

## Lecture 6. Bayesian estimation

### Prior and posterior distributions

- By Bayes' theorem,

$$\pi(\theta | \mathbf{x}) = \frac{f_{\mathbf{x}}(\mathbf{x} | \theta) \pi(\theta)}{f_{\mathbf{x}}(\mathbf{x})},$$

where  $f_{\mathbf{x}}(\mathbf{x}) = \int f_{\mathbf{x}}(\mathbf{x} | \theta) \pi(\theta) d\theta$  for continuous  $\theta$ , and  $f_{\mathbf{x}}(\mathbf{x}) = \sum f_{\mathbf{x}}(\mathbf{x} | \theta_i) \pi(\theta_i)$  in the discrete case.

- Thus

$$\begin{aligned} \pi(\theta | \mathbf{x}) &\propto f_{\mathbf{x}}(\mathbf{x} | \theta) \pi(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior}, \end{aligned} \quad (1)$$

where the constant of proportionality is chosen to make the total mass of the posterior distribution equal to one.

- In practice we use (1) and often we can recognise the family for  $\pi(\theta | \mathbf{x})$ .
- It should be clear that the data enter through the likelihood, and so the inference is automatically based on any sufficient statistic.

### The parameter as a random variable

- So far we have seen the *frequentist* approach to statistical inference
- i.e. inferential statements about  $\theta$  are interpreted in terms of repeat sampling.
- In contrast, the Bayesian approach treats  $\theta$  as a *random variable* taking values in  $\Theta$ .
- The investigator's information and beliefs about the possible values for  $\theta$ , before any observation of data, are summarised by a **prior distribution**  $\pi(\theta)$ .
- When data  $\mathbf{X} = \mathbf{x}$  are observed, the extra information about  $\theta$  is combined with the prior to obtain the **posterior distribution**  $\pi(\theta | \mathbf{x})$  for  $\theta$  given  $\mathbf{X} = \mathbf{x}$ .
- There has been a long-running argument between proponents of these different approaches to statistical inference
- Recently things have settled down, and Bayesian methods are seen to be appropriate in huge numbers of application where one seeks to assess a probability about a 'state of the world'.
- Examples are spam filters, text and speech recognition, machine learning, bioinformatics, health economics and (some) clinical trials.

### Inference about a discrete parameter

Suppose I have 3 coins in my pocket,

- biased 3:1 in favour of tails
- a fair coin,
- biased 3:1 in favour of heads

I randomly select one coin and flip it once, observing a head. What is the probability that I have chosen coin 3?

- Let  $X = 1$  denote the event that I observe a head,  $X = 0$  if a tail
- $\theta$  denote the probability of a head:  $\theta \in (0.25, 0.5, 0.75)$
- Prior:  $p(\theta = 0.25) = p(\theta = 0.5) = p(\theta = 0.75) = 0.33$
- Probability mass function:  $p(x | \theta) = \theta^x (1 - \theta)^{(1-x)}$

Coin	$\theta$	Prior $p(\theta)$	Likelihood $p(x=1 \theta)$	Un-normalised Posterior $p(x=1 \theta)p(\theta)$	Normalised Posterior $\frac{p(x=1 \theta)p(\theta)}{p(x)^\dagger}$
1	0.25	0.33	0.25	0.0825	0.167
2	0.50	0.33	0.50	0.1650	0.333
3	0.75	0.33	0.75	0.2475	0.500
Sum		1.00	1.50	0.495	1.000

$\dagger$  The normalising constant can be calculated as  $p(x) = \sum_i p(x|\theta_i)p(\theta_i)$

So observing a head on a single toss of the coin means that there is now a 50% probability that the chance of heads is 0.75 and only a 16.7% probability that the chance of heads is 0.25.

### Example 6.1

Suppose we are interested in the true mortality risk  $\theta$  in a hospital  $H$  which is about to try a new operation. On average in the country around 10% of people die, but mortality rates in different hospitals vary from around 3% to around 20%. Hospital  $H$  has no deaths in their first 10 operations. What should we believe about  $\theta$ ?

- Let  $X_i = 1$  if the  $i$ th patient dies in  $H$  (zero otherwise),  $i = 1, \dots, n$ .
- Then  $f_{\mathbf{x}}(\mathbf{x}|\theta) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$ .
- Suppose a priori that  $\theta \sim \text{Beta}(a, b)$  for some known  $a > 0$ ,  $b > 0$ , so that  $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ ,  $0 < \theta < 1$ .
- Then the posterior is

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto f_{\mathbf{x}}(\mathbf{x}|\theta)\pi(\theta) \\ &\propto \theta^{\sum x_i + a - 1} (1-\theta)^{n - \sum x_i + b - 1}, \quad 0 < \theta < 1.\end{aligned}$$

We recognise this as  $\text{Beta}(\sum x_i + a, n - \sum x_i + b)$  and so

$$\pi(\theta|\mathbf{x}) = \frac{\theta^{\sum x_i + a - 1} (1-\theta)^{n - \sum x_i + b - 1}}{B(\sum x_i + a, n - \sum x_i + b)} \quad \text{for } 0 < \theta < 1.$$

□

## Bayesian inference - how did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

### P R O B L E M.

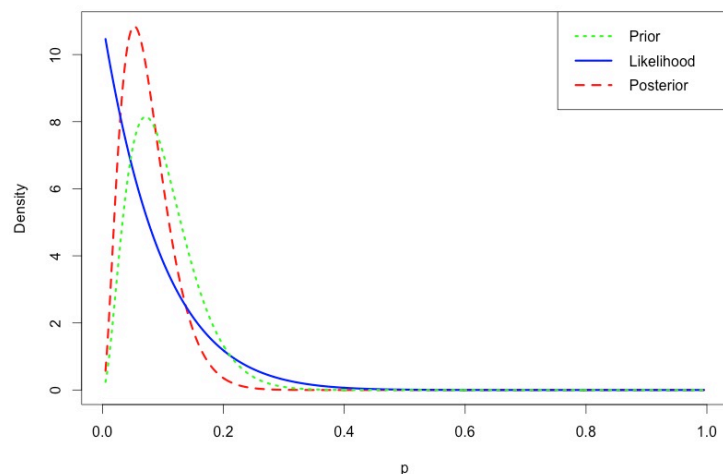
*Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.*

In modern language, given  $r \sim \text{Binomial}(\theta, n)$ , what is  $\mathbb{P}(\theta_1 < \theta < \theta_2 | r, n)$ ?

- In practice, we need to find a Beta prior distribution that matches our information from other hospitals.
- It turns out that a  $\text{Beta}(a=3, b=27)$  prior distribution has mean 0.1 and  $\mathbb{P}(0.03 < \theta < 0.20) = 0.9$ .
- The data is  $\sum x_i = 0$ ,  $n = 10$ .
- So the posterior is  $\text{Beta}(\sum x_i + a, n - \sum x_i + b) = \text{Beta}(3, 37)$
- This has mean  $3/40 = 0.075$ .
- NB Even though nobody has died so far, the mle  $\hat{\theta} = \sum x_i / n = 0$  (i.e. it is impossible that any will ever die) does not seem plausible.

```
install.packages("LearnBayes")
library(LearnBayes)
prior = c( a= 3, b = 27 )           # beta prior
data = c( s = 0, f = 10 )          # s events out of f trials
triplot(prior,data)
```

Bayes Triplot, beta( 3 , 27 ) prior, s= 0 , f= 10



## Bayesian approach to point estimation

- Let  $L(\theta, a)$  be the loss incurred in estimating the value of a parameter to be  $a$  when the true value is  $\theta$ .
- Common loss functions are quadratic loss  $L(\theta, a) = (\theta - a)^2$ , absolute error loss  $L(\theta, a) = |\theta - a|$ , but we can have others.
- When our estimate is  $a$ , the expected posterior loss is  $h(a) = \int L(\theta, a)\pi(\theta|\mathbf{x})d\theta$ .
- The **Bayes estimator**  $\hat{\theta}$  **minimises the expected posterior loss**.
- For **quadratic loss**

$$h(a) = \int (a - \theta)^2 \pi(\theta|\mathbf{x}) d\theta.$$

- $h'(a) = 0$  if

$$a \int \pi(\theta|\mathbf{x}) d\theta = \int \theta \pi(\theta|\mathbf{x}) d\theta.$$

- So  $\hat{\theta} = \int \theta \pi(\theta|\mathbf{x}) d\theta$ , the **posterior mean**, minimises  $h(a)$ .

## Conjugacy

- For this problem, a beta prior leads to a beta posterior. We say that the beta family is a **conjugate** family of prior distributions for Bernoulli samples.
- Suppose that  $a = b = 1$  so that  $\pi(\theta) = 1$ ,  $0 < \theta < 1$  - the uniform distribution (called the "principle of insufficient reason" by Laplace, 1774).
- Then the posterior is  $\text{Beta}(\sum x_i + 1, n - \sum x_i + 1)$ , with properties.

	mean	mode	variance
prior	1/2	non-unique	1/12
posterior	$\frac{\sum x_i + 1}{n + 2}$	$\frac{\sum x_i}{n}$	$\frac{(\sum x_i + 1)(n - \sum x_i + 1)}{(n + 2)^2 (n + 3)}$

- Notice that the mode of the posterior is the mle.
- The posterior mean estimator,  $\frac{\sum x_i + 1}{n + 2}$  is discussed in Lecture 2, where we showed that this estimator had smaller mse than the mle for non-extreme values of  $\theta$ . Known as Laplace's estimator.
- The posterior variance is bounded above by  $1/(4(n + 3))$ , and this is smaller than the prior variance, and is smaller for larger  $n$ .
- Again, note the posterior automatically depends on the data through the sufficient statistic.

- For **absolute error loss**,

$$\begin{aligned} h(a) &= \int |\theta - a| \pi(\theta|\mathbf{x}) d\theta = \int_{-\infty}^a (a - \theta) \pi(\theta|\mathbf{x}) d\theta + \int_a^{\infty} (\theta - a) \pi(\theta|\mathbf{x}) d\theta \\ &= a \int_{-\infty}^a \pi(\theta|\mathbf{x}) d\theta - \int_{-\infty}^a \theta \pi(\theta|\mathbf{x}) d\theta \\ &\quad + \int_a^{\infty} \theta \pi(\theta|\mathbf{x}) d\theta - a \int_a^{\infty} \pi(\theta|\mathbf{x}) d\theta \end{aligned}$$

Now  $h'(a) = 0$  if

$$\int_{-\infty}^a \pi(\theta|\mathbf{x}) d\theta = \int_a^{\infty} \pi(\theta|\mathbf{x}) d\theta.$$

- This occurs when each side is 1/2 (since the two integrals must sum to 1) so  $\hat{\theta}$  is the **posterior median**.

## Example 6.2

Suppose that  $X_1, \dots, X_n$  are iid  $N(\mu, 1)$ , and that a priori  $\mu \sim N(0, \tau^{-2})$  for known  $\tau^{-2}$ .

- The posterior is given by

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto f_{\mathbf{x}}(\mathbf{x}|\mu)\pi(\mu) \\ &\propto \exp\left[-\frac{1}{2}\sum (x_i - \mu)^2\right] \exp\left[-\frac{\mu^2\tau^2}{2}\right] \\ &\propto \exp\left[-\frac{1}{2}(n + \tau^2)\left\{\mu - \frac{\sum x_i}{n + \tau^2}\right\}^2\right] \quad (\text{check}).\end{aligned}$$

- So the posterior distribution of  $\mu$  given  $\mathbf{x}$  is a Normal distribution with mean  $\sum x_i / (n + \tau^2)$  and variance  $1 / (n + \tau^2)$ .
- The normal density is symmetric, and so the posterior mean and the posterior median have the same value  $\sum x_i / (n + \tau^2)$ .
- This is the optimal Bayes estimate of  $\mu$  under both quadratic and absolute error loss.

## Example 6.3

Suppose that  $X_1, \dots, X_n$  are iid  $\text{Poisson}(\lambda)$  rv's and that  $\lambda$  has an exponential distribution with mean 1, so that  $\pi(\lambda) = e^{-\lambda}$ ,  $\lambda > 0$ .

- The posterior distribution is given by

$$\pi(\lambda|\mathbf{x}) \propto e^{-n\lambda} \lambda^{\sum x_i} e^{-\lambda} = \lambda^{\sum x_i} e^{-(n+1)\lambda}, \quad \lambda > 0,$$

ie  $\text{Gamma}(\sum x_i + 1, n + 1)$ .

- Hence, under quadratic loss,  $\hat{\theta} = (\sum x_i + 1) / (n + 1)$ , the posterior mean.
- Under absolute error loss,  $\hat{\theta}$  solves

$$\int_0^{\hat{\theta}} \frac{(n+1)^{\sum x_i + 1} \lambda^{\sum x_i} e^{-(n+1)\lambda}}{(\sum x_i)!} d\lambda = \frac{1}{2}.$$