Lecture 6. Bayesian estimation

The parameter as a random variable

- So far we have seen the *frequentist* approach to statistical inference
- i.e. inferential statements about θ are interpreted in terms of repeat sampling.
- In contrast, the Bayesian approach treats θ *as a random variable* taking values in Θ.
- The investigator's information and beliefs about the possible values for θ, 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 θ is combined with the prior to obtain the **posterior distribution** $\pi(\theta \mid \mathbf{x})$ for θ 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.

Prior and posterior distributions

- By Bayes' theorem,

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

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

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

  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 $\pi(\theta \mid \mathbf{x})$.
- It should be clear that the data enter through the likelihood, and so the inference is automatically based on any sufficient statistic.

Inference about a discrete parameter

Suppose I have 3 coins in my pocket,

1. *biased 3:1 in favour of tails*
2. *a fair coin,*
3. *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
- θ 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: $p(x|\theta) = \theta^x(1-\theta)^{(1-x)}$

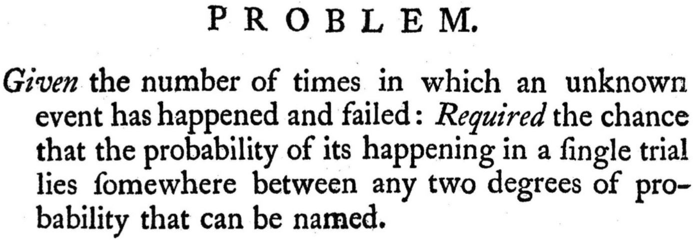
		Prior	**Likelihood**	**Un-normalised Posterior**	**Normalised Posterior**
Coin	θ	$p(\theta)$	$p(x=1\|\theta)$	$p(x=1\|\theta)p(\theta)$	$\frac{p(x=1\|\theta)p(\theta)}{p(x)^\dagger}$
1	0.25	0.33	0.25	0.0825	0.167
2	0.50	0.33	0.50	0.1650	0.333
3	0.75	0.33	0.75	0.2475	0.500
	Sum	1.00	1.50	0.495	1.000

† 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 in 0.25.

Bayesian inference - how did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote



PROBLEM.

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 $r \sim$ Binomial(θ, n), what is $\mathbb{P}(\theta_1 < \theta < \theta_2 | r, n)$?

Example 6.1

Suppose we are interested in the true mortality risk θ 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 θ?

- Let $X_i = 1$ if the ith patient dies in H (zero otherwise), $i = 1, \dots, n$.
- Then $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$.
- 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

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

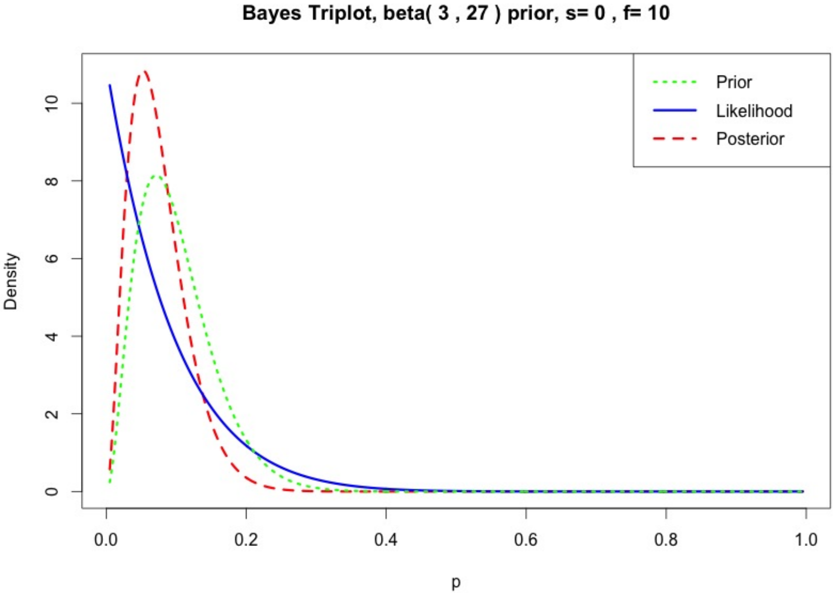
We recognise this as $\text{Beta}(\sum x_i + a, n - \sum x_i + b)$ and so

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

□

- In practice, we need to find a Beta prior distribution that matches our information from other hospitals.
- It turns out that a 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 Beta($\sum x_i + a, n - \sum x_i + b$) = 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
Prior
Likelihood
Posterior
Density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
p

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 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 θ. 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.

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 θ.
- 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 $h(a)$.

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

- The posterior is given by

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

- So the posterior distribution of μ 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 μ under both quadratic and absolute error loss.

Example 6.3

Suppose that $X_1, \ldots, X_n$ are iid Poisson(λ) rv's and that λ 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 Gamma$(\sum x_i + 1, n + 1)$.
- Hence, under quadratic loss, $\hat{\lambda} = (\sum x_i + 1)/(n+1)$, the posterior mean.
- Under absolute error loss, $\hat{\lambda}$ solves

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