Session 2.1: Bayesian inference

Imperial College London

Learning Objectives

After this session you should be able to:

- Understand how Bayes Theorem can be applied to random variables
- Describe what conjugacy means
- Obtain posterior distribution for the Beta-Binomial and Gamma-Poisson families

The topics treated in this lecture are presented in Chapter 3 of the book Blangiardo and Cameletti (2015) and in Chapter 2.3, 3.1-3.4 and 5.2-5.3 of the book Johnson, Ott, and Dogucu (2022).

Outline

1. Bayes Theorem for random variables
2. A quick recap
3. What is conjugacy?
4. Some conjugacy models: Beta-Binomial
5. Some conjugacy models: Gamma-Poisson

Bayes Theorem for random variables

An example

- A clinical trial is carried out to assess the efficacy of a preventive treatment for migraine
- A group of 20 patients are offered the new drug and they have to report if they have a migraine episode in the next week after taking the medication.
- Our aim is to estimate the probability of success of this new drug (θ)
- Our data here is Y: the number of patients reporting no migraine episodes.

  Note that Y is a **random variable** (look at recording 3 for a more detailed description of this concept)

Prior probability model

As a first step we need to assign a prior on θ.

As a simplification let's assume that θ is discrete and can only get thevalues 0.1,0.4,0.8 with the following probability function, which specifies the prior probability of each possible θ value:

θ	$p(\theta)$
0.1	0.2
0.4	0.5
0.8	0.3

Note that this prior reflects some sort of information from a previous study on a similar compound and put 50% probability on the event that 40% of the patients will report a reduction in migraine.

The Binomial data model

- Our random variable Y can take any discrete value between 0 and 20 (total number of patients)
- It will depend on the probability θ
- For our formal Bayesian analysis, we must model this dependence of Y on θ through a **conditional probability model**
- We make two assumptions about the trials: (1) the outcome of any one patient doesn't influence the outcome of another; and (2) the probability of success does not change among patients
- We can use the Binomial model

$Y \sim \text{Binomial}(n, \theta)$

with conditional probability function

$$p(y \mid \theta) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y}$$

- Mean of this distribution is $E(Y) = n\theta$
- Variance is $V(Y) = n\theta(1-\theta)$

Check out recording 4 for a recap on the Binomial distribution

The Binomial data model

This model allows to calculate ANY conditional probability. For instance, conditioning on $\theta = 0.4$ let's see the difference in the probability of getting 12 or 15 successes

1. $$p(y = 12 \mid \theta = 0.4) = \frac{20!}{12!8!} 0.4^{12}(1 - 0.4)^8 = 0.035$$

2. $$p(y = 15 \mid \theta = 0.4) = \frac{20!}{15!5!} 0.4^{15}(1 - 0.4)^5 = 0.0013$$

--

Note you can get these results in R using

```
> #1
> dbinom(12,20,0.4)
```

```
[1] 0.03549744
```

```
> #2
> dbinom(15,20,0.4)
```

```
[1] 0.001294494
```

Binomial likelihood function

- The Binomial provides a theoretical model of the data we might observe.
- In the end we observe 10 successes out of the 20 patients $(y = 10)$.
- The next step in our Bayesian analysis is to determine how compatible this particular data is with the various possible θ

We need to evaluate the *likelihood* of getting 10 successes in the trial under each possible value of θ

Similarly to last week with events, the likelihood function follows from evaluating the conditional probability function $p(y = 10 \mid \theta)$ for all the possible θ values:

1. $$p(y = 10 \mid \theta = 0.1) = \frac{20!}{10!10!} 0.1^{10}(1 - 0.1)^{10} = 0.00000644$$
2. $$p(y = 10 \mid \theta = 0.4) = \frac{20!}{10!10!} 0.4^{10}(1 - 0.4)^{10} = 0.117$$
3. $$p(y = 10 \mid \theta = 0.8) = \frac{20!}{10!10!} 0.8^{10}(1 - 0.8)^{10} = 0.002$$

Likelihood vs probability function

Let's recall the fundamental difference between probability function and likelihood function

When θ is known, the **conditional probability function** $p(\cdot \mid \theta)$ allows us to compare the probabilities of different values of the data Y occurring with θ:

$$p(y_1 \mid \theta) \text{ vs } p(y_2 \mid \theta)$$

When Y is known, the **likelihood function** $L(\cdot \mid y) = p(y \mid \cdot)$ allows us to compare the relative likelihood of the data y under different possible values of θ:

$$L(\theta_1 \mid y) = p(y \mid \theta_1) \text{ vs } L(\theta_2 \mid y) = p(y \mid \theta_2)$$

So $L(\cdot \mid y)$ provides the tool we need to evaluate the relative compatibility of data $Y = y$ with various θ values.

Normalising constant

- Now we have a prior for θ and a likelihood and as Bayesian we want to **balance** these two pieces of information to obtain the posterior.
- Something is missing...

Normalising constant

- Now we have a prior for θ and a likelihood and as Bayesian we want to **balance** these two pieces of information to obtain the posterior.
- Something is missing...

The normalising constant!

This is the total probability of having $y = 10$ successes. How do we get this?

Normalising constant

- Now we have a prior for θ and a likelihood and as Bayesian we want to **balance** these two pieces of information to obtain the posterior.
- Something is missing...

The normalising constant!

This is the total probability of having $y = 10$ successes. How do we get this?

Law of total probability:

$$P(A) = \sum_j P(A \mid B_j)P(B_j) \text{ (for events)}$$

$$P(Y = y) = \sum_j p(Y \mid \theta_j)p(\theta_j) \text{ (for discrete probability functions)}$$

So we get

$$P(Y = 10) = 0.00000644 \times 0.2 + 0.117 \times 0.5 + 0.002 \times 0.3 = 0.059$$

Interpretation: across all the possible θ there is only around 6% chance that there are 10 successes.

Posterior distribition

Finally we are ready to apply Bayes theorem and get the posterior distribution

$$p(\theta \mid y = 10) = \frac{p(\theta)L(\theta \mid y = 10)}{p(y = 10)} \text{ for } \theta \in \{0.1, 0.4, 0.8\}$$

which gives us:

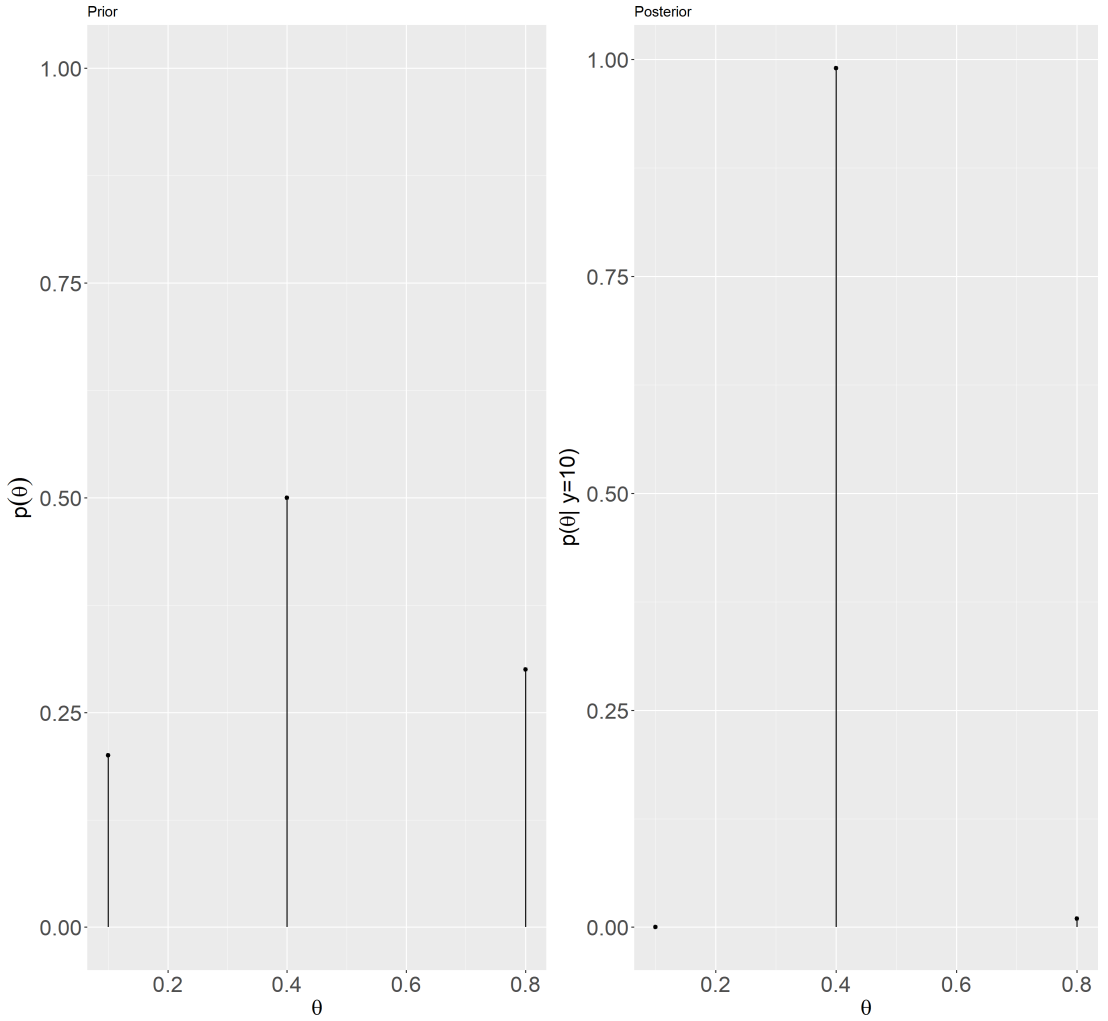
$$p(\theta = 0.1 \mid y = 10) = \frac{0.2 \times 0.0000064}{0.059} = 0$$

$$p(\theta = 0.4 \mid y = 10) = \frac{0.5 \times 0.117}{0.059} = 0.99$$

$$p(\theta = 0.8 \mid y = 10) = \frac{0.3 \times 0.002}{0.059} = 0.01$$

Comparing prior and posterior

θ	$p(\theta)$	$p(\theta \mid y = 10)$
0.1	0.2	0.00
0.4	0.5	0.99
0.8	0.3	0.01



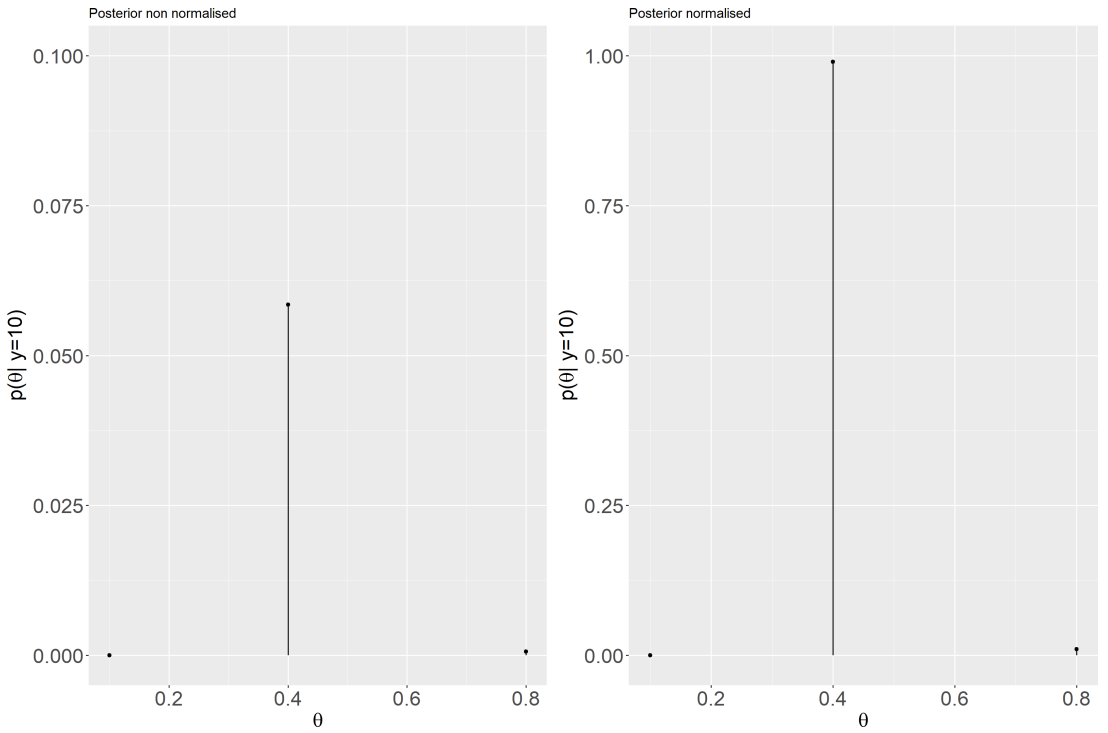
Posterior shortcut

- Moving forward we can actually forget about calculating the normalising constant
- Note that on slide 12 the $p(y = 10)$ appears at the denominator of all the $p(\theta \mid Y)$

  It normalises the posterior probabilities so they sum to 1
- So we can simply acknowledge that $p(y = 10) = 1/c$ and replace the posterior as:

$$p(\theta \mid y = 10) \propto p(\theta) \times L(\theta \mid y)$$

- The proportionality means that if we compare the normalised and unnormalised posterior they preserve their relative relationship



A quick recap

So a quick general recap: Bayesian inference

Makes fundamental distinction between

- Observable quantities y, i.e.~the data
- Unknown quantities θ
- θ can be statistical parameters, missing data, mismeasured data...

  $\rightarrow$ parameters are treated as random variables

  $\rightarrow$ in the Bayesian framework, we make probability statements about model parameters

So a quick general recap: Bayesian inference

Makes fundamental distinction between

- Observable quantities y, i.e.~the data
- Unknown quantities θ
- θ can be statistical parameters, missing data, mismeasured data...

  $\rightarrow$ parameters are treated as random variables

  $\rightarrow$ in the Bayesian framework, we make probability statements about model parameters

Note that in the Frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data

Bayesian inference

- As with any statistical analysis, we start building a model which specifies $p(Y = y \mid \theta)$
- This is the data distribution, which relates all variables into a **full probability model**
- The choice of data distribution depends on the nature of the data:

e.g. are we analysing continuous or discrete data, are the data symmetric or skewed, etc.

- As we observe the data, we can use descriptive tools (e.g. plots) to visualise the data and choose the best likelihood

Bayesian inference

From a Bayesian point of view

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data

$\rightarrow$ need to specify a prior distribution $p(\theta)$

- y is known so we should condition on it

  $\rightarrow$ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \propto p(\theta)\, p(y \mid \theta)$$

This is the posterior distribution

Bayesian inference

From a Bayesian point of view

- θ is unknown so should have a probability distribution reflecting our uncertainty about it before seeing the data

$\rightarrow$ need to specify a prior distribution $p(\theta)$

- y is known so we should condition on it

  $\rightarrow$ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \propto p(\theta)\, p(y \mid \theta)$$

This is the posterior distribution

- The prior distribution $p(\theta)$, expresses our uncertainty about θ before seeing the data
- The posterior distribution $p(\theta \mid y)$, expresses our uncertainty about θ after seeing the data

The posterior distribution

Posterior distribution forms basis for all inference --- can be summarised to provide

- point and interval estimates of Quantities of Interest (QOI), e.g. treatment effect, small area estimates, ...
- point and interval estimates of any function of the parameters
- probability that QOI (e.g. treatment effect) exceeds a critical threshold
- prediction of QOI in a new unit
- prior information for future experiments, trials, surveys, ...
- inputs for decision making
- ...

What is conjugacy?

How to select a prior

- Selecting the prior is crucial for a Bayesian analysis

  - There is no right way to select a prior
  - The choices often depend on the objective of the study and the nature of the data

- There are other criteria to consider when choosing a prior model:

Computational ease

Especially if we don't have access to computing power, it is helpful if the posterior model is easy to build.

Interpretability

The posterior balance (is a compromise) the data and the prior. A posterior model is interpretable, and thus more useful, when you can look at its formulation and identify the contribution of the data relative to that of the prior.

How to select a prior

- Selecting the prior is crucial for a Bayesian analysis

  - There is no right way to select a prior
  - The choices often depend on the objective of the study and the nature of the data

- There are other criteria to consider when choosing a prior model:

Computational ease

Especially if we don't have access to computing power, it is helpful if the posterior model is easy to build.

Interpretability

The posterior balance (is a compromise) the data and the prior. A posterior model is interpretable, and thus more useful, when you can look at its formulation and identify the contribution of the data relative to that of the prior.

We introduce now a type of models which satisfy both properties

Conjugate prior

Let the prior model for a parameter θ with $p(\theta)$ and the model of data Y conditioned on θ have likelihood function $L(\theta \mid y)$.

If the resulting posterior model $p(\theta \mid y) \propto p(\theta) \times p(y \mid \theta)$ is of the same family as the prior, then we say it is a **conjugate prior**.

Some conjugacy models: Beta-Binomial

Beta-Binomial model for proportions: example

- We consider an early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- We interpret this as a distribution with mean = 0.4 and standard deviation 0.1
- A Beta(9.2,13.8) distribution has these properties (Check recording 7 to see how to go from the mean and sd to the a and b parameters of a Beta distribution)
- Suppose we now treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses

Identifying the different model components

- Assuming patients are independent, with common unknown response rate θ, leads to a binomial data distribution

$$p(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \propto \theta^y (1-\theta)^{n-y}$$

- θ needs a continuous prior distribution:

$$\theta \sim \text{Beta}(a, b)$$
$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

Combining prior and data

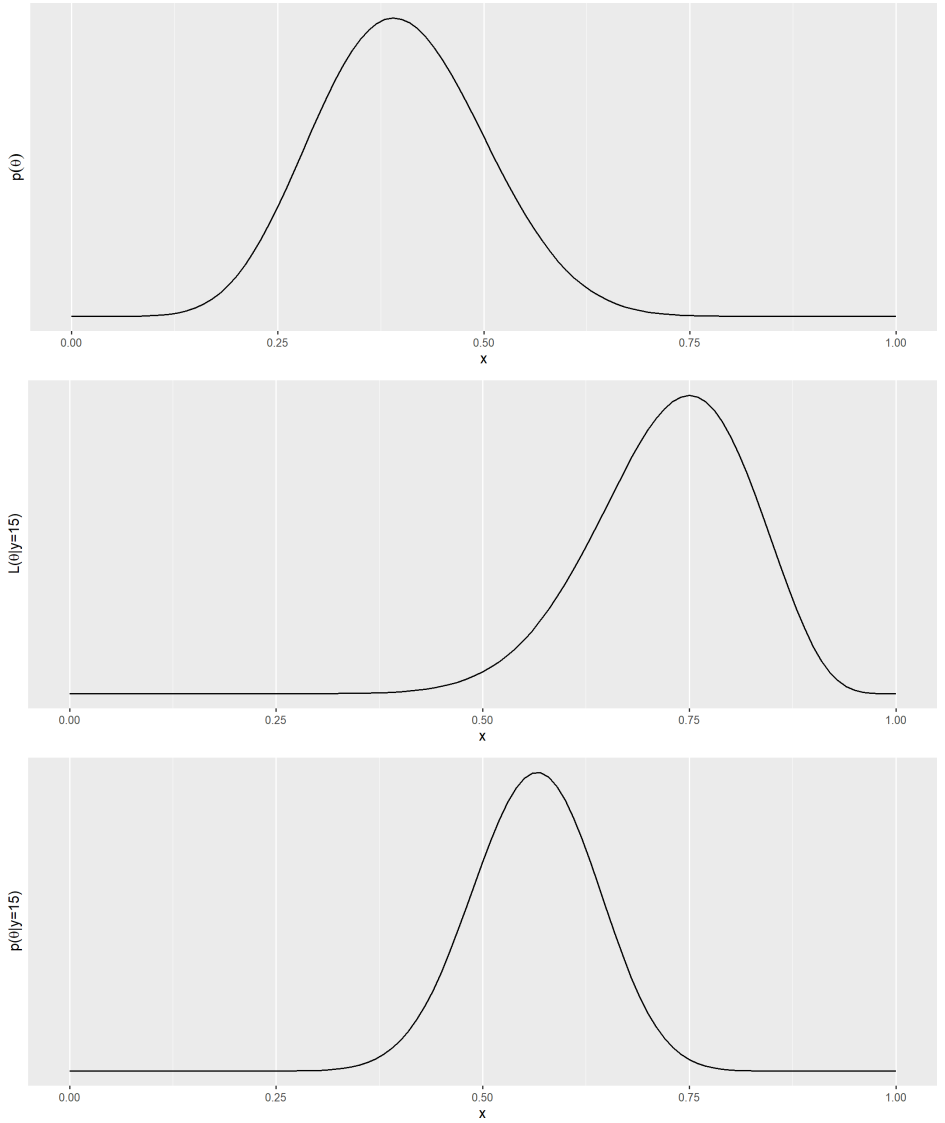
Combining the Binomial data and the Beta prior gives the following posterior distribution

$$\begin{aligned} p(\theta \mid y, n) &\propto p(y \mid \theta, n) p(\theta) \\ &\propto \theta^{y} (1-\theta)^{n-y} \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{y+a-1} (1-\theta)^{n-y+b-1} \end{aligned}$$

The posterior is still a Beta distribution (with different parameters):

$$p(\theta \mid y, n) \propto \text{Beta}(y + a,\ n - y + b)$$

Comparing prior, likelihood and posterior



Gamma-Poisson model for count data: example

- For a recap on the Poisson distribution see recording 5

- In epidemiology we are often interested in estimating the rate or relative risk rather than the mean for Poisson data:

- Suppose we observe $y = 7$ cases of leukaemia in one region;
- The expected number of cases is $E = 4$

- **Data distribution**: Poisson with mean $\theta = \lambda \times E$, where λ is the unknown incidence ratio:

$$p(y \mid \lambda, E) = \frac{(\lambda E)^y e^{-\lambda E}}{y!}$$

- **Prior**: $\text{Gamma}(a, b)$ on the the risk λ:

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

Check recording 8 for a recap on the Gamma distribution

Combining likelihood and prior

This implies the following posterior

$$
\begin{aligned}
p(\lambda \mid y) &\propto p(\lambda)\, p(y \mid \lambda) \\
&= \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b(\lambda)} e^{-(\lambda E)} \frac{(\lambda E)^y}{y!} \\
&\propto \lambda^{a+y-1} e^{-(b+E)\lambda} \\
&= \text{Gamma}(a+y,\, b+E)
\end{aligned}
$$

Combining likelihood and prior

This implies the following posterior

$$
\begin{aligned}
p(\lambda \mid y) &\propto p(\lambda)\, p(y \mid \lambda) \\
&= \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b(\lambda)} e^{-(\lambda E)} \frac{(\lambda E)^y}{y!} \\
&\propto \lambda^{a+y-1} e^{-(b+E)\lambda} \\
&= \text{Gamma}(a+y,\, b+E)
\end{aligned}
$$

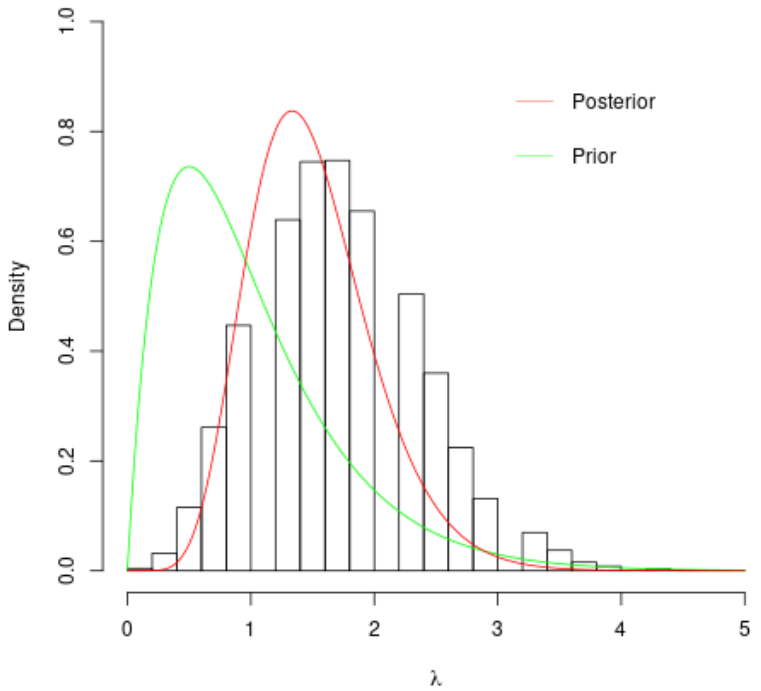
The posterior is another (different) Gamma distribution

$$
E(\theta \mid y) = \frac{a+y}{b+E}
$$

So posterior mean depends on the prior (a, b) and on the data (y, E)

Prior, likelihood, posterior for the Leukaemia example

- Assuming a prior $\lambda \sim \text{Gamma}(2, 2)$
- Considering the data $y = 7, E = 4$
- We obtain a posterior $\lambda \mid y \sim \text{Gamma}(9, 6)$ centered around 1.5



The posterior becomes a compromise between the prior and the data

References

Blangiardo, M. and M. Cameletti (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Johnson, A. A., M. Q. Ott, and M. Dogucu (2022). *Bayes Rules!: An Introduction to Applied Bayesian Modeling*. CRC Press.