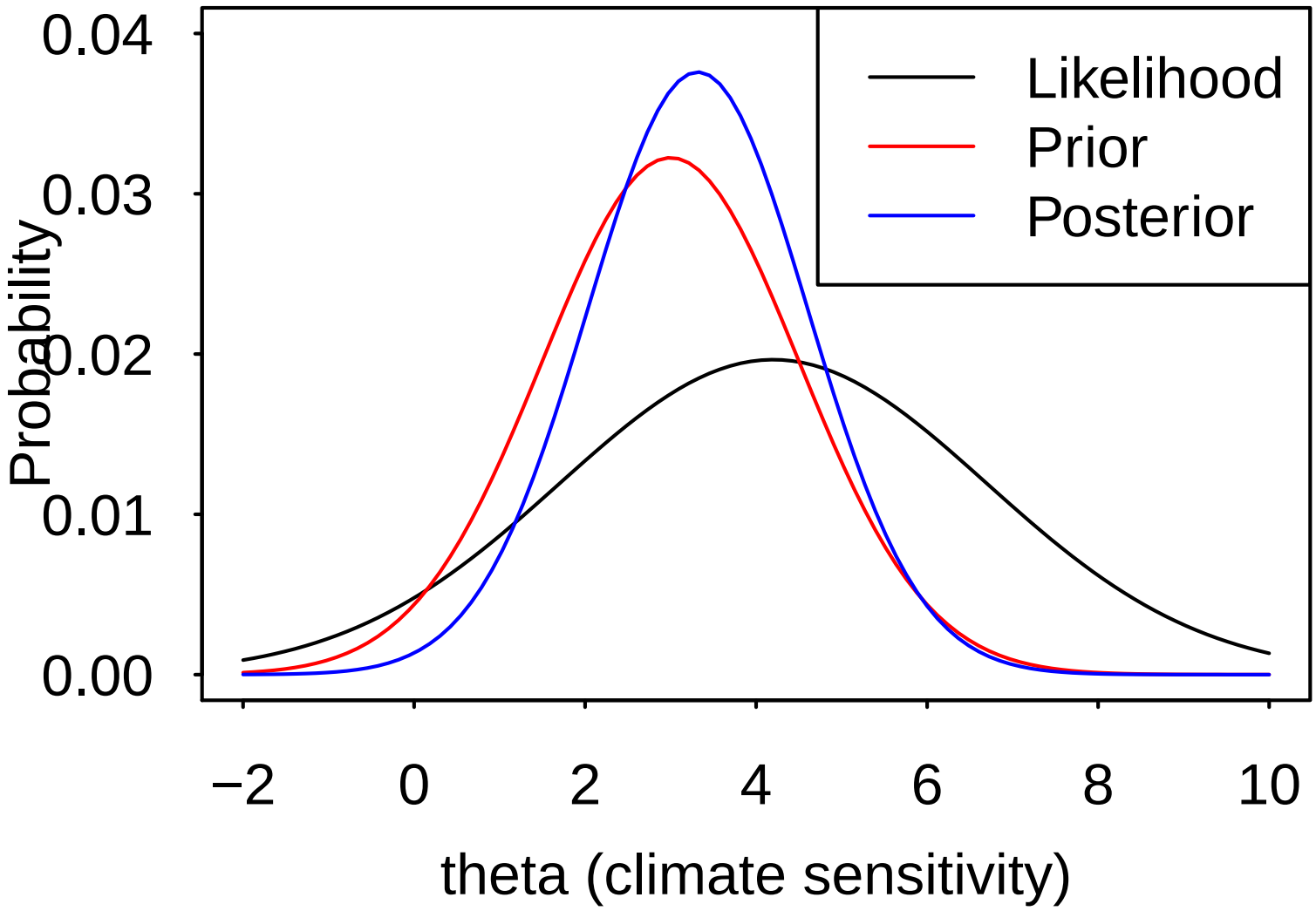
$$p(\theta|x) = \frac{p(x|\theta) \times p(\theta)}{p(x)}$$

- The *posterior* is the probability of the parameters given the data
- The *likelihood* is the probability of observing the data given the parameters (unknowns)
- The *prior* represents external knowledge about the parameters

Example 1: one parameter to estimate

- Let's suppose we want to estimate an unknown parameter θ which represents climate sensitivity
- A set of previous studies have estimated climate sensitivity to be $3 \pm 1.5^{\circ}C/W/m^2$ where 1.5 represents one standard deviation. We will assume this is normally distributed and use this as a prior distribution, so $\theta \sim N(3, 1.5^2)$
- We run a new climate model which gives us a new data point: $x = 4.2^{\circ}C/W/m^2$. Let's suppose that we know from previously running this model that the standard deviations of these runs is $2.5^{\circ}C/W/m^2$,
- Our likelihood states that, if we knew θ, the value x would be normally distributed around θ with standard deviation 2.5, so $x \sim N(\theta, 2.5^2)$
- We now need to use Bayes theorem to estimate the posterior distribution of climate sensitivity given our new run

Example 1 continued



Note: posterior mean is 3.32 $^\circ C/W/m^2$ and standard deviation is 1.29 $^\circ C/W/m^2$

Example 1 code

```
# Create grid for theta
theta = seq(-2, 10, length = 100)
# Evalutate prior, likelihood and posterior
prior = dnorm(theta, mean = 3, sd = 1.5)
likelihood = dnorm(4.2, mean = theta, sd = 2.5)
posterior = prior * likelihood
# Produce plot
plot(theta, likelihood / sum(likelihood), type = 'l',
     ylab = 'Probability', ylim = c(0, 0.04))
lines(theta, prior / sum(prior), col = 'red')
lines(theta, posterior / sum(posterior), col = 'blue')
legend('topright',
       legend = c('Likelihood', 'Prior', 'Posterior'),
       col = c('black', 'red', 'blue'),
       lty = 1)
```

Understanding the different parts of a Bayesian model

- The likelihood is the probability of observing the data given the parameters. It represents the *data generating process*
- The prior is the probability distribution of the parameters independent from the current data you have been generated. It often requires care (and philosophy) to choose. More on this later
- The posterior is the probability distribution of the parameters given the data. It is always the target of our inference.

In a Bayesian model we 'simply' specify the likelihood and the prior. JAGS (or other software) will calculate the posterior for us

Lots of probability distributions

Almost always, the likelihood and prior can be picked from the standard probability distributions:

Distribution	Range of parameter	Useful for:
Normal, $N(\mu, \sigma^2)$	$(-\infty, \infty)$	A good default choice
Uniform, $U(a, b)$	(a, b)	Vague priors when we only know the range of the parameter
Binomial, $Bin(k, \theta)$	$[0, k]$	Count or binary data restricted to have an upper value
Poisson, $Po(\lambda)$	$[0, \infty)$	Count data with no upper limit
Gamma, $Ga(\alpha, \beta)$	$(0, \infty)$	Continuous data with a lower bound of zero
Multivariate	$(-\infty, \infty)$	Multivariate

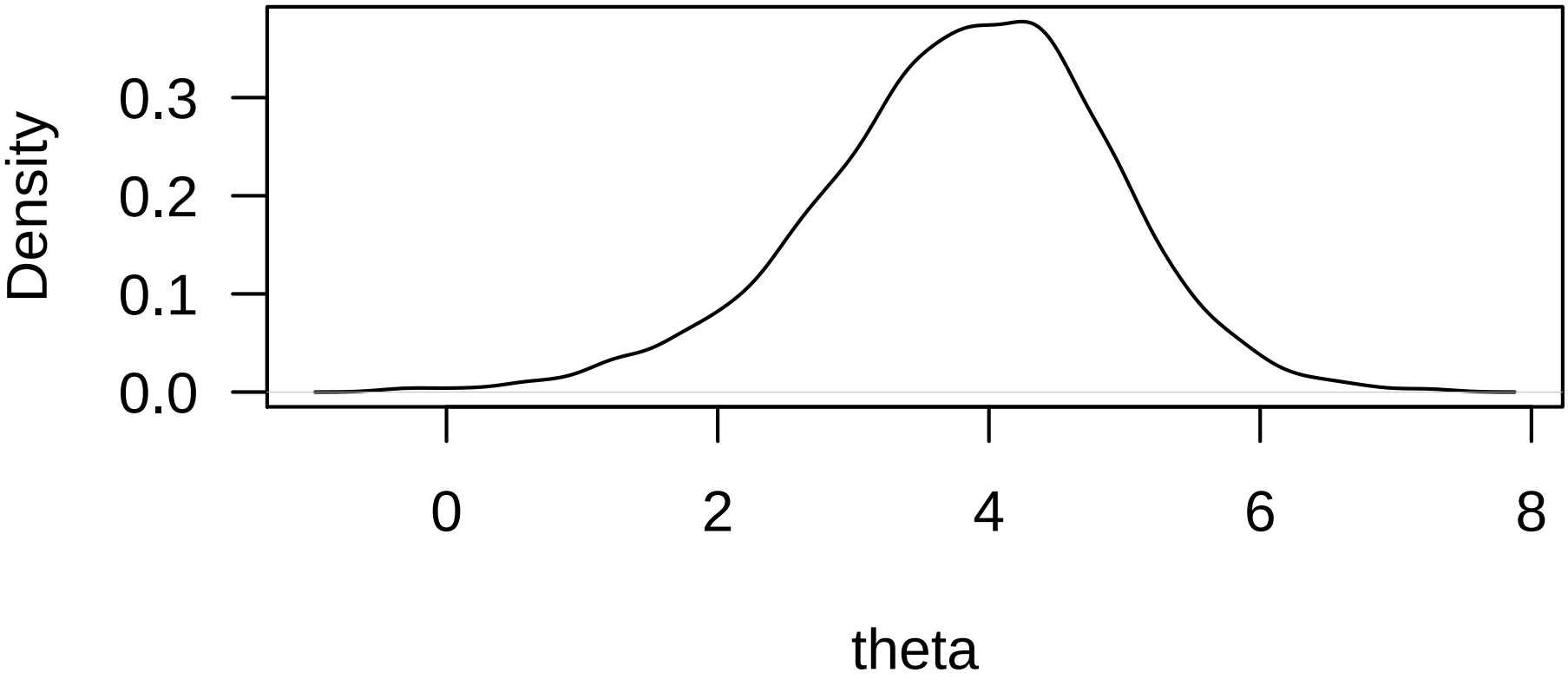
Choosing a likelihood and a prior

- Both can be hard choices
- For the likelihood, think about the range of the data values. Can they be positive and negative? Are they likely to be skewed? Are they counts or continuous numbers? Are there likely to be outliers?
- For the prior, think about what previous information is available. If very little information is available use a *vague* prior, such as $N(0, 100)$. If there are previous studies, use values from these. If there are experts around, ask them.
- If you can, use an *informative* prior

Example 1 (continued): Multiple observations and parameters

- Let's make example 1 more realistic
- The prior on climate sensitivity is the same as before: $\theta \sim N(3, 1.5^2)$
- Suppose now we have four new runs from climate models, so $x_1 = 4.2, x_2 = 1.6, x_3 = 6.7, x_4 = 5.8$
- We're not going to pretend that we know the true standard deviation of the runs any more so we have two parameters to estimate, the true climate sensitivity (θ), and the variability of our climate model runs (σ). Write this likelihood as $x \sim N(\theta, \sigma^2)$.
- We now also need a prior for σ. Suppose we have little information here, so we will use the vague prior $\sigma \sim U(0, 10)$
- We compute a *joint posterior distribution* of both θ and σ given the data

Example 1 (continued): posterior distribution



Posterior mean of θ is now 3.81 $^\circ C/W/m^2$ and standard deviation is 1.09 $^\circ C/W/m^2$

Showing your module

- Often all you need to do to describe a Bayesian model is define your notation, write down your likelihood and your prior
- For the example on the previous slide:

$$\text{Likelihood} : x_i \sim N(\theta, \sigma^2), i = 1, \ldots, N$$

$$\text{Priors} : \theta \sim N(3.15, 1.5^2), \sigma \sim U(0, 10)$$

- Note that it's very easy to simulate data values from this model:

```
sigma = runif(1, 0, 10)
theta = rnorm(1, 3.15, 1.5)
x = rnorm(1, theta, sigma)
```

- We can run this repeatedly. If the data look 'reasonable' then we know that our prior and likelihood are well chosen

Posterior computation in JAGS

- Here is the JAGS code to calculate the posterior distribution:

```
model
{
  # Likelihood
  for (i in 1:N) {
    y[i] ~ dnorm(theta, tau)
  }
  # Priors
  theta ~ dnorm(3, 1/pow(1.5, 2))
  tau <- 1/pow(sigma,2)
  sigma ~ dunif(0.0,10.0)
}
```

- It looks a lot like R code, but there are a few key differences:
  - You must use <- for assignment and ~ for the likelihood and the prior
  - The `dnorm` function in JAGS uses *precision* (1/variance) rather than standard deviation
  - I always transform the precision back into a standard deviation

Calculating the posterior vs sampling from it

- There are two ways to get at a posterior:
  1. Calculate it directly using hard maths
  2. Use a simulation method
- Number 1 is impractical once you move beyond a few parameters, so number 2 is used by almost everybody
- This means that we create *samples* from the posterior distribution. Here are three samples from the previous example:

```
##          sigma    theta
## [1,] 3.623002 3.467479
## [2,] 1.672596 3.868244
## [3,] 5.458645 3.488686
```

- We often create thousands of posterior samples to represent the posterior distribution

Things you can do with posterior samples

- Create histograms or density plots:
- Individual summaries such as means, standard deviations, and quantiles (e.g. 95% confidence intervals)
- Joint summaries such as scatter plots or correlations
- Transformations such as logs/exponents, squares/square roots, etc

The posterior distribution will usually be stored in a matrix where each row is a sample, and each column is a different parameter. Having the posterior distribution enables you to get at exactly the quantities you are interested in

Summary so far: for and against Bayes

For:

- A Bayesian model can be simply displayed as a likelihood and a prior. Everything is explicit
- JAGS finds the posterior distribution for us so we don't need to do any maths
- We can get exactly the quantity we are interested in, the probability distribution of our unknowns given our knowns

Against:

- It can be hard to create a prior distribution (and a likelihood)
- Not having p-values can make papers harder to publish (but this is changing)

Example 1 (continued again): hierarchical version

- The likelihood we used previously, $x_i \sim N(\theta, \sigma^2)$ states that each model run (perhaps from separate GCMs) is providing evidence for *one* climate sensitivity parameter. Is this realistic?
- A slightly more realistic model would be $x_i \sim N(\theta_i, \sigma^2)$ where each climate model has its *own climate sensitivity parameter*
- We can't fit this model because we don't have enough information to estimate a parameter from 1 data point from each climate model. We instead put a *prior* such that $\theta_i \sim N(\theta_0, \sigma_0^2)$ and can put further prior distributions on θ_0 and σ_0
- We can now estimate each climate model's own climate sensitivity (θ_i) and the overall climate sensitivity θ_0
- This is known as a hierarchical model

Bayesian time series, some general notation

Throughout this course we use the following general notation:

- Roman letters for data, Greek for parameters
- y_t for a the time series values we're interested in modelling at time t.
- x for explanatory variables
- $\theta, \phi, ...$ for unknown parameters
- σ for parameters representing standard deviations

There are a few exceptions later in the course (e.g. state-space models)

General tips:

- If you have lots of disparate data, try to build one model for all it. You'll be able to *borrow strength* from the data (e.g. in a hierarchical model) and reduce the uncertainty in your parameter estimates
- Try your hardest to use informative priors, and always justify the values you use (especially when trying to publish). In this course we're presenting generic versions so have almost always used vague priors
- Check your model. Many of the usual requirements from traditional statistics (e.g. residual checks) are still relevant in the Bayesian world. There are some extra Bayesian tricks we can also do; discussed in later lectures

Summary

- Bayesian statistical models involve a *likelihood* and a *prior*. These both need to be carefully chosen. From these we create a posterior distribution
- The likelihood represents the information about the data generating process; the prior information about the unknown parameters
- We usually create and analyse samples from the posterior probability distribution of the unknowns (the parameters) given the knowns (the data)
- From the posterior distribution we can create means, medians, standard deviations, credible intervals, etc, from samples we take using e.g. JAGS