Bayesian Inference and Computation

Dr Rowland Seymour

Semester 2, 2023

Contents

Practicalities

0.1 Module aims

The module aims to give you an overview of the Bayesian paradigm. By the end of the course, you should

1. Be able to conceptualise a Bayesian approach for statistics
2. Be able to derive posterior and posterior predictive distributions for uni- and multivariate models
3. Identify suitable prior distributions and understand how the choice of prior distribution may affect the final result
4. Understand the principles of Markov Chain Monte Carlo and be able to construct an MCMC algorithm

0.2 Module structure

The module is split between theory and programming. Each week (excluding week 6) will have two lectures and two computer labs.

0.3 Assessment

The module is 55% coursework and 45% exam. The exam will last 1h 30m and take place during the summer exam period. More details about the coursework will be announced during the semester.

0.4 Getting help

There are lots of ways of getting help throughout the module. You can visit my office hour (Watson 317) on ….. or email me at r.g.seymour@bham.ac.uk. Each week, there will also be a problem class.

0.5 Recommended books and videos

No books are required for this course and the whole material is contained in these notes. However, you may find it useful to use other resources in your studies. I recommend the following:

1. A First Course in Bayesian Statistical Methods - Peter D. Hoff. This is a short book that covers the basics of Bayesian inference and computation. To the point and well written, it's a useful place to look topics up.
2. Bayesian Data Analysis - Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. This is a thorough book explaining everything you'd need to know to carry out Bayesian data analysis. It's a fairly look and in-depth book, but the authors are authoritative and give good advice throughout. Example code on the website is in R, Python and Stan.
3. Statistical Rethinking - Richard McElrath. This book provides a friendly intuitive understanding of Bayesian inference and computation. Aimed at social and natural scientists, it has less theory that the other two books but is perhaps more approachable. A set of video lectures for this book can be found on YouTube.

Chapter 1

Fundamentals of Bayesian inference

Bayesian inference is built on a different way of thinking about probability that methods you have learnt so far. In the past 30 years or so, Bayesian inference has become much more popular. This is partly due to increased computational power becoming available. In this first chapter, we are going to set out:

1. What are the fundamental principles of Bayesian inference?

2. What makes Bayesian inference different from other methods?

1.1 Statistical Inference

Three purpose of statistical inference is to "draw conclusions, from numerical data, about quantities that are not observed" (Bayesian Data Analysis, chapter 1). Generally speaking there are two kinds of inference: inference for quantities that are not possible to observe, and inference for quantities that unobserved or hasn't happened yet. We can apply our inference methods to a huge range of applications, from the effectiveness of drugs in a clinical trial, to estimating the number of victims of violence against women and girls.

In this module, we are going to look at a different way of carrying out statistical inference, one that doesn't depend on long running event. Instead, we're going to introduce the definition of probability that allows us to Interpret the subjective chance that an event occurs.

1.2 Frequentist Theory

Frequentist theory is built on the theory on long run events. Probabilities must be interpretable as frequencies over multiple repetitions of the experiment that is being analysed, and are calculated from the sampling distributions of measured quantities.

Definition 1.1. The long run relative frequency of an event is the **probability** of that event.

Example 1.1. For example, if a frequentist wanted to assign a probability to rolling a 6 on a particular dice, then they would roll the dice a large number of times and compute the relative frequency.

Definition 1.2. The **sampling distribution** of a statistic is the distribution based on a long run of samples of a fixed size from the population.

These two definitions form the basis of frequentist statistics and give rise to higher-level objects. However, due to the long run condition, these objects are often misunderstood.

Definition 1.3. Let X be a random sample from a probability distribution θ. A **95% confidence interval** for θ is an interval $(u(X), v(X))$ such that

$$P(u(x) < \theta < v(X)) = 0.95$$

This means that if you had an infinite number of samples for X and the corresponding infinite number of confidence intervals, then 95% of them would contain the true value of θ. It does *not* mean that there is a 0.95 probability a particular interval contains the true value of θ.

Example 1.2. Suppose that a ornithologist believes that for the sex of an offspring of a certain species of bird is equally likely to be male or female and is independent of the sex of any siblings. They selects a random sample of 100 broods with 4 offspring and counts the number of male offspring in each. The results are summarised in the following table.

```
offspring.df <- data.frame("males" = c(0, 1, 2, 3, 4),
                           "observed" = c(5, 20, 50, 20, 5))
knitr::kable(
  offspring.df, booktabs = TRUE,
  caption = 'The number of males in each of the broods.'
  )
```

If X denotes the number of male offspring then the scientist's null hypothesis H_0, is that $X \sim Bin(4, 1/2)$. This can be tested by calculating the so called χ^2 statistic - which measures how well the observations conform to the supposed distribution. This is computed by

Table 1.1: The number of males in each of the broods.

males	observed
0	5
1	20
2	50
3	20
4	5

```
expected <- 100*dbinom(offspring.df$males, 4, 0.5)
observed <- offspring.df$observed

## Chi sq = sum of (expected - observed)^2/expected
chi.sq <- sum((expected - observed)^2/expected)
```

Under H_0, over repeated sampling the distribution of the χ^2 statistic follows a χ^2 distribution on 4 degrees of freedom. Therefore, the probability of seeing something as extreme as this (the p-value) is

```
1 - pchisq(chi.sq, 4)
```

```
## [1] 0.1545873
```

Consider now a second scientist who is more open-minded than their col- league. They hypothesise only that $X \sim Bin(4, p)$ where p is unknown. To test their H_0 they first calculates the maximum likelihood estimator of p, $\hat{p} = 0.5$, computes the values of $E[X]$ using this estimate and calculates χ^2 to be 6.67 like their colleague.

However, because they are estimating p as part of the process of computing χ^2 over repeated sampling the distribution of their χ^2 is (approximately) a χ^2 distribution on 3 degrees of freedom. Their p-value is therefore

```
1 - pchisq(chi.sq, 3)
```

```
## [1] 0.08331631
```

We now have the somewhat confusing scenario whereby the second scientist finds that the evidence against their null hypothesis is stronger than that found by the first scientist against their null hypothesis, despite the second hypothesis being weaker (i.e. a logical consequence of the first). Thus, whatever a p-value

represents, it must never be interpreted as the probability that the null hypothesis is true conditional on the observed data. In this example it simply tells us how extreme the observed values of χ^2 are when compared to their distribution over many repetitions of the experiment.

1.3 Bayesian probability

The Bayesian paradigm is built about a different definition of probability. We are simply building our view of statistics upon a different set of axioms.

Definition 1.4. The subjective belief of an event is the **probability** of that event.

This definition means we can assign probabilities to events that frequentists do not recognise as valid.

Example 1.3. For example,

1. The probability that I vote for the labour party at the next election
2. A photo taken from the James Watt telescope contains a new planet.
3. The real identify of Banksy is Robin Gunningham.

These are not events that can be repeated in the long run.

1.4 Conditional Probability and Exchangability

Before we derive Bayes' theorem, we recap some important definitions in probability.

Definition 1.5. Given two events A and B, the **conditional probability** that event A occurs given the event B has already occurred is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

when $P(B) > 0$.

Definition 1.6. Two events A and B are **independent** given event C if

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C).$$

Definition 1.7. Let $p(y_1, \dots, y_N)$ be the joint density of $Y_1, \dots, Y_N$. If $p(y_1, \dots, y_N) = p(y_{\pi_1}, \dots, y_{\pi_N})$ for a permutations π of $\{1, \dots, N\}$, then $Y_1, \dots, Y_N$ are **exchangeable**.

Exchangability means that the labels of the random variables don't contain any information about the outcomes. This is an important idea in many areas of probability and statistics, and we often model exchangeable events as iid.

Example 1.4. If $Y_i \sim Bin(n, p)$ are independent and identically distributed for $i = 1, 2, 3$, then $p(Y_1, Y_2, Y_3) = p(Y_3, Y_1, Y_2)$.

Example 1.5. Let (X, Y) follow a bivariate normal distribution with mean $\mathbf{0}$, variances $\sigma_x = \sigma_y = 1$ and a correlation parameter $\rho \in [-1, 1]$. (X, Y) are exchangable, but only independent if $\rho = 0$.

Proposition 1.1. *If $\theta \sim p(\theta)$ and $(Y_1, \dots, Y_N)$ from a sample space $\mathcal{Y}$ are conditionally iid given some parameter θ, then marginally $Y_1, \dots, Y_N$ are exchangable.*

Proof. Suppose $(Y_1, \dots, Y_N)$ are conditionally iid given some parameter θ. Then for any permutation π of $\{1, \dots, N\}$ and observations $\{y_1, \dots, y_N\}$

$$\begin{aligned} p(y_1, \dots, y_N) &= \int p(y_1, \dots, y_N \mid \theta) p(\theta)\, d\theta \qquad \text{(definition of marginal distribution)} \\ &= \int \left\{ \prod_{i=1}^{N} p(y_i \mid \theta) \right\} p(\theta)\, d\theta \qquad \text{(definition of conditionally iid)} \\ &= \int \left\{ \prod_{i=1}^{N} p(y_{\pi_i} \mid \theta) \right\} p(\theta)\, d\theta \qquad \text{(product is commutative)} \\ &= p(y_{\pi_1}, \dots, y_{\pi_N}) \qquad \text{(definition of marginal distribution)} \end{aligned} \tag{1.1}$$

□

This tells us that if we have some conditionally iid random variables and a subjective prior belief about some parameter θ, then we have exchangeability. This is nice to have, but the implication in the other direction is much more interesting and powerful.

Theorem 1.1 (de Finetti)**.** *If a sequence of random variables $(Y_1, \dots, Y_N)$ from a sample space $\mathcal{Y}$ is exchangeable, then its joint distribution can be written as*

$$p(y_1, \dots, y_N) = \int \left\{ \prod_{i=1}^{N} p(y_i \mid \theta) \right\} p(\theta)\, d\theta$$

for some parameter θ, some distribution on θ, and some sampling model $p(y_i \mid \theta)$.

This is a kind of existence theorem for Bayesian inference. It says that if we have exchangeable random varibales, then a parameter θ must exist and a subjective probability distribution $p(\theta)$ must also exist. The argument against Bayesian inference is that it doesn't guarantee a *good* subjective probability distribution $p(\theta)$ exists.

1.5 Bayes' Theorem

Now we have an understanding of conditional probability and exchange ability, we can put these two together to understand Bayes' Theorem. Bayes' theorem is concerned with the distirbution of the parameter θ given some observed data y. It tries to answer the question: what does the data tell us about the model parameters?

Theorem 1.2 (Bayes)**.** *The distribution of the model parameter θ given the data y is*

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

Proof.

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} \tag{1.2}$$

$$\implies p(\theta, y) = p(\theta \mid y)p(y) \tag{1.3}$$

Analogously, using $\pi(y \mid \theta)$ we can derive

$$p(\theta, y) = p(y \mid \theta)p(\theta)$$

Putting these two terms equal to each other and dividing by $\pi(y)$ gives

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

□

There are four terms in Bayes' theorem:

1. The **posterior distribution** $p(\theta \mid y)$. This tells us our beleif about the model parameter θ given the data we have observed y.
2. The **likelihood function** $p(y \mid \theta)$. The likelihood function is common to both frequentist and Bayesian methods. By the likelihood principle, the likleihood funciton contains all the information the data can tell us about the model parameter θ.
3. The **prior distribution**.

Some history about Thomas Bayes.

Chapter 2

Programming in R

Chapter 3

Bayesian inference

Chapter 4

Sampling

4.1 Uniform random numbers

4.2 Inverse transform sampling

4.3 Rejection sampling

4.4 Monte Carlo

4.5 Markov Chain Monte Carlo

Chapter 5

Advanced computation