

Bayesian Methods - Continuous Functions

Dr. David Lucy

d.lucy@lancaster.ac.uk

Lancaster University

Bayesian Methods - p.1/33

Bayesian inference

So far:

- We have looked at examples of the use of Bayes' theorem which involve discrete observations and events.
- Very useful for scientists and people concerned with propositions and data.
- **Not really what statisticians mean by Bayesian methods though.**

Bayesian Methods - p.2/33

Let us examine an example:

- It has been observed that the proportion of female births is not 0.5 in human populations.
- A local maternity ward recorded 98 female live births and 105 male live births in 2006.

What inferences can we make about the probability of a female live birth from all live births?

Bayesian Methods - p.3/33

Proportion of female births

Conventionally:

- The probability of a female live birth could be given by the expectation $98/(98 + 105) \approx 0.48$.

Is this true?

- Not really. It is a bit naive.
- Frequentist estimate could be $se = \sqrt{\frac{p(1-p)}{N}}$, where p is the observed proportion, and N the sample size.
- **Problems with what the resultant CI means and how to use it.**

Bayesian Methods - p.4/33

Proportion of female births

Bayesian methods offer different solutions.

- Consideration of the parameter of interest, denoted by θ , in this case θ is the probability of a female live birth, as in itself a random variable.
- Information about the nature of this random variable comes from the sample.

How do we go about calculating a suitable distribution for θ ?

Bayesian Methods – p.5/33

Notation

- Let θ be the parameter of interest with prior function $\pi(\theta)$.
- The observations are $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, and have a marginal density function $f(\mathbf{x})$.
- The likelihood function $f(\mathbf{x}|\theta)$ is the distribution of \mathbf{x} conditional on specific values of θ .

Then:

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

Note: π is used to denote functions of θ , and f functions of the data.

Bayesian Methods – p.6/33

Likelihood function

Critical to Bayesian inference are likelihood functions are:

- the likelihoods describing the observations given parameter, which,
- serve to link the sampling space to the parameter space.

The probability:

$$f(\mathbf{x}|\theta) = L(\theta) = \prod_i f(x_i|\theta)$$

is a general expression for a likelihood function.

Bayesian Methods – p.7/33

Data marginal

- The function $f(\mathbf{x})$ is the marginal distribution of the data.
- For most inference problems we are not that interested in \mathbf{x} , but in θ .

Because of this we can write the posterior distribution as:

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{1}{f(\mathbf{x})} \pi(\theta)L(\theta) \\ &\propto \pi(\theta)L(\theta)\end{aligned}$$

as $f(\mathbf{x})$ can be regarded as a normalising constant.

Bayesian Methods – p.8/33

Kernels

The kernel of a function:

- Probability density functions of a random variable \mathbf{X} usually have the form $cg(\mathbf{x})$.
- c can be regarded as a normalising function which is there simply to ensure $\int g(\mathbf{x}) = 1$.

$g(\mathbf{x})$ is the kernel of a density, or mass function, and is the bit which has data.

Kernels

A gamma pdf is:

$$f(\mathbf{x}|\lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha \mathbf{x}^{\alpha-1} e^{-\lambda \mathbf{x}}$$

The bits which have data in are of interest, so the kernel is:

$$f(\mathbf{x}|\lambda, \alpha) \propto \mathbf{x}^{\alpha-1} e^{-\lambda \mathbf{x}}$$

as \mathbf{x} is a function solely of λ and α .

Bernoulli likelihood

- Returning to our data.
- The observation of 98 female live births from 203 live births is a series of Bernoulli trials with $\mathbf{x} = \{0, 1\}$. For any single trial the probability mass function is:

$$f(x_i|\theta) = \theta^{x_i} (1 - \theta)^{1-x_i}$$

- The likelihood function is therefore:

$$f(\mathbf{x}|\theta) = \prod_i \theta^{x_i} (1 - \theta)^{1-x_i}$$

Bernoulli likelihood

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_i \theta^{x_i} (1 - \theta)^{1-x_i} \\ &= [\theta^{x_1} (1 - \theta)^{1-x_1}] \times [\theta^{x_2} (1 - \theta)^{1-x_2}] \dots [\theta^{x_n} (1 - \theta)^{1-x_n}] \\ &= \theta^{x_1} \theta^{x_2} \dots \theta^{x_n} (1 - \theta)^{1-x_1} (1 - \theta)^{1-x_2} \dots (1 - \theta)^{1-x_n} \\ &= \theta^{\sum_i x_i} (1 - \theta)^{(1-x_1)+(1-x_2)+\dots+(1-x_n)} \\ &= \theta^{\sum_i x_i} (1 - \theta)^{n - \sum_i x_i} \end{aligned}$$

Bernoulli likelihood

- The likelihood for a series of iid Bernoulli trials can be modelled:

$$L(\theta) = \theta^{\sum_i x_i} (1 - \theta)^{n - \sum_i x_i}$$

- As for a series of $\sum x_i$ where $\mathbf{x} = \{1, 0\}$ is the number of “successes”, and n the number of trials, then $n - \sum x_i$ can be seen as the number of “failures”.
 $\sum x_i$ can be denoted s .
- This is the kernel of the *Beta* distribution. θ is restricted such that $0 \leq \theta \leq 1$.

Bayesian Methods – p.13/33

Posterior distribution

- Let a suitable prior, $\pi(\theta)$, also be distributed *Beta*.
- The parameters for $\pi(\theta)$ may be denoted $a - 1$, and $b - 1$. These are called hyperparameters to distinguish them from the parameters of the sampling space.

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \pi(\theta)f(\mathbf{x}|\theta) \\ &= c [\theta^{a-1}(1-\theta)^{b-1}] \times [\theta^s(1-\theta)^{n-s}] \\ &= c [\theta^{a-1}\theta^s(1-\theta)^{b-1}(1-\theta)^{n-s}] \\ &= c \theta^{a+s-1}(1-\theta)^{b+n-s-1}\end{aligned}$$

Bayesian Methods – p.14/33

Beta summaries

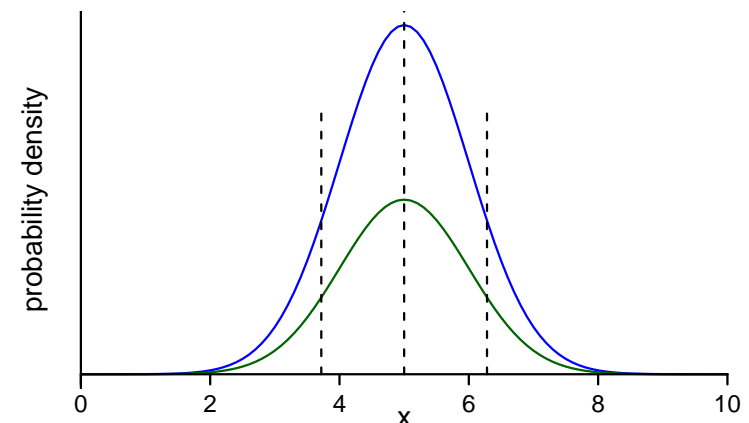
- We have the kernel of the posterior distribution, and in many cases this would be sufficient as, for the *Beta* distribution there are a number of summaries which are readily available:

- $E(\theta) = (a + s)/(a + b + n)$
- mode = $(a + s - 1)/(a + b + n - 1)$
- var(θ) = $(\alpha\beta)/(\alpha + \beta)^2(\alpha + \beta + 1)$

where: $\alpha = a + s$, and $\beta = b + n - s$

Bayesian Methods – p.15/33

Normalising constant



Bayesian Methods – p.16/33

Normalising constant

- It may not be absolutely necessary, but in the case of the β we do know.
- In the case of the β distribution constant c is:

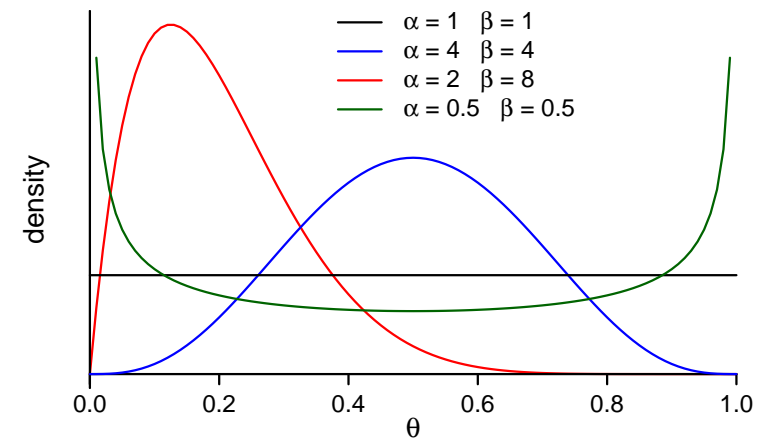
$$c = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)}$$

which, iff α and β take on integer values:

$$c = \frac{(a + b + n - 1)!}{(a + s - 1)!(b + n - s - 1)!}$$

Bayesian Methods – p.17/33

The β distribution



Bayesian Methods – p.19/33

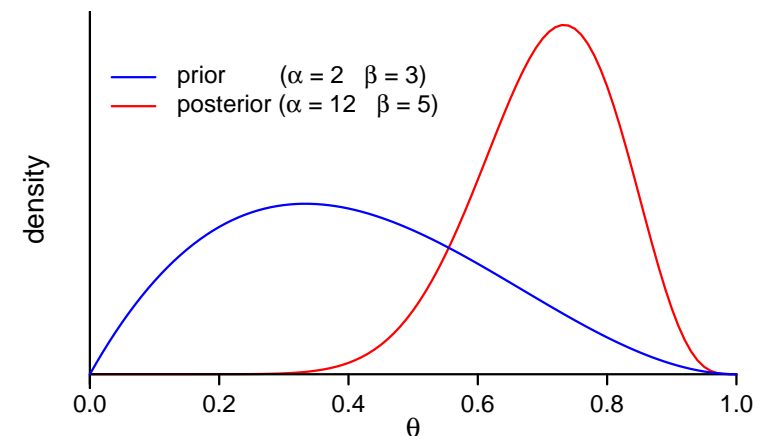
The β distribution

The β distribution is uniquely suited to providing $\pi(\theta)$ if θ is a proportion.

1. Any non-zero values for $\pi(\theta)$ are confined to $\{0, 1\}$.
2. It can take on a large variety of shapes to represent different ideas for $\pi(\theta)$.
3. Only two parameters, α and β , which can be taken to represent the number of “successes” and “failures” from a run of n Bernoulli trials.
4. Experts can define prior information in terms of “successes” and “failures”.

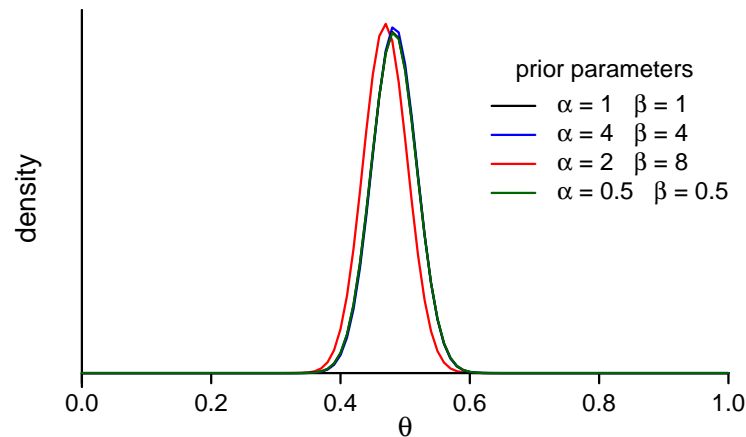
Bayesian Methods – p.18/33

The β distribution



Bayesian Methods – p.20/33

Female live births



Bayesian Methods – p.21/33

Inferences about means

- Inferences about proportions important, and useful, part of Bayesian statistical approaches.
- Now investigate methods for making inferences about parameters for other distributions.
- Most univariate discrete distributions relatively simple as plausible prior distribution functions tend to be of exactly the same form as their likelihood functions.
- The prior distribution is then said to be a “prior conjugate” distribution.

Bayesian Methods – p.22/33

Normal distribution

Assume \mathbf{X} is distributed with known variance, such that $\mathbf{X} \sim N(\mu, \sigma^2)$. Suppose we are interested in $\mu = \{-\infty \leq \mu \leq \infty\}$ then, assuming a uniform prior for μ :

$$\begin{aligned} \pi(\mu|\mathbf{x}) &\propto \pi(\mu)L(\mu) \\ &= \exp\left\{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right\} \times 1 \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(n\mu^2 - 2n\mu\bar{x} + n\bar{x}^2)\right\} \\ &= \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right\} \end{aligned}$$

Bayesian Methods – p.23/33

Normal distribution

$$\begin{aligned} \sum (x_i - \mu)^2 &= \sum (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum [(x_i - \bar{x}) + (\bar{x} - \mu)][(x_i - \bar{x}) + (\bar{x} - \mu)] \\ &= \sum [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2] \end{aligned}$$

Because $\sum (x_i - \bar{x}) = 0$, and as we are thinking about this as a function of μ then we can concentrate on the right hand term of above as the $(x_i - \bar{x})^2$ bit has no μ term in it.

$$\begin{aligned} \sum (x_i - \mu)^2 &\propto \sum (\bar{x} - \mu)^2 \\ &= \sum (\bar{x}^2 - 2\bar{x}\mu + \mu^2) \\ &= n\bar{x}^2 - 2n\bar{x}\mu + n\mu^2 \\ &= n\mu^2 - 2n\bar{x}\mu + n\bar{x}^2 \\ &= n(\mu - \bar{x})^2 \end{aligned}$$

Bayesian Methods – p.24/33

Normal distribution

- Leaves us with a normal kernel:

$$\pi(\mu|\mathbf{x}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right\}$$

- The posterior density function for $\mu \sim N(\bar{x}, \sigma^2/n)$, (the $n/2\sigma^2$ gives us the variance of σ^2/n).
- Very similar to the standard conventional result for the “standard error of the mean”, where the mean is \bar{x} , with standard error σ/\sqrt{n} .
- Has advantage of straightforward Bayesian interpretation for $(1 - \alpha)$ interval.

Bayesian Methods – p.25/33

Normal distribution σ known

Usual to use an “informative prior” for normal distribution.

Let us say: $\mathbf{x} \sim N(\mu_1, \sigma_1)$, and, $\pi(\mu_1) \sim N(\mu_0, \sigma_0)$.

The posterior is derived:

$$\begin{aligned}\pi(\mu_1|\mathbf{x}) &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu_1 - \mu_0)^2\right\} \times \exp\left\{-\frac{1}{2\sigma_1^2}(\mu_1 - \bar{x})^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mu_1 - \mu)^2\right\}\end{aligned}$$

Which can be recognised as $\mu_1 \sim N(\mu, \sigma^2)$.

Bayesian Methods – p.26/33

Normal distribution σ known

$$\pi(\mu|\mathbf{x}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\mu_1 - \mu)^2\right\}$$

Where:

$$\mu = \frac{(1/\sigma_0^2)\mu_0 + (n/\sigma_1^2)\bar{x}}{(1/\sigma_0^2) + (n/\sigma_1^2)} = w\mu_0 + (1-w)\bar{x}$$

If:

$$w = \frac{1/\sigma_0^2}{1/\sigma_0^2 + n/\sigma_1^2}$$

Bayesian Methods – p.27/33

Normal distribution σ known

The posterior density function:

$$\pi(\mu|\mathbf{x}) \propto \exp\left\{-\frac{1}{2\sigma^2}(\mu_1 - \mu)^2\right\}$$

Where:

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma_1^2}$$

For which the precision can be seen as a sum of the precisions of the prior and data.

Bayesian Methods – p.28/33

Normal distribution

- Using a non-informative (uniform) prior yields exactly equivalent result to classical approaches.
- An informative prior, in the case above another normal, gives us a posterior distribution which is some weighted sum of the parameters of the prior and data.
- Classical approaches are in some senses a subset of Bayesian methods.
- **Can use similar approaches for when the posterior variance cannot be assumed known.**

Bayesian Methods – p.29/33

Posterior summaries

- HDR region also known as a “credible interval”.
- Select the smallest interval for the parameter of interest which contains $(1 - \alpha) \times 100\%$ of the area.
- Interpretation is: there is a probability of $(1 - \alpha) \times 100\%$ that the parameter of interest falls within the region.
- Classical Confidence Interval is interpreted as: in the long run $(1 - \alpha) \times 100\%$ of the observations of the parameter of interest will fall within the CI.

Bayesian Methods – p.31/33

Posterior summaries

- Point estimation all very well and useful, but only for specific posterior distributions can points and variances be sufficient statistics.
- More general form of summary is the $(1 - \alpha)$ highest density region (HDR).
- Analogous to the classical “confidence interval”.

Bayesian Methods – p.30/33

Posterior summaries

- Bayesian credible intervals far more intuitive for the scientist to interpret.
- Are open for further manipulation.
- As such are a far more powerful device to quantify uncertainty.

Bayesian Methods – p.32/33

Credible intervals

- For Normal posterior probability density functions the HDR is easily calculated.
- No so for other density functions, especially non-parametric posterior probability density functions such as kernel density functions.
- Numerical integration and computer intensive methods prevail.
- The subject of all our sessions tomorrow.