Session 2.2: Posterior Predictive Distribution and Monte Carlo computation

Bayesian modelling for Spatial and Spatio-temporal data, Imperial College

Learning objectives

After this lecture you should be able to

- Describe what the posterior predictive distribution (PPD) is
- Explain how the PPD is computable
- Describe what Monte Carlo simulation is, and why it is useful
- The topic of posterior prediction is treated in Section 8.3 of Johnson, Ott, and Dogucu (2022)
- The topic of Monte Carlo simulation is presented in Sections 4.1-4.4 of Blangiardo and Cameletti (2015).

Outline

Posterior Predictive Distribution

Bayesian prediction

- Often the objective of our analysis is to predict a future event, based upon the data currently available.

Bayesian prediction

- Often the objective of our analysis is to predict a future event, based upon the data currently available.
- Consider this example:

We estimate the prevalence of a disease in a UK hospital using a sample of n = 58 individuals.

We find that $y = 10$ individuals have the disease.

What is the probability that, if we additionally sample `(k = 30)` individuals this year, at least 5 will have the disease?

Bayesian prediction

- Often the objective of our analysis is to predict a future event, based upon the data currently available.
- Consider this example:

We estimate the prevalence of a disease in a UK hospital using a sample of n = 58 individuals.

We find that $y = 10$ individuals have the disease.

What is the probability that, if we additionally sample `(k = 30)` individuals this year, at least 5 will have the disease?

- As usual we start specifying the data distribution:

$$y \sim \text{Binomial}(\theta, n = 58)$$

- Let's θ be the true disease prevalence and y^* be the predicted value

Bayesian prediction

- Often the objective of our analysis is to predict a future event, based upon the data currently available.
- Consider this example:

We estimate the prevalence of a disease in a UK hospital using a sample of n = 58 individuals.

We find that $y = 10$ individuals have the disease.

What is the probability that, if we additionally sample `(k = 30)` individuals this year, at least 5 will have the disease?

- As usual we start specifying the data distribution:

$$y \sim \text{Binomial}(\theta, n = 58)$$

- Let's θ be the true disease prevalence and y^* be the predicted value
- If θ were known, then we would predict

$$y^*|\theta \sim \text{Binomial}(30, \theta)$$

thus $\mathrm{P}(y \geq 5) = 1 - (\sum_{j=0}^{4} \theta^j (1-\theta)^{30-j})$

BUT ... θ is unknown

Source of variation in prediction:

- We don't know the true value of the parameters and we specify a prior on it:

$$\theta \sim \mathrm{Beta}(a, b)$$

- There is sampling variability ($\rightarrow$ choice of the data distribution)

Source of variation in prediction:

- We don't know the true value of the parameters and we specify a prior on it:

$$\theta \sim \mathrm{Beta}(a, b)$$

- There is sampling variability ($\rightarrow$ choice of the data distribution)

To account for the sources of variation we iterate the following steps:

1. Sample from the posterior distribution $\theta \sim p(\theta \mid y)$
2. Sample new values $y^* \sim p(y \mid \theta)$

- By repeating these steps a large number of times, we eventually obtain a reasonable approximation to the posterior predictive distribution.

Posterior Predictive Distribution (PPD)

- The PPD represents our uncertainty over the outcome of a future data collection, accounting for the observed data and model choice

- For the sake of prediction, the parameters are not of interest. They are vehicles by which the data inform about the predictive model

- The PPD averages over their posterior uncertainty

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta$$

- This properly accounts for parametric uncertainty

- The input is data, the output is a prediction distribution

Computation

Computing the PPD

- Say $\theta^{(1)}, \ldots, \theta^{(M)}$ are samples from the posterior

- If we make a sample for y^* for each $\theta^{(m)}$,

$$y^{*(m)} \sim p(y|\theta^{(m)})$$

  then the $y^{*(m)}$ are samples from the PPD

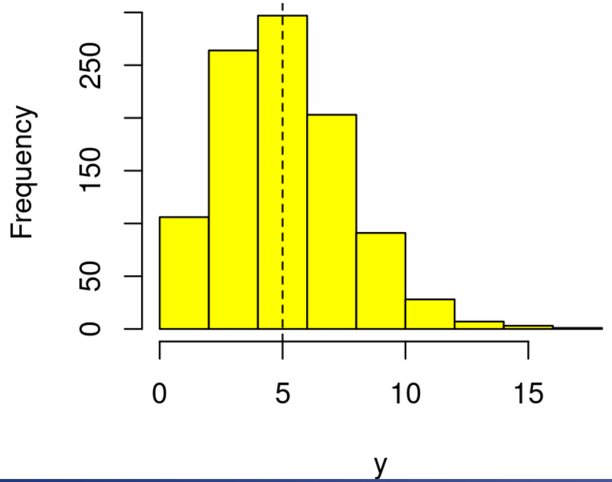
- The posterior predictive mean is approximated by the sample mean of the $y^{*(m)}$

- The probability that $y^* \geq 5$ is approximated by the sample proportion of the $y^{*(m)}$ that are equal or above 5

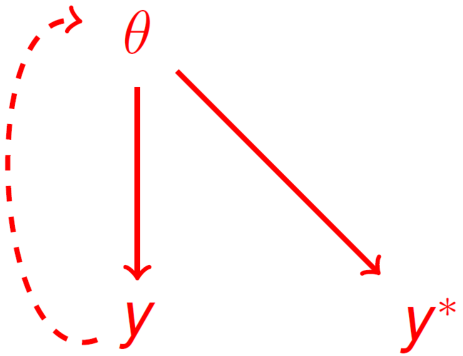
Example

- We estimate the prevalence of a disease in the UK population using a sample of $n = 58$ individuals.
- We find that $y = 10$ individuals have the diseases.
- What is the probability that, if we additionally sample `(k=30)` individuals this year, at least 5 will have the disease?

1. Likelihood: $y \sim \text{Binomial}(\theta, 58)$
2. Prior: $\theta \sim \text{Beta}(1, 1)$
3. Posterior: $\theta \mid y \sim \text{Beta}(10 + 1, 58 - 10 + 1)$
4. PPD: $y^* \sim \text{Binomial}(\theta \mid y, 30)$
5. $P(y \geq 5) = \sum_{j=5}^{30} P(y^* = j)$



$p(\theta \mid y)$



$p(y^* \mid y) = \int p(y^* \mid \theta) p(\theta \mid y) d\theta$

Introduction to Bayesian computing: Monte Carlo simulations

Bayesian computing

- In Session 2.1 we have introduced the concept of conjugacy, and we said that if the the prior and posterior come from the same family of distributions, the prior is said to be **conjugate** to the likelihood $\rightarrow$ the posterior is a known distribution.

- In real life it is (almost) impossible to use conjugacy so we need to resort to simulative approaches or approximations to perform computation:
  - Monte Carlo methods
  - Markov Chain Monte Carlo (MCMC) methods
  - Integrated Nested Laplace Approximation (INLA)

Monte Carlo simulation

- A Monte Carlo (MC) simulation is a randomly evolving simulation.
- MC sampling is based on the idea that if you have a large random sample from a certain distribution, the statistics that you can calculate in this sample (mean, standard deviation, percentiles...) will be very similar to the corresponding theoretical values in the distribution.
- If you have a complicated mathematical expression for a distribution and you cannot calculate algebraically important parameters, you could get the computer to generate a large random sample from such a distribution.
- By calculating the mean of that parameter in the sample you could estimate the mean in the original distribution with great precision.

Example: a Monte Carlo approach to estimating tail-areas of distributions

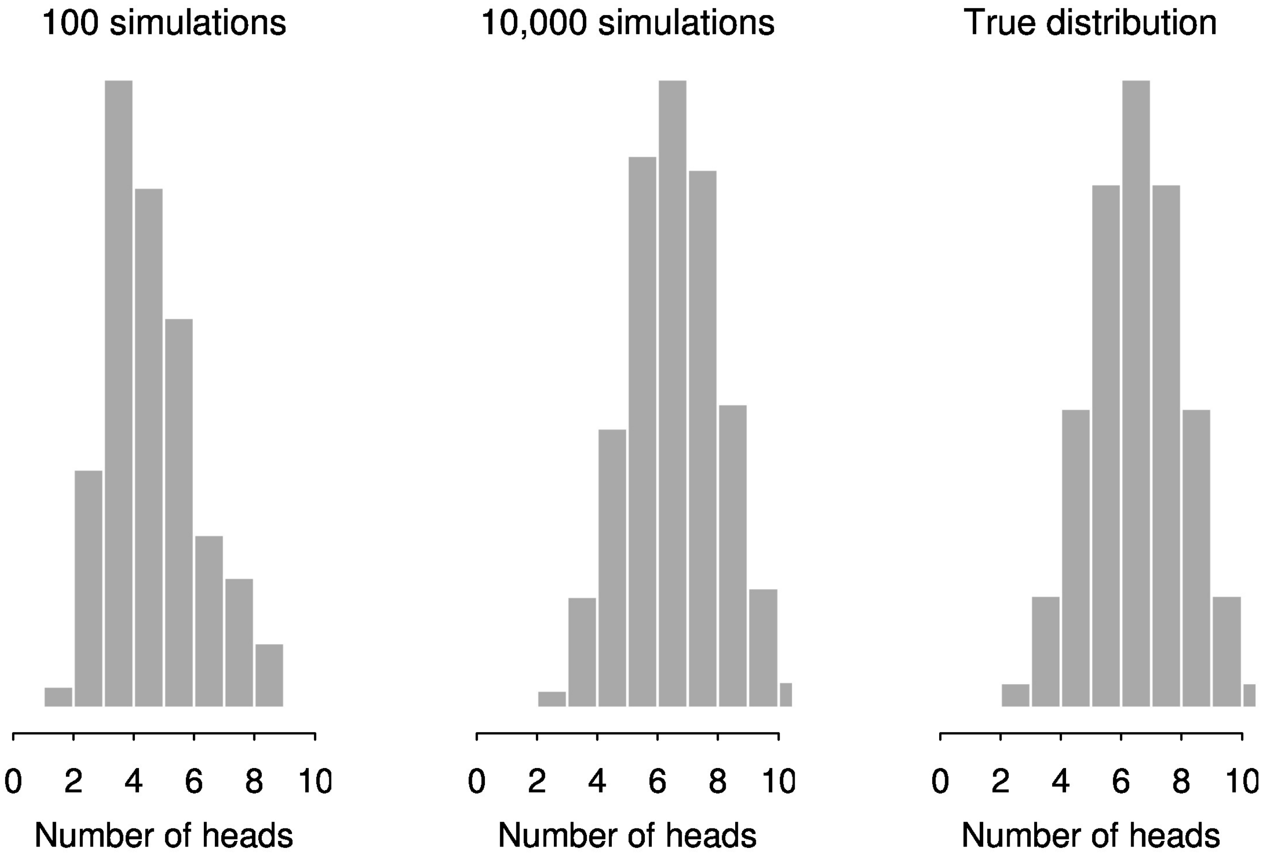
Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times.

- An *algebraic* approach would be:

$$\begin{aligned}
y &= \text{Number of heads} \\
y &\sim \text{Binomial}(\theta, n) \\
\Pr(\geq 8 \text{ heads}) &= \sum_{y=8}^{10} p\left(y \mid \theta = \frac{1}{2}, n = 10\right) \\
&= \binom{10}{8}\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{2} + \binom{10}{9}\left(\frac{1}{2}\right)^{9}\left(\frac{1}{2}\right)^{1} + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0} \\
&= 0.0547
\end{aligned}$$

- A *physical* approach would be to repeatedly throw a set of 10 coins and count the proportion of throws that there were 8 or more heads.

- A *simulation* approach uses a computer to toss the coins!



Proportion with 8 or more heads in 10 tosses: (a) After 100 throws (0.02); (b) after 10,000 throws (0.0577); (c) the true Binomial distribution (0.0547).

Monte Carlo approach to approximate log-odds

- We start with a Binomial likelihood

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

combined with a

$$\text{Beta}(a, b)$$

as prior for the probability of success θ.

- We are interested in the log-odds function of θ defined as

$$\log\left(\frac{\theta}{1-\theta}\right)$$

- The integral

$$\int_0^1 \log\left(\frac{\theta}{1-\theta}\right) p(\theta \mid y)\mathrm{d}\theta$$

cannot be computed analytically; we resort to Monte Carlo approximation.

Example of MC in practice

- We simulate m independent values $\left\{\theta^{(1)}, \ldots, \theta^{(m)}\right\}$ from the

$$\mathrm{Beta}(a_1 = y + a, b_1 = n - y + b)$$

posterior distribution using the property of conjugacy (Beta prior is conjugate to the Binomial likelihood).

- We apply the log-odds transformation to each value obtaining the set of values

$$\left\{\log\left(\frac{\theta^{(1)}}{1-\theta^{(1)}}\right), \ldots, \log\left(\frac{\theta^{(m)}}{1-\theta^{(m)}}\right)\right\}$$

- Finally, we compute the sample mean

$$\frac{\sum_{i=1}^{m} \log\left(\frac{\theta^{(i)}}{1-\theta^{(i)}}\right)}{m}$$

which is the Monte Carlo approximation to

$$\log\left(\frac{\theta}{1-\theta}\right)$$

Example of MC: R code

In R:

```
> a <- 1
> b <- 1
> theta <- rbeta(1,a,b)
> n <- 1000
> y <- rbinom(1, size=n, p=theta)
```

Example of MC: R code

In R:

```
> a <- 1
> b <- 1
> theta <- rbeta(1,a,b)
> n <- 1000
> y <- rbinom(1, size=n, p=theta)
```

- With this setting the exact posterior distribution of θ is

$$\text{Beta}(a_1 = a + y, b_1 = n - y + b)$$

- To approximate the log-odds, we simulate $m = 50000$ values from this Beta posterior distribution using the `rbeta` function.

```
> a1 <- a + y
> b1 <- n - y + b
> sim <- rbeta(n=50000, shape1=a1, shape2=b1)
> logodds <- log(sim/(1-sim))
```

The empirical distribution of the Monte Carlo sample is plotted below together with the exact posterior distribution of θ.