## Session 2

# Bayesian Methods

Iraj Kazemi

i.kazemi@lancaster.ac.uk

Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YF, England.

March 10-11, 2005



## Bayesian Inference

- Suppose we have a sample of $n$ variables $X' = (X_1, \cdots, X_n)$, generated from a density function $f(x|\theta)$.
- We assume that the cases are independent given $\theta$, and hence the joint probability density of the sample is

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

- The likelihood function $L(\theta) = f(\mathbf{x}|\theta)$ plays a central role in classical methods.
- Under ML estimation, we would compute the mode (the maximal value of $L$, as a function of $\theta$ given the data $\mathbf{x}$) of the likelihood function.
- For Bayesian methods, the likelihood function is the instrument to pass from the prior density $\pi(\theta)$ to the posterior density $\pi(\theta|\mathbf{x})$ via Bayes' Theorem.

## Bayesian Point Estimation

- How do we extract a Bayes estimator for some unknown parameter $\theta$?
- There are a number of candidates:
- We could follow ML and use the **mode of the distribution** (its maximal value), with

$$\hat{\theta} = \max_{\theta} \pi(\theta|x)$$

- The **median of the posterior distribution**, where the estimator satisfies $Pr(\theta > \hat{\theta}|x) = Pr(\theta < \hat{\theta}|x) = 0.5$, hence

$$\int_{\hat{\theta}}^{\infty} \pi(\theta|x)d\theta = \int_{-\infty}^{\hat{\theta}} \pi(\theta|x)d\theta = \frac{1}{2}.$$

- We could take the **expected value** of $\theta$ given the data,

$$\hat{\theta} = E(\theta|x) = \int \theta\pi(\theta|x)d\theta$$

- This estimate is defined as the **posterior Bayes estimate** of $\theta$ with respect to the prior $\pi(\theta)$.
- The full form of the posterior distribution is not easy to obtain for some complex models, but
- it may still be possible to obtain one of the three above estimators.
- We can generally obtain the posterior by simulation using **Gibbs** sampling, and hence the Bayes estimate can be found.

## Inferences for a Binomial Probability

- Let $\theta$ denotes the proportion of people in England with genotype $Z$.
- Consider the binary responses $X_1, \cdots, X_n$, where, for $i = 1, \cdots, n$,
- $X_i = 1$ if the $i$th person in the sample possesses the genotype $Z$, and $X_i = 0$ otherwise.
- The number of persons with genotype $Z$ in the sample, $S = \sum_i X_i$, has a binomial distribution with probability $\theta$ and sample size $n$.
- Assume now that based upon a random sample of size $n = 2,500$, a sample geotype frequency $s = 50$ is observed.
- the ML estimate of $\theta$ is $\frac{s}{n} = 0.02$ with the standard error $\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.0028$.

- Assuming the prior distribution of $\theta$ is $Beta\,(a, b)$, the posterior density would be $Beta\,(a + s, b + n - s)$.
- Suppose that the prior mean is $E(\theta) = 0.06$, and the precision of the prior distribution, $a + b$, is 400.
- It follows from $E(\theta) = \frac{a}{a+b}$ that $a = 24$ and $b = 376$,
- i.e., $\theta \sim Beta(24, 376)$ and $\theta|\mathbf{x} \sim Beta(74, 2826)$.
- The posterior mean is

$$E\,(\theta|\mathbf{x}) = \frac{a+s}{a+b+n} = 0.0255$$

and the posterior variance

$$var(\theta|\mathbf{x}) = \frac{E(\theta|x)\,[1 - E(\theta|x)]}{a+b+n+1} = (0.00293)^2$$

therefore, the posterior s.d. of $\theta$ is 0.00293.

## WinBUGS program

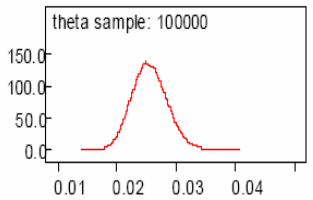
```
model{
for(i in 1:N) {y[i]~dbin(theta,2500)}
 theta~dbeta(24,376)
}
#data
list(y=c(50), N=1)
#Initial values
list(theta=850)
```

- Line 2 specifies that the variable **y** is distributed binomially with the parameter **theta** and $n$=**2500**.
- In WinBUGS distributional relationships are described by the ~symbol
- Line 3 specifies the prior for theta.
- Comments (starting with #) are inserted for ease of reading.
- WinBUGS borrows its notation from S-plus using the convention **c(..)** to represent a vector of observations.



## WinBUGS Results

| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| theta | 0.02551 | 0.002928 | 9.656E-6 | 0.02011 | 0.02539 | 0.03156 | 1 | 100000 |



**Inference for Poisson Data**

- Suppose that the sample $X_1, X_2, \cdots, X_n$ be i.i.d. $Poisson(\theta)$.
- The ML estimator of $\theta$ is the sample mean $\overline{X}$ :

$$\hat{\theta}_{MLE} = \overline{X}$$

- Assuming the prior distribution of $\theta$ is Gamma$(a, b)$, the prior mean is

$$E(\theta) = \frac{a}{b}.$$

- We can show that the posterior distribution is $Gamma(\sum x_i + a, n + b)$.
- The posterior Bayes estimate of $\theta$ with respect to the gamma prior is

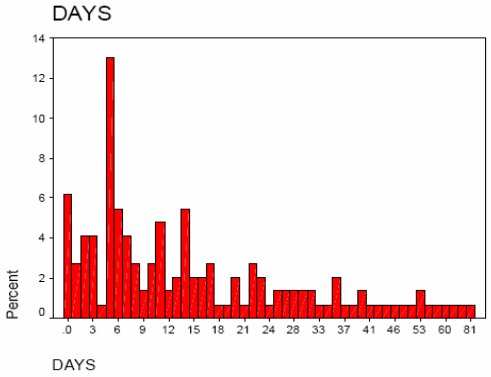
$$\hat{\theta}_{Bayes} = E(\theta|\mathbf{x}) = \frac{\sum x_i + a}{n + b} = w\overline{x} + (1 - w) E(\theta)$$

where $w = \frac{n}{n + b}$.

the posterior mean = weighted mean of data value and prior mean.

**Example: Poisson Data**

- Suppose that the number of days $(X)$ absent from school during the school year for each child is $Poisson$ with mean $\theta$.
- Assume that the prior distribution of $\theta$ is $Gamma(1, 0.04)$.
- $E(\theta) = \frac{1}{\theta} = 25.$
- Based on a sample size of $n = 146$ we wish to estimate $\theta$.




**Example: Poisson Data (cont.)**

- It can be found that $\sum_i x_i = 2403$, thus $\hat{\theta}_{MLE} = 16.459$,
- Since $var(X) = \theta$, then $var\left(\overline{X}\right) = \frac{\theta}{n}$, and $std\left(\hat{\theta}\right) = \sqrt{\frac{\theta}{n}} = 0.336.$
- **The posterior Bayes estimate of** $\theta$ :

$$\begin{aligned} \hat{\theta}_{Bayes} &= \frac{\sum x_i + a}{n + b} = \frac{2403 + 1}{146 + 0.04} = 16.461 \\ var(\theta|\mathbf{x}) &= \frac{2403 + 1}{(146 + 0.04)^2} = 0.1127 \\ &\Rightarrow sd(\theta|\mathbf{x}) = \sqrt{0.1127} = 0.336 \end{aligned}$$



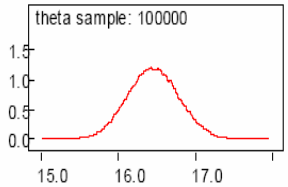
**WinBUGS program:**

```
model{
 for(i in 1:N){days[i]~dpois(theta)}
 theta~dgamma(1,0.04)
}
#initial value
list(theta=16.)
#data
list(N=146, days=c( 2, 11,...,22,37))
```

- Lines 2 specifies that the distribution of observations is Poisson with a mean theta.
- Line 3 specifies that the prior for theta is gamma.

## WinBUGS Results:

| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| theta | 16.46 | 0.3363 | 0.001082 | 15.81 | 16.46 | 17.13 | 1 | 100000 |






## Bayesian Interval Estimation

- Posterior mean and mode provide simple summaries of the posterior distribution.
- It can be further useful to find a region that contains $\theta$ with a specified probability $1-\alpha$.
- Given the posterior distribution, construction of confidence intervals is obvious.
- For example, a $100(1-\alpha)\%$ *Bayesian confidence interval* is given by any $(L_{\alpha/2}, H_{\alpha/2})$ satisfying

$$\int_{L_{\alpha/2}}^{H_{\alpha/2}} \pi(\theta|x)d\theta = 1-\alpha.$$



- Bayesian confidence intervals are also called **credible intervals**.
- Shortest Bayesian confidence regions are called posterior **highest density regions** ($HDR$s), i.e.,
- Regions with the smallest volume in the parameter space.
- For a single parameter, $\theta$, the region reduces to the interval.



**Example: Normal with known variance**

- Conditional on $\theta$, consider a random sample $X_1, X_2, \cdots, X_n$ drawn from a $N(\mu, \sigma^2)$ distribution with known $\sigma^2$.
- Suppose $\pi(\mu) \propto 1$.
- The posterior distribution of $\mu$ is given by

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

- Thus $\mu|\mathbf{x} \sim N\left(\overline{x}, \frac{\sigma^2}{n}\right)$. A $(1-\alpha)\%$ HPD interval is

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where for $Z \sim N(0,1)$, $Pr\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}$.

- This HPD interval is identical to the classical $(1-\alpha)\%$ confidence interval for the mean of a normal population when the variance is known.
- However, the meaning is different.
- The interpretation of the $(1-\alpha)\%$ confidence interval is based on a repeatation of the sampling process, so that if we could take, say, $100$ samples, we would expect that in $(1-\alpha)\%$ of times the interval $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ contains the true value of $\mu$.
- With the Bayesian $HDR$, conditional on the information currently available we believe that, with probability $(1-\alpha)$, $\theta$ belongs to the interval $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.



## Bayesian Hypothesis Testing

- In the classical hypothesis testing framework, we have two alternatives.
- The null hypothesis $H_0$ that the unknown parameter $\theta$ belongs to some set or interval $\Theta_0$ $(\theta \in \Theta_0)$, versus
- the alternative hypothesis $H_1$ that $\theta$ belongs to the alternative set $\Theta_1$ $(\theta \in \Theta_1)$.
- Two sets $\Theta_0$ and $\Theta_1$ contain no common elements $(\Theta_0 \cap \Theta_1 = \varnothing)$ and the union of $\Theta_0$ and $\Theta_1$ contains the entire space of values for $\theta$ (i.e., $\Theta_0 \cup \Theta_1 = \Theta$).



- In the classical statistical framework, we use the observed data to test the significant of a particular hypothesis, and compute a p-value.
- In a Bayesian framework, as using the posterior distribution

$$Pr(\theta > \theta_0) = \int_{\theta_0} \pi(\theta|x)d\theta$$

and

$$Pr(\theta_0 < \theta < \theta_1) = \int_{\theta_0}^{\theta_1} \pi(\theta|x)d\theta$$



- To operationalize, let

$$p_0 = Pr(\theta \in \Theta_0|\mathbf{x}); \quad p_1 = Pr(\theta \in \Theta_1|\mathbf{x})$$

denote the posterior probability that $\theta$ is in the null ($p_0$) and alternative ($p_1$) hypothesis sets.
- Since $\Theta_0 \cap \Theta_1 = \varnothing$ and $\Theta_0 \cup \Theta_1 = \Theta$, it follows that $p_0 + p_1 = 1$.
- For the prior probabilities we have

$$\pi_0 = Pr(\theta \in \Theta_0); \quad \pi_1 = Pr(\theta \in \Theta_1)$$

- Thus the **prior odds** of $H_0$ versus $H_1$ are $\pi_0/\pi_1$, while the **posterior odds** are $p_0/p_1$.

- The **Bayes factor** $B_0$ in favor of $H_0$ versus $H_1$ is given by the ratio of the posterior odds divided by the prior odds,

$$B_0 = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0\pi_1}{\pi_0 p_1} \;\; or \;\; B_0 = \frac{p_0\,(1-\pi_0)}{\pi_0\,(1-p_0)}$$

- By symmetry note that the Bayes factor $B_1$ in favor of $H_1$ versus $H_0$ is just

$$B_1 = 1/B_0$$



- When the hypotheses are simple, say $\Theta_0 = \theta_0$ and $\Theta_1 = \theta_1$, then for $i = 0, 1$,

$$p_i \propto \pi(\theta_i)L(\theta_i) = \pi_i L(\theta_i)$$

Thus

$$\frac{p_0}{p_1} = \frac{\pi_0 L(\theta_0)}{\pi_1 L(\theta_1)}$$

and the Bayes factor (in favor of the null) reduces the

$$B_0 = \frac{L(\theta_0)}{L(\theta_1)}$$

which is simply a *likelihood ratio*.



**Example: Poisson Distribution**

- Let $X_1, \cdots, X_n$ be i.i.d. Poisson with mean $\theta$. Thus, the likelihood function

$$L(\theta) \propto \theta^{\sum x_i} e^{-n\theta},$$

- Let $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ be two simple hypotheses, with $p(H_0) = p(H_1)$.
- The Bayes factor is

$$B_0 = \left(\frac{\theta_0}{\theta_1}\right)^{\sum x_i} e^{n(\theta_1 - \theta_0)}$$

and hence, since the prior odds are equal to 1, the decision rule is to accept $H_0$ if the Bayes factor is greater than 1.



**Example: Normal Distribution**

- Suppose $X_1, \cdots, X_n$ are i.i.d. $N(\theta, 1)$, and we wish to test

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta = 1 \end{cases}$$

- Then Bayes factor is

$$\begin{aligned} B_0 &= \frac{(2\pi)^{-n/2}\, e^{-\frac{1}{2}\sum x_i^2}}{(2\pi)^{-n/2}\, e^{-\frac{1}{2}\sum (x_i-1)^2}} \\ &= e^{\left(\frac{n}{2} - \sum x_i\right)} \end{aligned}$$

- Suppose $n = 10$ and $\sum x_i = 4.5$. Then

$$B_0 = e^{\frac{10}{2} - 4.5} = e^{0.5} = 1.65,$$

which is weak evidence in favour of $H_0 : \theta = 0$.

- Note if $\sum x_i = 4.5$, then $B_0 = e^{5-1} = e^4 = 55$, which is strong evidence in favour of $H_0 : \theta = 0$.

## A Primer on MCMC and The Gibbs Sampler

**What is a Markov Chain?**

- Let $\theta_t$ denote the value of a random variable at time $t$, and let the state space refer to the range of possible $\theta$ values.
- The random variable $\theta$ is a Markov process if the transition probabilities between different values in the state space depend only on the random variable's current state, i.e.,

$$Pr(\theta_t|\theta_{t-1}, \cdots, \theta_1, \theta_0) = Pr(\theta_t|\theta_{t-1})$$

- Thus for a Markov random variable the only information about the past needed to predict the future is the current state of the random variable.



- A *Markov chain* refers to a sequence of random variables $(\theta_0, \cdots, \theta_n)$ generated by a Markov process.
- A particular chain is defined most critically by its *transition probabilities*, $Pr(i, j) = Pr(i \rightarrow j)$, which is the probability that a process at state space $s_i$ moves to state $s_j$ in a single step,

$$Pr(i, j) = Pr(\theta_t = s_j|\theta_{t-1} = s_i).$$



**A Simple Example:**

**A discrete vote choice between two political parties**

- Suppose that voters that normally select $\theta_1$ have an $80\%$ chance of continuing to do so, and
- voters that normally select $\theta_2$ have only a $40\%$ chance of continuing to do so.
- Since there are only two choices, this leads the transition matrix $P$:

$$\text{current period} \begin{cases} \theta_1 \\ \theta_2 \end{cases} \overbrace{\begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix}}^{\text{next period}} \quad \begin{matrix} \\ \theta_1 \;\; \theta_2 \end{matrix}$$



- All Markov chains begin with a starting point assigned by the researcher.
- This initial state defines the proportion of individuals selecting $\theta_1$ and $\theta_1$ before beginning the chain. Consider the starting point:

$$S_0 = \begin{bmatrix} 0.50 & 0.5 \end{bmatrix}.$$

- That is, before running the Markov chain $50\%$ of the observed population select each alternative.
- For the first state we simply multiply the initial state by the transition matrix:

$$S_1 = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \end{bmatrix}.$$

- So after the first iteration we have the new proportions: $70\%$ select $\theta_1$ and $30\%$ select $\theta_2$.

- This process continues multiplicatively as long as we like:

$$\text{second state} : S_2 = \begin{bmatrix} 0.7 & 0.3 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.74 & 0.26 \end{bmatrix}.$$

$$\text{third state} : S_3 = \begin{bmatrix} 0.74 & 0.26 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.748 & 0.252 \end{bmatrix}.$$

$$\text{fourth state} : S_4 = \begin{bmatrix} 0.748 & 0.253 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.7496 & 0.2504 \end{bmatrix}$$

- As you might guess, the choice proportions are converging to $[0.75, 0.25]$.
- This is because the transition matrix is pushing toward a steady state or more appropriately "stationary" distribution of the proportions.
- The operation of running a Markov chain until it reaches its stationary distribution is exactly the process employed in $MCMC$.

**The Gibbs Sampler**

- The aim is to specify how to construct a Markov Chain whose values converge to the target distribution.
- To introduce the Gibbs sampler, consider a bivariate random variable $(X, Y)$, and suppose we wish to compute one or both marginals, $f(x)$ and $f(y)$.
- The idea behind the sampler is that
- it is so easy to consider a sequence of conditional distributions, $f(x|y)$ and $f(y|x)$, than obtain marginal densities by integration of the joint density $f(x, y)$, e.g., $f(x) = \int p(x, y)dy$.



(1) The sampler starts with some initial value $y_0$ for $y$ and obtains $x_0$ by generating a random variable from the conditional distribution $f(x|y = y_0)$.

(2) The sampler then uses $x_0$ to generate a new value of $y_1$, drawing from the conditional distribution based on the value $x_0$, $f(y|x = x_0)$.

(3) The sampler proceeds as follows

$$\begin{aligned} x_t &\sim f(x|y = y_{t-1}) \\ y_t &\sim f(y|x = x_t) \end{aligned}$$

(4) Repeating this process $T$ times, generates a *Gibbs sequence* of length $T$, where a subset of points $(x_t, y_t)$ for $1 \leq t \leq m < T$ are taken as our simulated draws from the full joint distribution.

- The Gibbs sequence converges to a stationary distribution that is independent of the starting values, and by construction this stationary distribution is the target distribution we are trying to simulate.



**Example**

- Suppose the joint distribution of $(X, \theta)$ is given by

$$\begin{aligned} f(x, \theta) &= \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \\ \text{for } x &= 0, 1, \cdots, n \text{ and } 0 < \theta < 1 \end{aligned}$$

- We denote that the conditional distribution of $X|\theta \sim Bin(n, \theta)$, while $\theta|x \sim Beta(x + \alpha, n - x + \beta)$.
- By computing a sequence of a binomial and then a beta we can compute marginal distributions.

- Suppose $\alpha = 1, \beta = 2$, and $n = 10$.
- Start the sampler with (say) $\theta_0 = 1/2$,

(i) $x_0$ is obtained by generating a random $Bin(n, \theta_0) = Bin(10, 1/2)$ random variable, giving $x_0 = 5$ in our simulation.

(ii) $\theta_1$ is obtained from a $Beta(x_0 + \alpha, n - x_0 + \beta) = Beta(5 + 1, 10 - 5 + 2)$ random variable, giving $\theta_1 = 0.33$.

(iii) $x_1$ is a realization of a $Bin(n, \theta_1) = Bin(10, 0.33)$ random variable, giving $x_1 = 3$.

(iv) $\theta_2$ is obtained from a $Beta(x_1 + \alpha, n - x_1 + \beta) = Beta(3 + 1, 10 - 3 + 2)$ random variable, giving $\theta_2 = 0.56$.

(v) $x_2$ is obtained from a $Bin(n, \theta_2) = Bin(10, 0.56)$ random variable, giving $x_2 = 0.7$.

- So our realization of the Gibbs sequence after three iterations is $(5, 0.5), (3, 0.33), (7, 0.56)$.
- We can continue this process to generate a chain of the desired length.

This suggests an iterative algorithm of the following form:

- Starting with an initial estimate $\left(\theta_1^{(0)}, \cdots, \theta_k^{(0)}\right)$

(1) Draw $\theta_1^{(1)}$ at random from $f\left(\theta_1 | \mathbf{x}, \theta_2^{(0)}, \cdots, \theta_k^{(0)}\right)$

(2) Draw $\theta_2^{(1)}$ at random from $f\left(\theta_2 | \mathbf{x}, \theta_1^{(1)}, \theta_3^{(0)}, \cdots, \theta_k^{(0)}\right)$; and so on down to

k. Draw $\theta_k^{(1)}$ at random from $f\left(\theta_k | \mathbf{x}, \theta_1^{(1)}, \cdots, \theta_{k-1}^{(1)}\right)$

- This is only one iteration of the Gibbs sampler.
- Repeat until convergence to the stationary distribution.



# End of Session