

# Introduction to Bayesian Statistics and Computational Methods

Antonietta Mira

Swiss Finance Institute

University of Lugano, Switzerland

Cotonou, Benin, March 2013

Thanks to the organizers  
for the invitation

*Thanks to Kerrie Mengersen (QUT Brisbane)  
for a draft version of the teaching material*

# Aim of the course

- Provide basic concepts of the **Bayesian approach**
  - Subjective interpretation of probability
  - Use of Bayes theorem in updating information
  - Type of prior distributions
- Introduce the use of **Bayesian methods** for data analysis
- Introduce to **Monte Carlo (MC)** and **Markov chain Monte Carlo (MCMC)** simulations

- Three approaches to Probability
  - **Axiomatic** (Kolmogorov)
    - Probability by definition and properties
  - **Relative Frequency** (classical, objective)
    - Repeated trials
  - **Degree of belief** (Bayesian, subjective)
    - Personal measure of uncertainty
- Problems
  - The chance that the next Italian government will succeed
  - The probability of rain today

# Bayes Theorem

- Thomas Bayes
- Published works by Bayes
- Background on probabilities
- Bayes Theorem
- Applications

# Thomas Bayes

- Born in **1702**, London
- Little childhood information
- Presbyterian Minister
- In 1742, elected fellow by the Royal Society of London
- Retired in 1752
- Died in April of **1761**



# Written Work

- Only two works published during his life
  - *Divine Benevolence* (1731)
  - *Introduction to the Doctrine of Fluxions* (1736)
- He never published his mathematical works
- “The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the chance of the thing expected upon it’s happening”

# Publishing of Bayes Theorem



- **Richard Price** examined Bayes' work after his death
- Responsible for the communication to the Royal Society on Bayes' work
- *An Essay Toward Solving a Problem in the Doctrine of Chances*



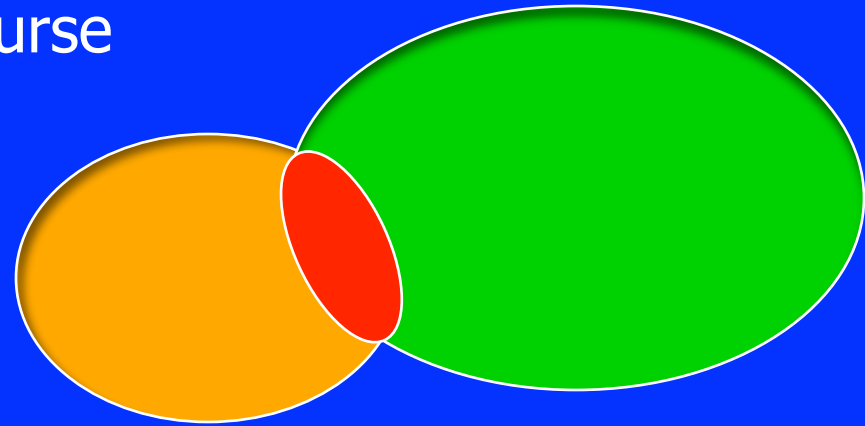
# *An Essay Toward Solving a Problem in the Doctrine of Chances*

- a version of what becomes Bayes Theorem
- *the definition of conditional probability*
- ***If  $P(B) > 0$ , the conditional probability of A given B, denoted by  $P(A | B)$ , is***

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

Among the conference participants,

- 70% attended this morning course
- 55% attended my course and
- 45% attended both



If a randomly selected participant attended this morning course, what is the probability he or she also attended my course?

$$P(\text{afternoon} \mid \text{morning}) = \frac{P(\text{morning and afternoon})}{P(\text{morning})}$$

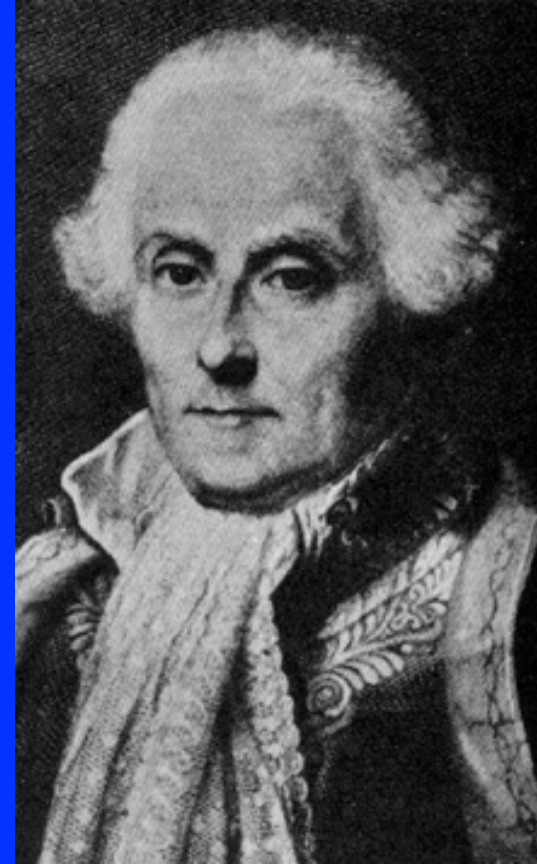
$$= .45 / .7 \approx .6429 \text{ or } 64.29\%$$

- **Pierre Simon Laplace**

French mathematician

Responsible for current form of Bayes Theorem

- **Bayes** found the probability that  $x$  is between two values given a number of successes and failures
- **Laplace** found an expression for the probability of a number of future successes and future failures given the number of successes and failures
- **Richard von Mises** states, “We owe Bayes only the statement of the problem and the principle of the solution. The theorem itself was first formulated by Laplace”

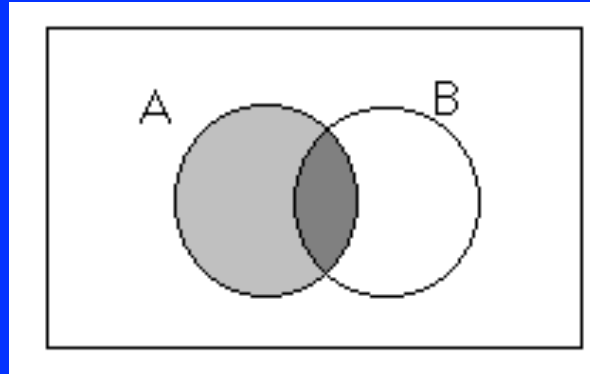


# Law of Total Probability

- Sometimes it is not possible to calculate  $P(A)$  however, it may be possible to find  $P(A | B)$  and  $P(A | B^c)$  for some event  $B$
- *Let  $B$  be an event with  $P(B) > 0$  and  $P(B^c) > 0$   
Then for any event  $A$ :*

$$P(A) = P(A | B) P(B) + P(A | B^c) P(B^c)$$

- We know that  $P(A) = P(AB) + P(AB^c)$



Substitute in the conditional probability

$$P(AB) = P(A | B)P(B)$$

$$P(AB^c) = P(A | B^c)P(B^c)$$

The Law of Total Probability then becomes:

$$P(A) = P(A | B)P(B) + P(A | B^c)P(B^c)$$

An insurance company rents

- 40% of the cars for its customers from agency I
- 60% from agency II
  
- If 6% of the cars from agency I
- and 5% of the cars from agency II  
break down

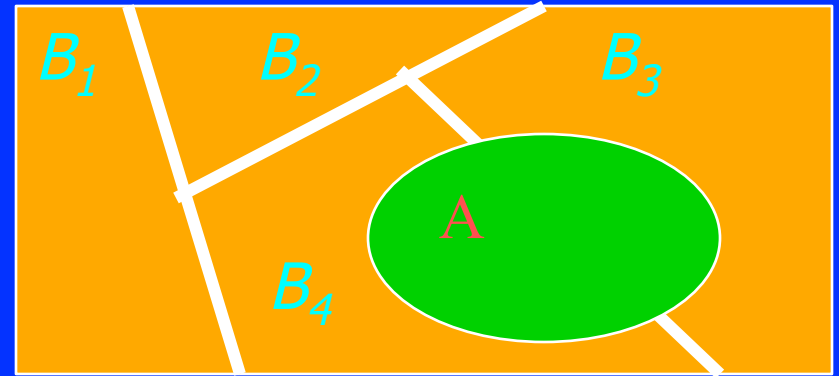
What is the probability that a car rented by this company breaks down?

$$\begin{aligned} P(\text{car rented by insurance breaks down}) &= (.4)(.06) + (.6)(.05) \\ &= .054 = 5.4\% \end{aligned}$$



# Bayes Theorem

Consider the sample space  $S$  of an experiment and a partition  $\{B_1, B_2, \dots, B_n\}$ , with  $P(B_i) > 0$ , for  $i = 1, 2, \dots, n$



Then, for any *event*  $A$  of  $S$ , with  $P(A) > 0$

$$P(B_k | A) = \frac{P(A | B_k)P(B_k)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \dots + P(A | B_n)P(B_n)}$$

$P(B_k)$  is the prior probability

$P(B_k | A)$  is the posterior probability

**LTP**

- Proof is very simple

- $P(A | B) = \frac{P(AB)}{P(B)}$  , conditional probability

- Rearranged becomes:

- $P(AB) = P(B) P(A | B)$

- $P(BA) = P(A) P(B | A)$

- Therefore  $P(B)P(A | B) = P(A)P(B | A)$

- Solve for the  $P(B | A)$

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)}$$



- A box contains 7 red and 13 blue balls.
- Two balls are selected at random and are discarded without their colors being seen.
- If a third ball is drawn randomly and observed to be red, what is the probability that both of the discarded balls were blue?

- Solve  $P(BB | R)$

$$\begin{aligned}
 &= \frac{P(R | BB)P(BB)}{P(R | BB)P(BB) + P(R | BR)P(BR) + P(R | RR)P(RR)} \\
 &= \frac{\frac{7}{18} * \frac{39}{95}}{\frac{7}{18} * \frac{39}{95} + \frac{6}{18} * \frac{91}{190} + \frac{5}{18} * \frac{21}{190}} \approx 0.46
 \end{aligned}$$

# Applications

- Diagnostic testing
  - Tests identify if a person has a particular disease or not
  - Tests are not always 100% correct
  - If a person tests positive for a disease is he or she truly sick?
- A cancer is found in 1 in every 2000 people. If a person has the disease w.p. 90% the test will result positive. If a person does not have the disease, the test will result in a false positive 1% of the time.
- The probability that a person with a positive test really has cancer:

$$P(\text{Cancer} \mid \text{Positive Test}) = \frac{\frac{1}{2000} * .90}{\frac{1}{2000} * .90 + \frac{1999}{2000} * .01} \approx .043$$

# Introduction to Bayesian Statistical Inference

- Statistical inference is used to draw conclusions from known data in samples to populations for which data is unknown
- **EXAMPLE:** Find the probability that an African man's height is over 1.75 meters

- If we have no info about this man, the probability is based on the proportion of Africans taller than 1.75 meters
  - Frequentist approach
- However, if we have prior knowledge about the man, it must be factored into the probability
  - If he plays basketball, the probability will be larger than population proportion
  - Bayesian approach

# Frequentist approach to modelling

We have some data  $Y$ , and want to know about  $\theta$

$\theta$  can be unknown parameters, missing data, latent variables, etc. Eg: *sample of data from a normal distribution, what is the population mean?*

Frequentist: estimate  $\theta$  through the likelihood:  $p(Y|\theta)$

*How likely is  $Y$  for given values of  $\theta$ ?*

*Use moment estimators or maximum likelihood.*

But we *really* want to know about  $p(\theta|Y)$

# Bayesian approach to modelling

$$p(\theta|Y) = p(Y|\theta) p(\theta) / p(Y)$$

$p(\theta)$  is the *prior* for  $\theta$


*What do we know about  $\theta$  independently of the data?*

$$p(Y) = \sum p(\theta) p(y|\theta) \quad \text{or} \quad \int p(\theta) p(y|\theta) d\theta$$

i.e., the probability of the data for all values of  $\theta$   
*(constant - calculate analytically or numerically)*

# Why Bayes?

Bayesian methods allow us to:

- Think differently about interpreting and estimating parameters (unknown  random)  
“what are possible values of this parameter, based directly on the posterior distribution  $p(\theta|Y)$ ?”
- Combine prior information with the data  
“what else do I know about  $\theta$ ?”
- Describe many sources of uncertainty in the model  
“how sure am I about the inputs to my model?”
- Describe complex systems using hierarchical or multi-level models

# Why Bayes?

## Bayesian computational methods

(such as MCMC) allow us to:

- Use non-standard distributions as LHD
- Fit very complex non-linear models
- Obtain a very wide variety of estimates
- Make a very wide range of inferences, based directly on posterior probabilities  
(CI, HP, Prediction)
- Avoids averaging over the unobserved values of  $\mathbf{x}$
- Coherent update of the info on  $\theta$

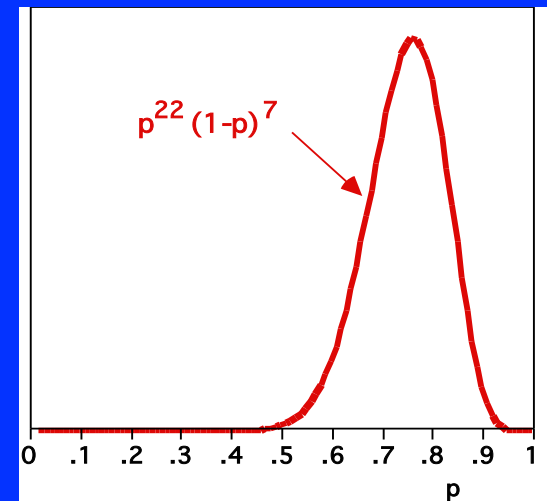


# Example: Estimating a proportion

- **Data:** Suppose that we observe  $n=29$  sites with  $Y=22$  presences of a species (and 7 absences). What is the probability  $\theta$  of presence of the species?

(Or: Suppose that we have  $n=29$  patients;  $Y=22$  survive and 7 die. What is the probability  $p$  of survival?)

- **Unobserved:**  $\theta$  probability of success
- **Likelihood:**  $Y$  has a Binomial distribution



$$p(Y|\theta) \sim \text{Bin}(29, \theta)$$
$$p(Y=22|\theta) \propto \theta^{22}(1-\theta)^7$$

- **Posterior:**  $P(Y|\theta) \propto \theta^y (1-\theta)^{n-y}$

# The Beta Distribution

*Prior for  $\theta$* : Many choices:

- point estimate
- Beta distribution (continuous over range 0,1)

$$\theta \sim \text{Beta}(a,b)$$

$$p(\theta) \propto \theta^{a-1} (1-\theta)^{b-1}$$

$$E(\theta) = a/(a+b) \quad \text{“unbiased Bayesian est.”}$$

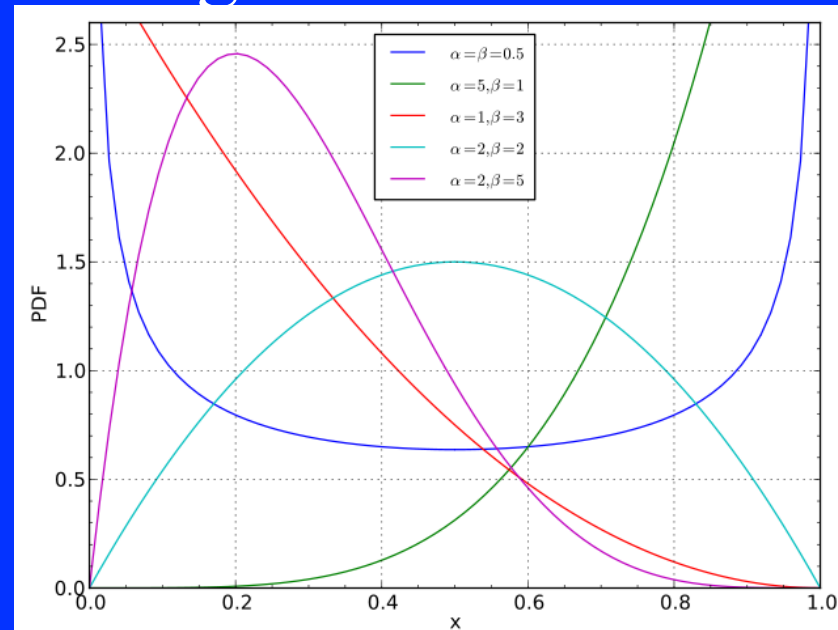
$$\text{Var}(\theta) = ab/\{(a+b)^2(a+b+1)\}$$

$$\text{Mode}(\theta) = (a-1)/(a+b-2) \quad \text{“Bayesian MLE”}$$

$$a=b \implies \text{mean} = 0.5 \implies \text{symmetric}$$

# Beta distribution

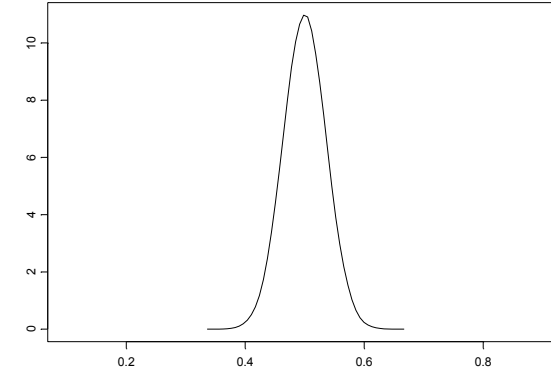
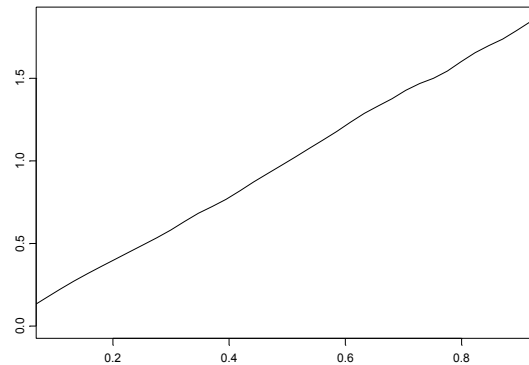
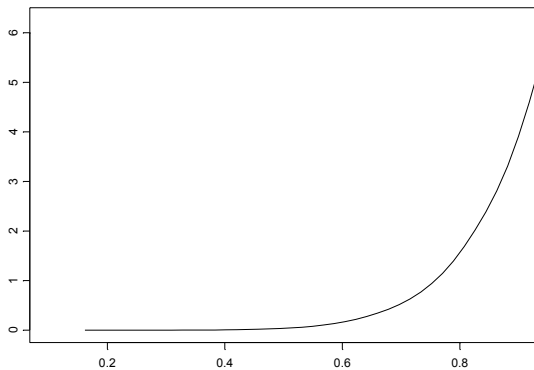
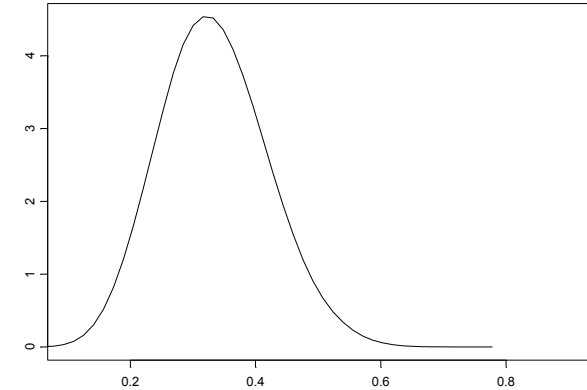
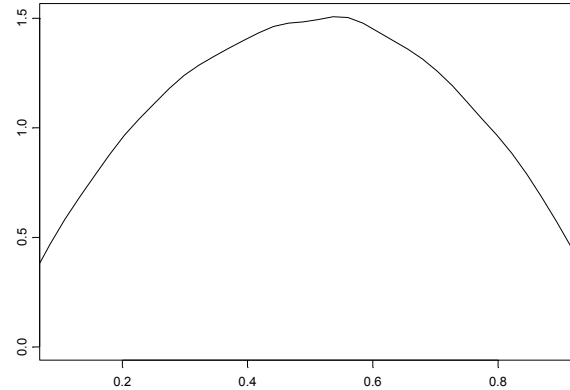
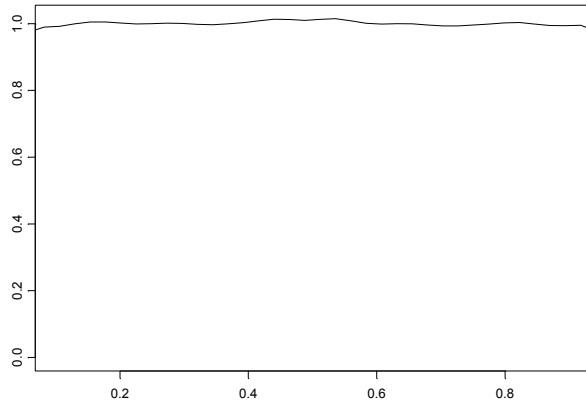
- $a < 1, b < 1$  → U-shaped
- $a > 1, b > 1$  → Unimodal
- $a < 1, b \geq 1$  → Positive skewed
- $a \geq 1, b < 1$  → Negative skewed
- $a = 1, b > 1$  → Strictly increasing
- $a > 1, b = 1$  → Strictly decreasing



# The Beta Distribution

- Match the plots to the distributions
- What are the posterior means, modes and variances?

Beta(1,1) Beta(2,2) Beta(100,100) Beta(2,1) Beta(10,20) Beta(9,1)

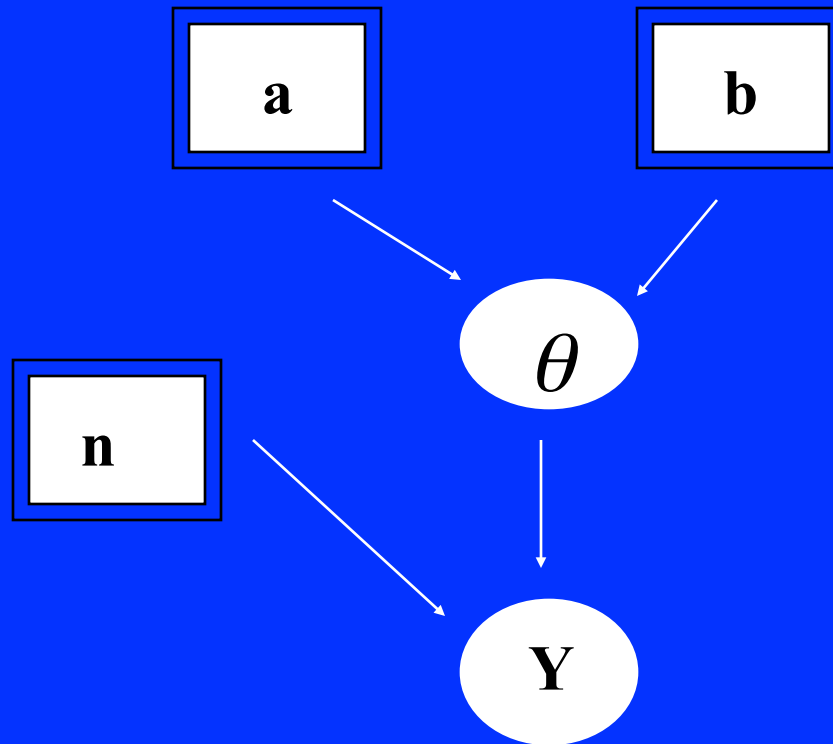


# DAG: Binomial model

- **Model**

$Y \sim \text{Binomial}(\theta, n)$

$\theta \sim \text{Beta}(a, b)$



# Priors for binomial model

- *Prior: Some alternatives*

1. Assume we ‘know nothing’ about  $\theta$ , so we set a uniform prior  $\theta \sim U[0,1]$ , equivalently,  $\theta \sim \text{Beta}(1,1)$
2. Based on past information, adopt a **Beta(9,1)** prior
3. Based on expert info, assume a **Beta(100,100)** prior

# Posterior for binomial model

$$\begin{aligned} P(\theta | y) &\propto \textit{likelihood} \times \textit{prior} \\ &= \theta^y (1-\theta)^{n-y} \times \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{y+a-1} (1-\theta)^{n-y+b-1} \\ &= \textit{Beta}(y+a, n-y+b) \end{aligned}$$

$$E(\theta) = a / (a + b)$$

$$E(\theta | y) = a + y / (a + b + n)$$

# Influence of prior on posterior

$$E(\theta | y) = a + y / (a + b + n)$$

- The posterior mean is a compromise between prior mean and sample mean
- The stronger the prior ( $a+b$ ), the more weight the prior has in the posterior
- The larger the sample size, the more weight the likelihood has in the posterior



# Your turn!

Binomial example with 22 presences, 7 absences:

Consider the following priors for  $\theta$ :

Beta(1,1)

Beta(9,1)

Beta(100,100)

*For each of these:*

- 1. What is the prior mean for  $\theta$ ?*
- 2. What is the posterior distribution for  $\theta$ ?*
- 3. What is the posterior mean for  $\theta$ ?*
- 4. What general conclusions can you make about the influence of priors and sample size?*

# Answers:

Sample proportion =  $22/29 = 0.76$

## ***Beta(1,1):***

Prior mean =  $(1)/(1+1) = 0.5$

Posterior mean =  $(22+1)/(29+2) = 0.74$

## ***Beta(9,1):***

Prior mean =  $(9)/(9+1) = 0.90$

Posterior mean =  $(22+9)/(29+10) = 0.79$

## ***Beta(100,100):***

Prior mean =  $(100)/(100+100) = 0.5$

Posterior mean =  $(22+100)/(29+200) = 0.53$

# Biased coin

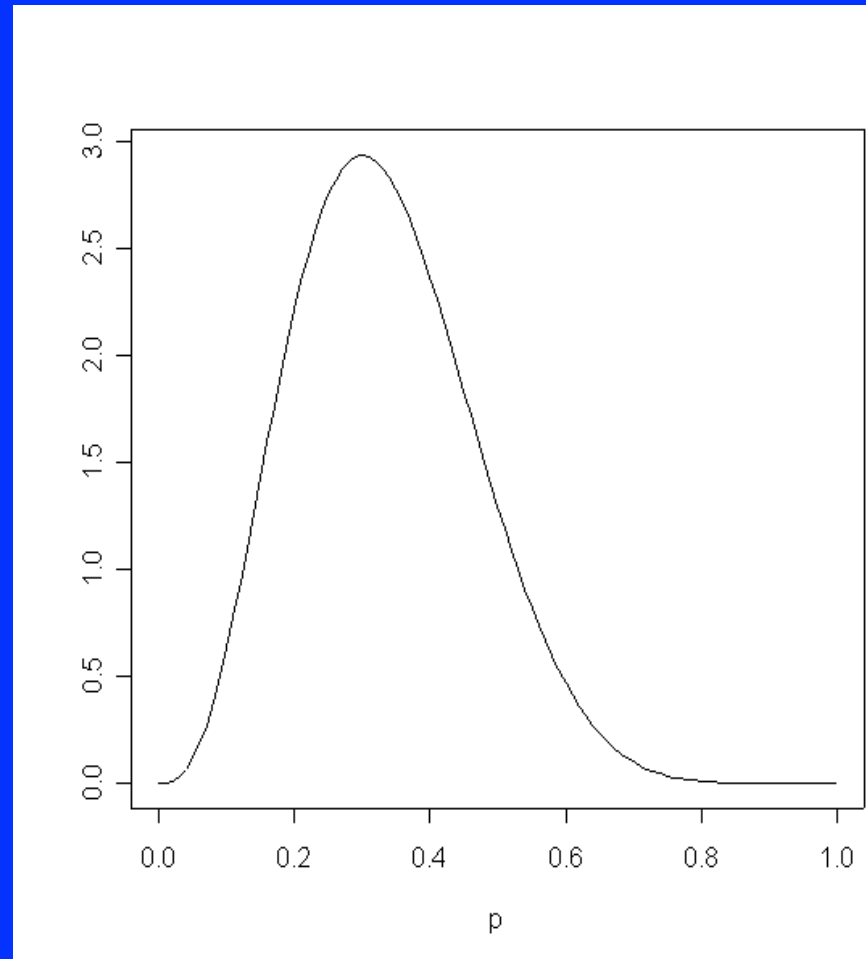
- $P(\text{Heads}) = P(X = 1) = p = ?$
- $X_1, \dots, X_n$  0-1 i.i.d. Bernoulli( $p$ ) trials
- Let  $X = \sum X_i$  be the number of heads in  $n$  trials
- Likelihood is  $f(X | p) = p^X (1 - p)^{n - X}$
- For prior use *uninformative* distribution on  $(0, 1)$
- So posterior distribution is proportional to

$$f(X|p) f(p) \propto p^X (1 - p)^{n - X}$$

- $f(p|X) \propto p^X (1 - p)^{n - X}$

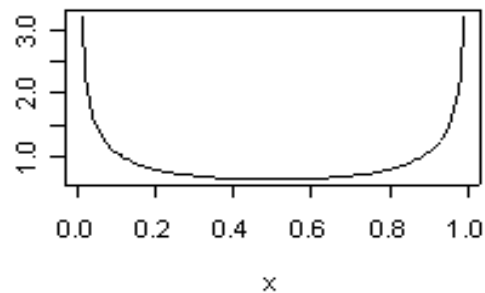
- Toss the coin 10 times
- Data: T, T, H, T, T, T, T, H, T, H  
0, 0, 1, 0, 0, 0, 0, 1, 0, 1
- $n=10$  and  $x=3$
- Posterior distribution is  
 $\text{Beta}(3+1, 7+1) = \text{Beta}(4, 8)$

- Posterior distribution Beta(4,8)
- Mean: 0.33
- Mode: 0.30
- $qbeta(.025,4,8) = 0.11$
- $qbeta(.975,4,8) = 0.61$
- $[.11, .61]$  is the 95% *credible* interval for  $p$
- $P(.11 < p < .61 | X) = .95$

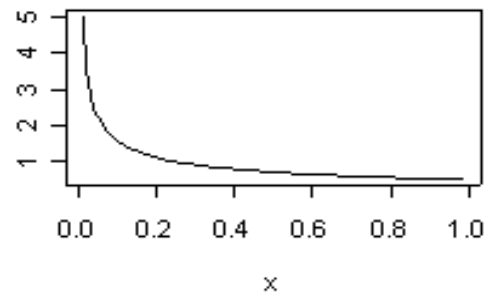


# Choice of Prior distributions

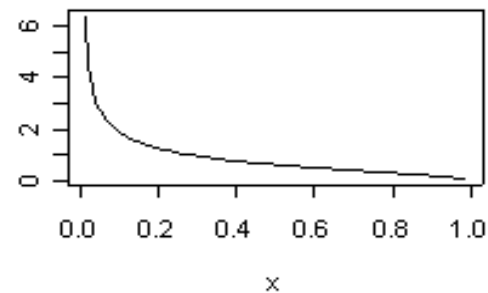
**a=.5, b=.5**



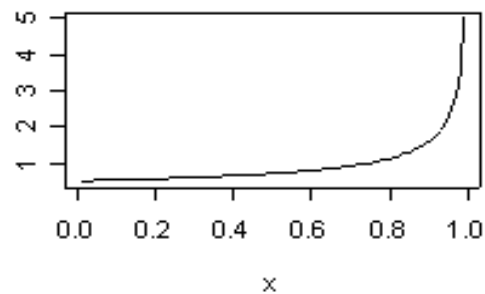
**a=.5, b=1**



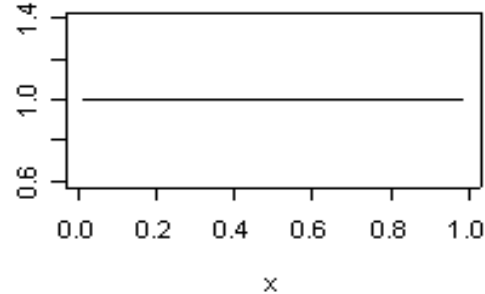
**a=.5, b=1.5**



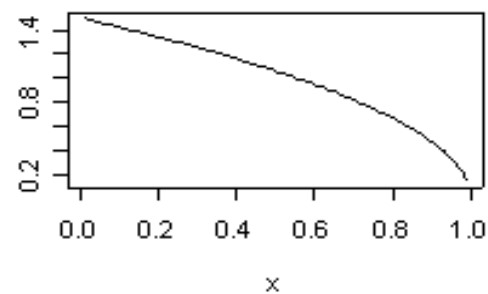
**a=1, b=.5**



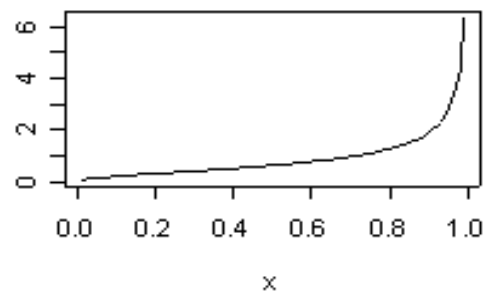
**a=1, b=1**



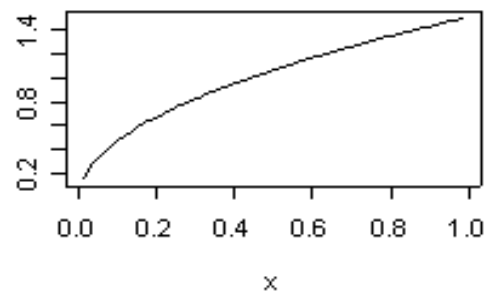
**a=1, b=1.5**



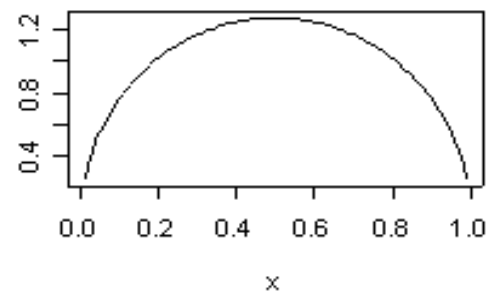
**a=1.5, b=.5**



**a=1.5, b=1**



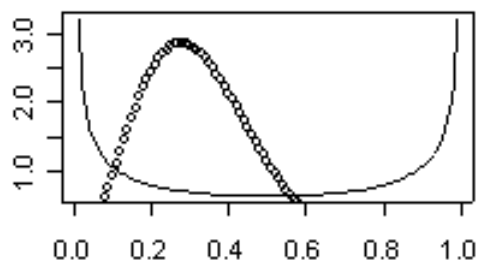
**a=1.5, b=1.5**



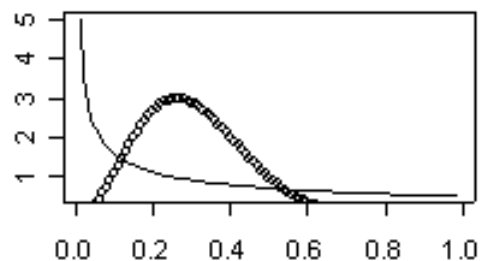
# Posterior distribution is $\text{Beta}(x+a, n-x+b)$

$\propto$

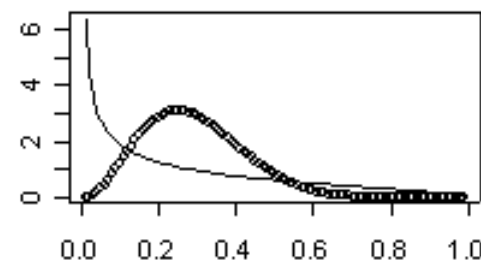
**a=0.5 b=0.5**  
postmean=0.32 postmax=0.28



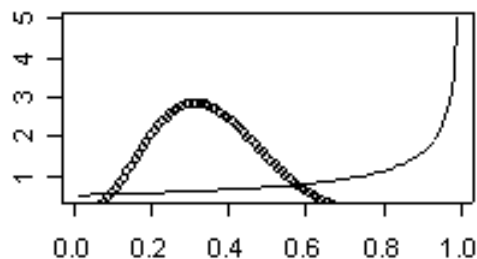
**a=0.5 b=1**  
postmean=0.3 postmax=0.26



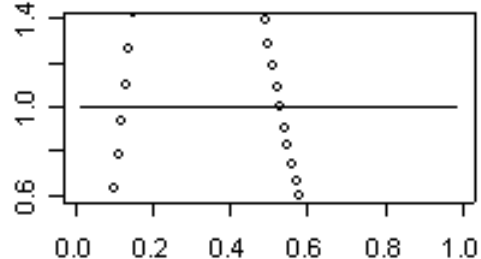
**a=0.5 b=1.5**  
postmean=0.29 postmax=0.25



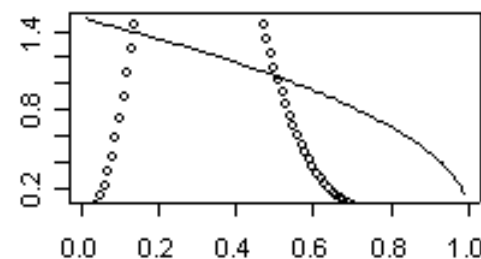
**a=1 b=0.5**  
postmean=0.35 postmax=0.32



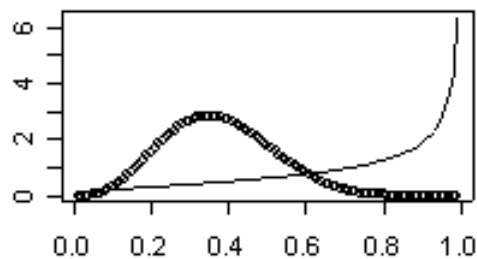
**a=1 b=1**  
postmean=0.33 postmax=0.3



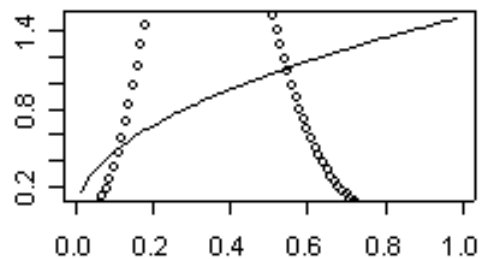
**a=1 b=1.5**  
postmean=0.32 postmax=0.29



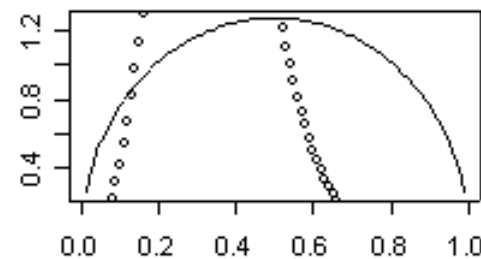
**a=1.5 b=0.5**  
postmean=0.38 postmax=0.35



**a=1.5 b=1**  
postmean=0.36 postmax=0.33



**a=1.5 b=1.5**  
postmean=0.35 postmax=0.32



# Conjugate priors

- It might be reasonable to expect the posterior distribution to be of the same form as the prior distribution
- This is the principle of *conjugacy*
- For a **Binomial likelihood** a **Beta prior** distribution is conjugate: a **Beta posterior** is obtained as a result of Bayes theorem



# Conjugate priors

<u>Family</u>	<u>Conjugate Prior</u>
Binomial( $N, \theta$ )	$\theta \sim \text{beta}(\alpha, \lambda)$
Poisson( $\theta$ )	$\theta \sim \text{gamma}(\delta_0, \gamma_0)$
$N(\mu, \sigma^2)$ , $\sigma^2$ known	$\mu \sim N(\mu_0, \sigma_0^2)$
$N(\mu, \sigma^2)$ , $\mu$ known, $\tau = 1/\sigma^2$	$\tau \sim \text{gamma}(\delta_0, \gamma_0)$
$\text{gamma}(\alpha, \lambda)$ , $\alpha$ known	$\lambda \sim \text{gamma}(\delta_0, \gamma_0)$
Beta( $\alpha, \lambda$ ), $\lambda$ known	$\alpha \sim \text{gamma}(\delta_0, \gamma_0)$

# Strategies for prior determination

1. Partition  $\Theta$  in sets (e.g., intervals)
  2. Determine the probability of each set
  3. Approach  $\pi$  by an histogram
- Select significant elements of  $\Theta$  and evaluate their respective likelihood
  - Empirical Bayes
  - Hierarchical Bayes

# Improper prior

- Not a distribution
- Often only way to derive a prior in noninformative settings
- Performances of associated estimators usually good
- Often occur as limits of proper distributions
- Check if the posterior is still proper

# Noninformative priors

Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing.

—Kass and Wasserman, 1996—

# Jeffreys' prior (1891–1989)

Proportional to  $\sqrt{|\text{Fisher information}|}$

Parameterization invariant

Suffers from dimensionality curse

# Dynamic Updating

- If we obtain more data, we do not have to redo all of the analysis: our posterior from the first analysis simply becomes our prior for this next analysis

- Binomial example:

**Stage 0:** Prior  $p(\theta) \sim \text{Beta}(1,1)$ ; ie  $E(\theta)=0.5$

**Stage 1:** Observe  $y=22$  ‘presences’ from 29 sites

Likelihood:  $p(y|\theta) \sim \text{Bin}(n=29, \theta)$

Posterior:  $p(\theta|y) \sim \text{Beta}(23,8)$ ; ie  $E(\theta|y) = 0.74$

**Stage 2:** Observe 5 more ‘presences’ from 10 sites

Likelihood:  $p(y|\theta) \sim \text{Bin}(n=10, \theta)$

Prior  $p(\theta) \sim \text{Beta}(23,8)$

Posterior  $p(\theta|y) \sim \text{Beta}(28,13)$ ; ie  $E(\theta|y) = 0.68$

# Example: Estimating a normal mean

$n$  observations  $Y = (y_1, \dots, y_n)$  from a normal distribution,  
unknown mean  $\mu$ , known variance  $\sigma^2$

- Likelihood:  $p(y_i|\theta) \quad y_i \sim N(\mu, \sigma^2)$   
 $p(Y|\theta) = (2\pi\sigma^2)^{-n/2} \exp(-0.5\sum_i (y_i - \theta)^2 / \sigma^2)$
- Prior: A conjugate prior is  $\theta \sim N(\mu_0, \sigma_0^2)$   
 $p(\theta) = (2\pi\sigma_0^2)^{-1/2} \exp(-(\theta - \mu_0)^2 / \sigma_0^2)$

$\mu_0, \sigma_0^2$  can be specified values representing  
our “best guess” at the true mean and how  
certain we are of this.

Or we can put priors on these values as well: hierarchical model

# Normal Model, known variance

- Posterior:

$$p(\mu|Y) \sim N(\mu_1, \sigma_1^2)$$

Posterior mean is a weighted average of prior and data

$$\mu_1 = (\mu_0 / \sigma_0^2 + n \bar{y} / \sigma^2) / (1/\sigma_0^2 + 1/\sigma^2)$$

$$1/\sigma_1^2 = 1 / \sigma_0^2 + n / \sigma^2$$

Posterior variance also combines variances from prior and data



# Your turn!

1. Suppose that we observe  $y = 2$  and wish to estimate the population mean  $\mu$
2. Assume **model**:  $p(y|\mu) \sim N(\mu, \sigma^2=3)$   
Assume **prior**:  $p(\mu) \sim N(\mu_0=0, \sigma_0^2=1)$
3. What is the **posterior** distribution for  $\mu$ ?
4. What if the prior is  $N(2,1)$ ?  $N(0,10)$ ?
5. Verify the equations on the previous slide

# Answers

- Observe  $y = 2$  ;  $p(y|\mu) \sim N(\mu, \sigma^2=3)$
- If prior  $p(\mu) \sim N(\mu_0=0, \sigma_0^2=1)$   
then the posterior is  $p(\mu|y) \sim N(\mu_1, \sigma_1^2)$   
posterior mean:  $\mu_1 = (0/1 + 2/3) / (1/1 + 1/3) = 0.50$   
posterior variance:  $1/\sigma_1^2 = 1/1 + 1/3 = 1.333$  so  $\sigma_1^2 = 0.75$
- If prior  $p(\mu) \sim N(\mu_0=2, \sigma_0^2=1)$   
 $\mu_1 = (2/1 + 2/3) / (1/1 + 1/3) = 2$   
 $1/\sigma_1^2 = 1/1 + 1/3 = 1.333$  so  $\sigma_1^2 = 0.75$
- If prior  $p(\mu) \sim N(\mu_0=0, \sigma_0^2=10)$   
 $\mu_1 = (0/10 + 2/3) / (1/10 + 1/3) = 1.54$   
 $1/\sigma_1^2 = 1/10 + 1/3 = 0.4433$  so  $\sigma_1^2 = 2.31$

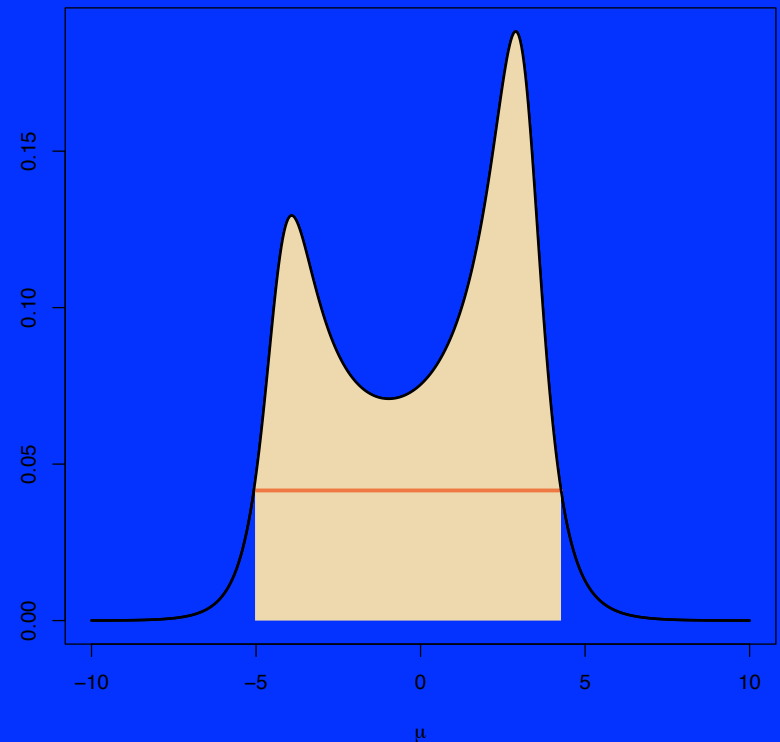
# Normal model, unknown mean unknown variance

$$\sigma \sim \text{Uniform}(a, b)$$



# Credible Intervals and Regions

- The Bayesian equivalent of frequentist confidence intervals
- Highest posterior density (HPD) regions: give the highest probabilities of containing  $\theta$  for a given volume



# Hypothesis testing

- Decide about validity of assumptions or restrictions on the parameter

$$H_0 : \theta \text{ in } \Theta_0$$

$$H_1 : \theta \text{ in } \Theta_1$$

## Binary decision process:

- accept coded by 1
- reject coded by 0

## 0-1 loss function

Accept the null if

the posterior probability of  $\Theta_0$  is  $> 0.5$

# Example: linear regression

**Model:**  $y = X\beta + e$ ;  $e \sim N(0, \sigma^2)$

Alternative representation:

$$y \sim N(\mu, \sigma^2); \quad \mu = X\beta$$

**OLS estimates:**

$$\beta_{\text{ols}} = (X^T X)^{-1} X^T y$$

$$\sigma^2_{\text{ols}} = y^T y - 2\beta_{\text{ols}}^T X^T y + \beta_{\text{ols}}^T X^T X \beta_{\text{ols}} / (n - p)$$

# Example: linear regression

**Likelihood:**

$$p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp[-\sum_i (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 / 2\sigma^2]$$

i.e. prop. to  $\exp [ (\mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} ) / 2\sigma^2 ]$

**MVN prior for  $\boldsymbol{\beta}$ :**  $p(\boldsymbol{\beta}) \sim \text{MVN}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$

Then  $p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}, \sigma^2)$  is prop. to  $p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta})$

i.e.  $\exp[\boldsymbol{\beta}^T (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^T \mathbf{y} / \sigma^2) - 0.5 \boldsymbol{\beta}^T (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X} / \sigma^2) \boldsymbol{\beta}]$

i.e. **posterior is MVN**

$$\text{mean} = (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^T \mathbf{y} / \sigma^2) / (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X} / \sigma^2)$$

$$\text{var} = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X} / \sigma^2)^{-1}$$

# Example: linear regression

IG prior for  $\sigma^2$  or equivalently, gamma prior on precision:  $1/\sigma^2$

$$p(\sigma^2) \sim \text{IG}(v/2, \delta/2)$$

Then  $p(\sigma^2 | y, X, \beta)$  is prop. to  $p(y | X, \beta, \sigma^2) p(\sigma^2)$

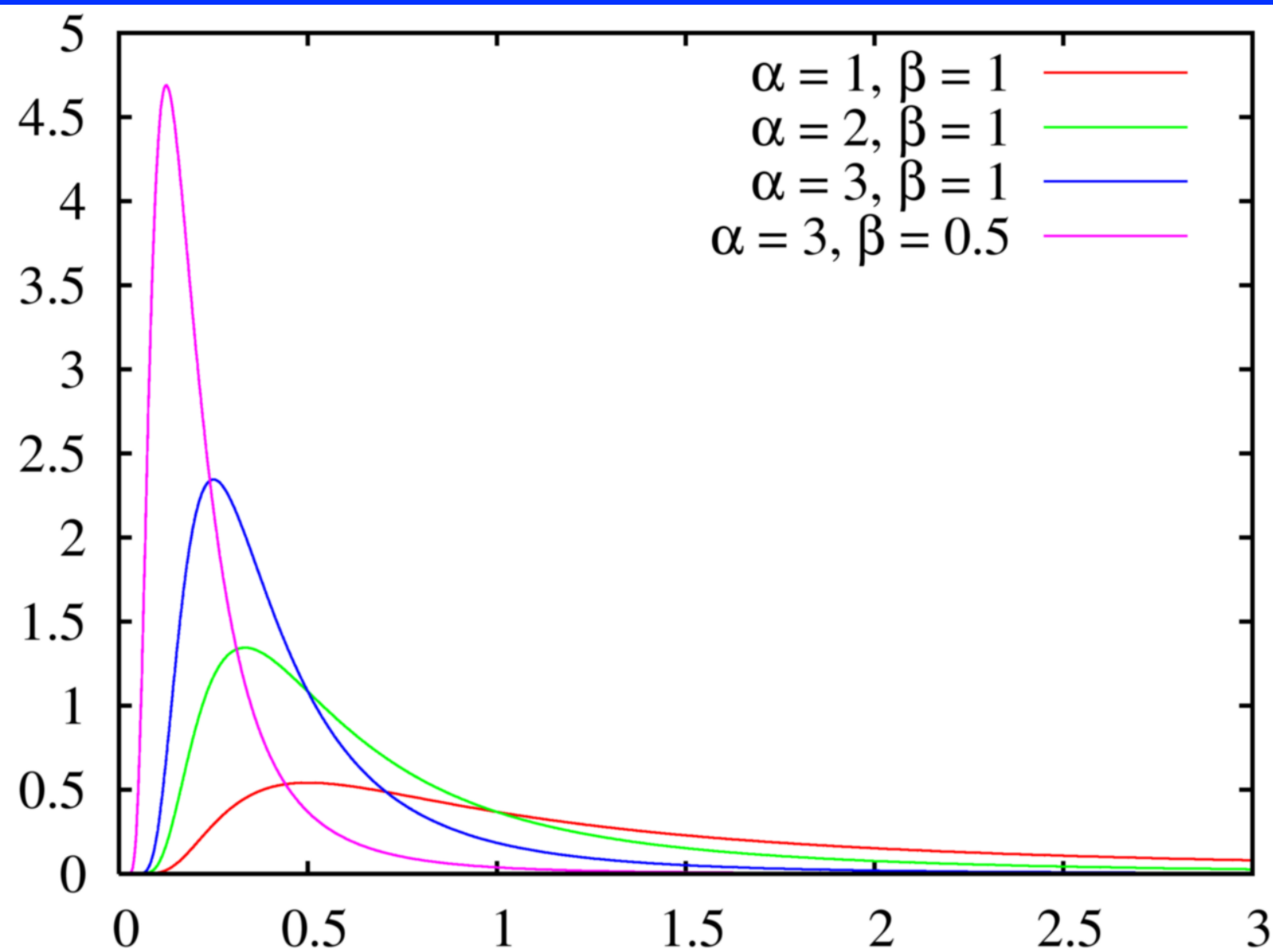
i.e. IG posterior with parameters

$$(v + n) / 2$$

$$(\delta + y^T y - 2\beta^T X^T y + \beta^T X^T X \beta) / 2$$



IG mean =  $\alpha / (\alpha - 1)$     variance =  $\beta^2 / (\alpha - 1)^2 (\alpha - 2)$



# Example: logistic model

Experiment: proportion of seeds that germinated on each of 21 plates arranged as a 2 by 2 factorial layout by **seed** and type of **root extract**

$r_i$  and  $n_i$  are the number of germinated and the total number of seeds on the  $i$ th plate,  $i = 1, \dots, N$

Here  $y_i = r_i$ ;  $\theta_i = p_i =$  probability of germination on the  $i$ th plate

seed <i>O. aegyptiaco</i> 75			seed <i>O. aegyptiaco</i> 73								
Bean			Cucumber			Bean			Cucumber		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	55	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.44	46	79	0.58	0	4	0.00	3	7	0.43
			10	13	0.77						

# Logistic model

$$r_i \sim \text{Binomial}(p_i, n_i)$$

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i$$

$$b_i \sim \text{Normal}(0, \sigma^2) \quad \text{Extra-binomial variation}$$

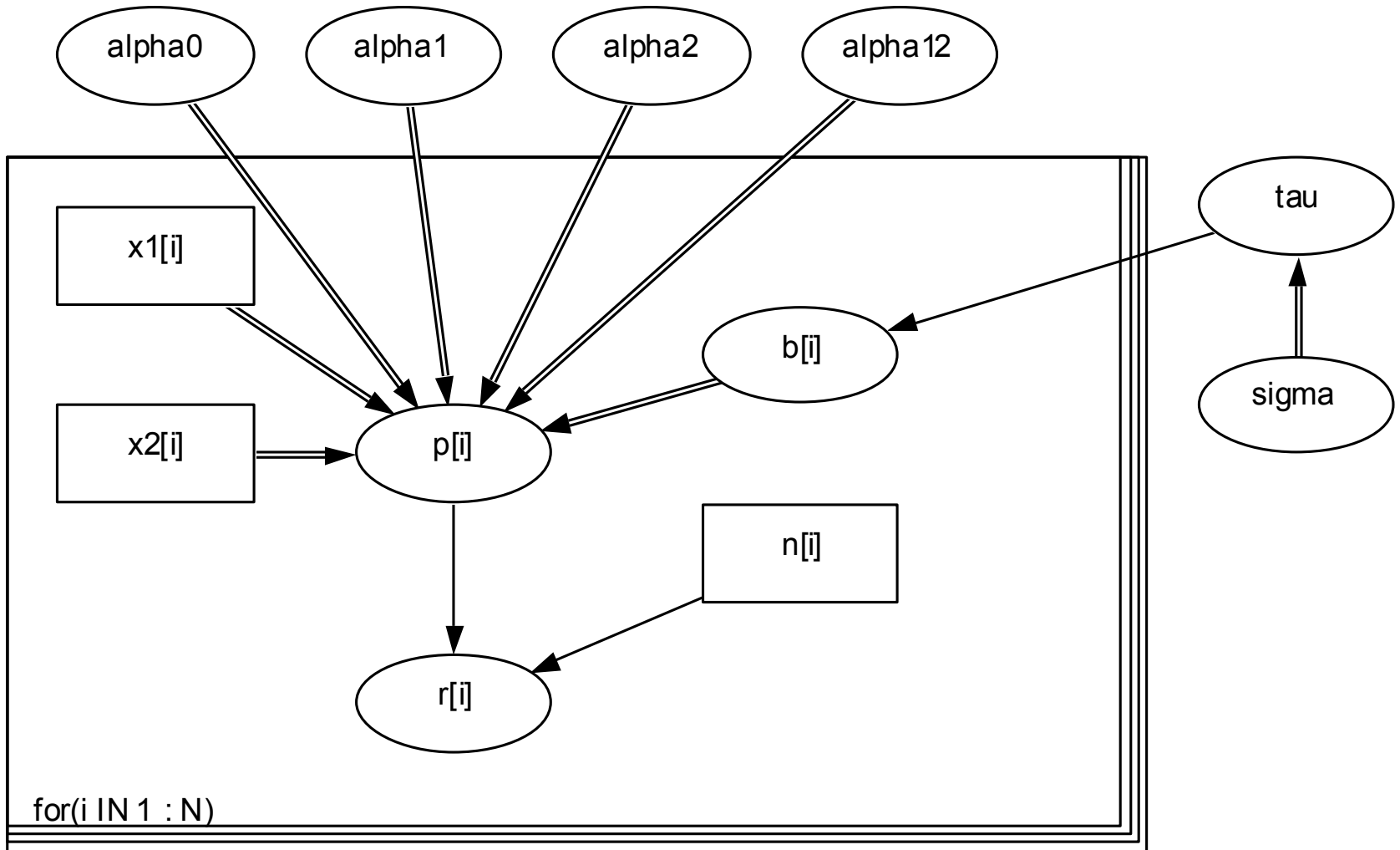
$$\alpha_0, \alpha_1, \alpha_2, \alpha_{12} \sim \text{Normal}(0, 1.0\text{E}6)$$

$$\sigma \sim \text{U}(0, 100)$$

1.0 E6 = 1000000

1.0 E-6=0.000001

# DAG for logistic model



# Example: random effects linear model

30 rats, weighed weekly for 5 weeks

Model as random effects linear growth curve

**Weight  $Y_{ij}$  of rat  $i$  on day  $x_j$**

**$x_j =$     8        15        22        29        36**

<b>Rat 1</b>	151	199	246	283	320
<b>Rat 2</b>	145	199	249	293	354
<b>...</b>					
<b>Rat 30</b>	153	200	244	286	324

# Rats model

- **Model**

$$y_{ij} \sim \text{Normal} (\alpha_i + \beta_i x_j, \sigma_C^2)$$

- **Priors**

$$\alpha_i \sim \text{Normal} (\alpha_C, \sigma_\alpha^2)$$

$$\beta_i \sim \text{Normal} (\beta_C, \sigma_\beta^2)$$

$$\alpha_C \sim \text{Normal} (0, 1E4)$$

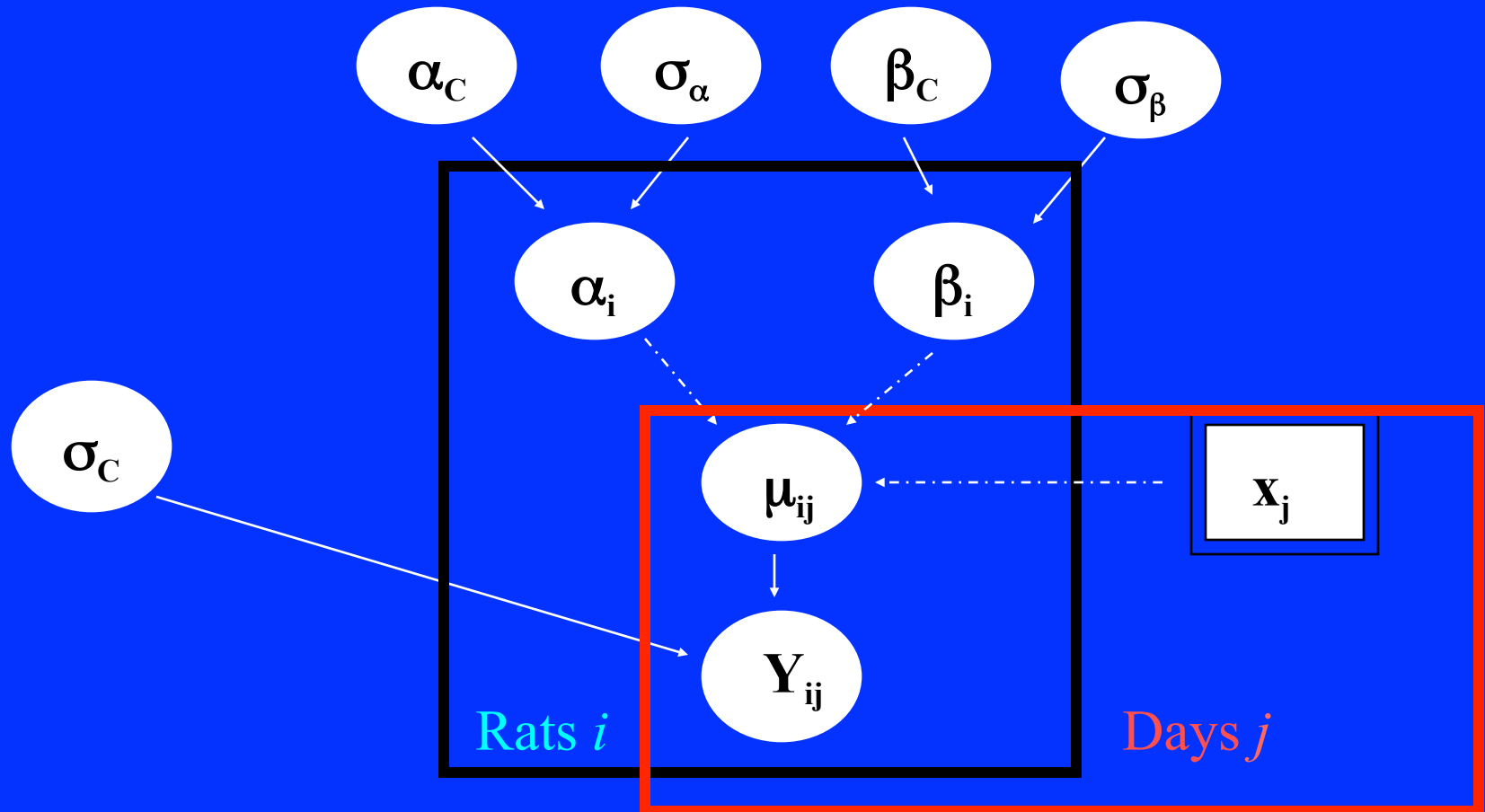
$$\beta_C \sim \text{Normal} (0, 1E4)$$

$$\sigma_C \sim \text{Uniform} (0, 100)$$

$$\sigma_\alpha \sim \text{Uniform} (0, 100)$$

$$\sigma_\beta \sim \text{Uniform} (0, 100)$$

# DAG for rats example



End of session 1

First part of session 2: MC and MCMC

Second part of session 2: back to these slides



# Recall linear regression

$$\beta \mid y, X, \sigma^2 \sim \text{MVN}\left( \frac{(\Sigma_0^{-1} \beta_0 + X^T y / \sigma^2)}{(\Sigma_0^{-1} + X^T X / \sigma^2)}, \right. \\ \left. (\Sigma_0^{-1} + X^T X / \sigma^2)^{-1} \right)$$

$$\sigma^2 \mid y, X, \beta \sim \text{IG}\left( (v + n) / 2, \right. \\ \left. (\delta + y^T y - 2\beta^T X^T y + \beta^T X^T X \beta) / 2 \right)$$

# Computation: linear regression

Constructing a Gibbs sampler to sample from the joint posterior distribution of  $\beta$  and  $\sigma^2$  is straightforward

## 1. Update $\beta$

1.1 Compute  $V = \text{Var}[\beta \mid y, X, \sigma^{2(s)}] = (\Sigma_0^{-1} + X^T X / \sigma^{2(s)})^{-1}$

1.2 Compute  $m = \text{E}[\beta \mid y, X, \sigma^{2(s)}] = \frac{\Sigma_0^{-1} \beta_0 + X^T y / \sigma^{2(s)}}{\Sigma_0^{-1} + X^T X / \sigma^{2(s)}}$

1.3 Sample  $\beta^{(s+1)} \sim \text{MVN}(m, V)$

## 2. Update $\sigma^2$

2.1 Compute  $b = y^T y - 2\beta^{(s+1)T} X^T y + \beta^{(s+1)T} X^T X \beta^{(s+1)}$

2.2 Sample  $\sigma^{2(s+1)} \sim \text{IG}([\nu + n]/2, [\delta + b]/2)$

# WinBUGS

- Windows version of ‘Bayesian Analysis Using Gibbs Sampling’
- Also available: OpenBUGS
- Can call WinBUGS from R, Matlab, etc
- Can program MCMC using R, Fortran, C etc
- See also First Bayes and other specialist programs

# Running WinBUGS

## 1. Model: Specification

1. Check model
2. Load data
3. Compile
4. Load or generate initial values for simulations

## 2. Inference

1. **Model: Update:** run chain for short time (burn-in)
2. **Inference: Samples:** Monitor parameters of interest
3. **Model: Update:** run chain for longer time (collection)
4. **Inference: Samples:** Summary statistics, plots etc

# Example: BUGS code for seeds

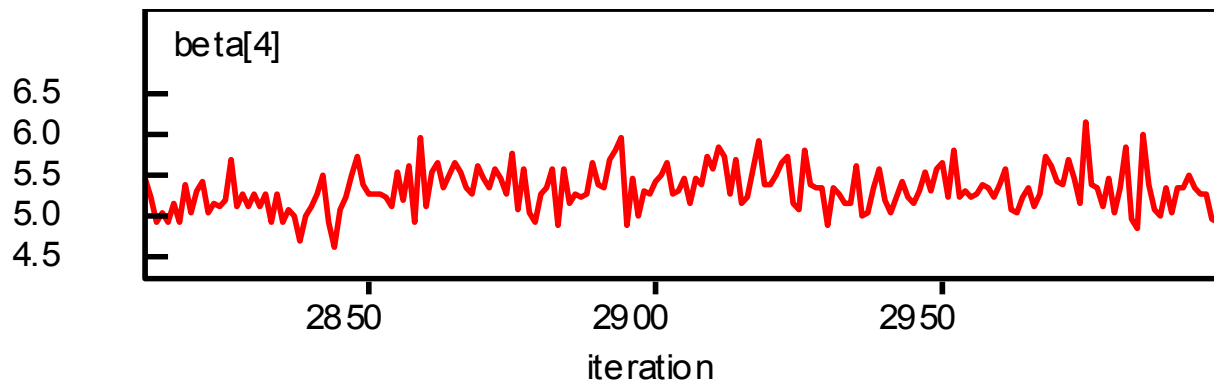
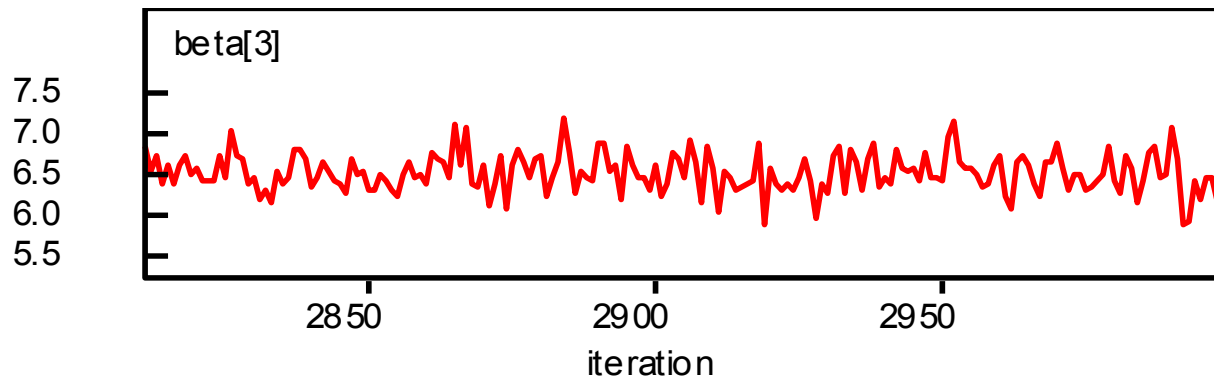
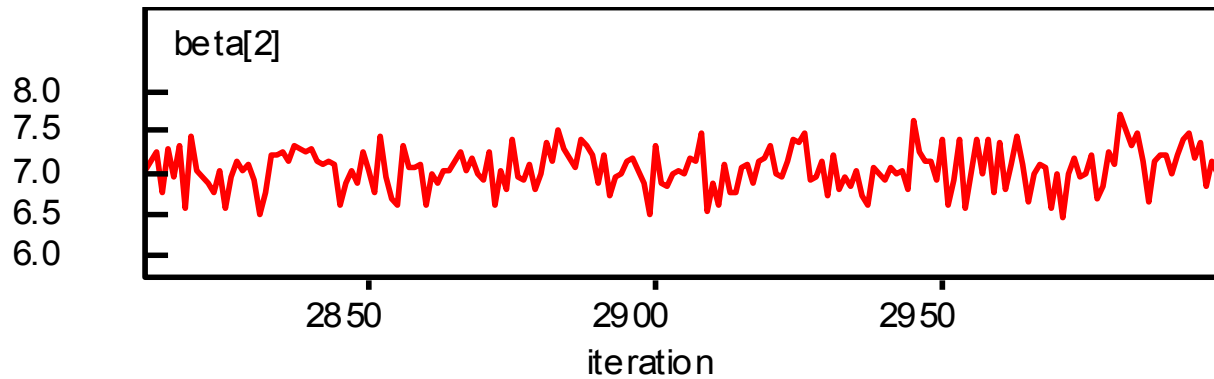
```
model {  
  
  for (i in 1:N) {  
    r[i] ~ dbin(p[i],n[i])  
    b[i] ~ dnorm(0.0,tau)  
    logit(p[i]) =  
      alpha0+alpha1*x1[i]+alpha2*x2[i]+alpha12*x1[i]*x2[i]+b[i]  
  }  
  
  alpha0 ~ dnorm(0 , 1.0E-6)  
  alpha1 ~ dnorm(0 , 1.0E-6)  
  alpha2 ~ dnorm(0 , 1.0E-6)  
  alpha12 ~ dnorm(0 , 1.0E-6)  
  sigma ~ dunif(0 , 100)  
  tau <- 1/(sigma*sigma)  
}
```

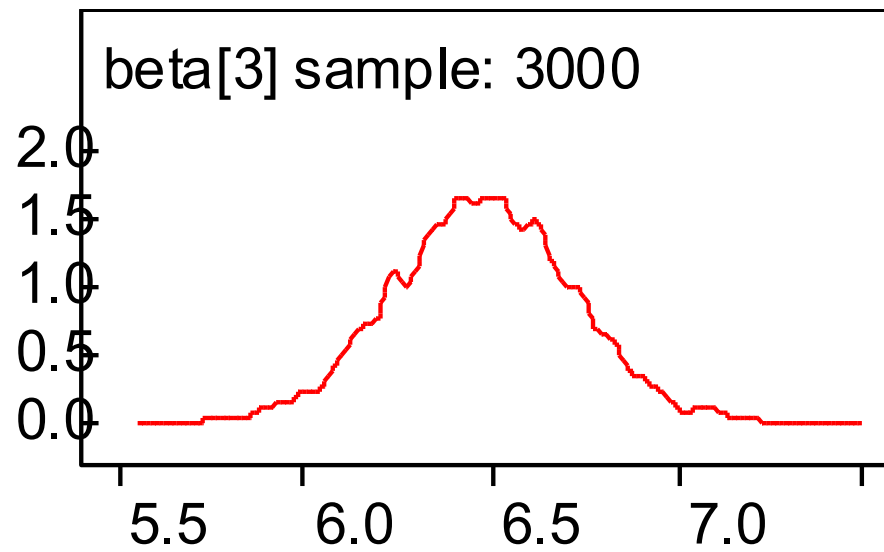
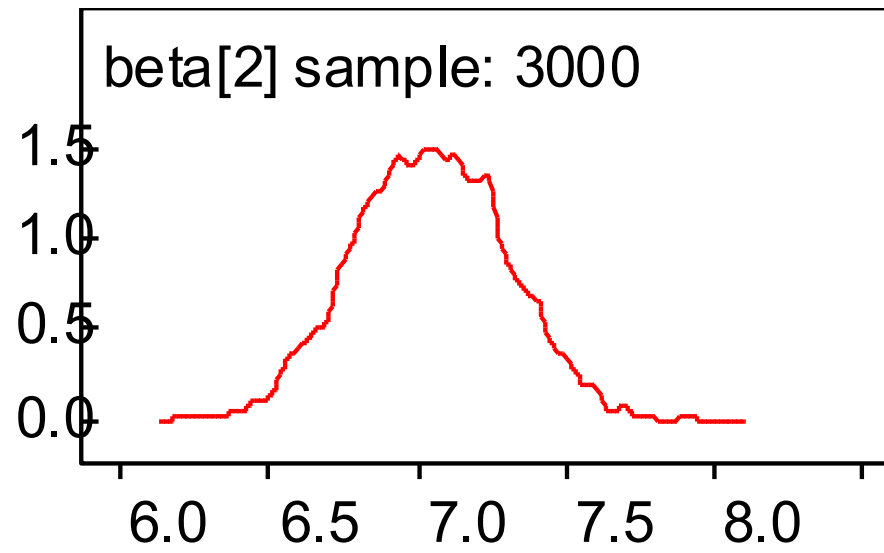
*Beware! WinBugs uses precisions  $1/\sigma^2$  for normal distributions!*

# Example: BUGS code for rats

```
model {
    N = no. rats, T = no. time periods
    for (i in 1:N) {
        for (j in 1:T) {
            mu[i,j] <- alpha[i] + beta[i] * (x[j] - x.bar)
            Y[i,j] ~ dnorm(mu[i,j], tau.c)
        }
        alpha[i] ~ dnorm(alpha.c, tau.alpha)
        beta[i] ~ dnorm(beta.c, tau.beta)
    }
    alpha.c ~ dnorm (0, 1.0E-4)
    beta.c ~ dnorm (0, 1.0E-4)
    tau.c <- 1/(sigma.c*sigma.c)
    sigma.c ~ dunif (0, 100)
    tau.alpha <- 1/(sigma.alpha*sigma.alpha)
    sigma.alpha ~ dunif (0, 100)
    tau.beta <- 1/(sigma.beta*sigma.beta)
    sigma.beta ~ dunif (0, 100)
    x.bar <- mean( x[] )
    alpha0 <- alpha.c - beta.c * x.bar
}
```

# Trace plots for rats







# Bayesian analysis via R

Many packages are now available:

<http://cran.r-project.org/web/views/Bayesian.html>

Example: bayesm

**runireg**: Gibbs Sampler for Univariate Linear Model

**rhierLinearModel**: Gibbs Sampler for Hierarchical Linear Model

**hierLinearModel**: Gibbs Sampler for Hierarchical Linear Model

Example: MCMCPack

**MCMCLogit**: MCMC for logistic regression

# Your turn:

## Linear modelling with WinBUGS

Read the RATS example in the Help Examples Vol 1 and do the following:

- Make sure that you understand the model.
- Make sure that you understand the WinBUGS code for the model and priors.
- Run this model. (ie check the model, enter the data, validate the model, update, monitor, update, summarise, etc).
- Write down the posterior estimates of the overall growth parameters.

# Your turn: More with WinBugs

Continue with the RATS example and experiment with changing the code to reanalyse the data excluding the first 10 rats. To do this:

1. Open a new file in WinBugs (File New).
2. Copy the original data, model and initial values to the new file.
3. Delete the first 10 rats in the dataset.
4. Modify the Model and Initial Values files to allow for this change.
5. Run the analysis again. What difference do you see in the results?

# Your turn: using R (*bayesm*) for linear modelling

- Install and load *bayesm*
- Read the documentation for *runiregGibbs*
- Run the example given in the documentation (see next slide)
- Use this package to analyse the environmental health study data given in the previous example

# bayesm code for linear modelling

```
# set number of iterations
```

```
R = 10000
```

```
# simulate data
```

```
X=cbind(rep(1,n),runif(n))
```

```
beta=c(1,2)
```

```
sigsq=.25
```

```
y=X%*%beta+rnorm(n,sd=sqrt(sigsq))
```

```
# set data
```

```
Data1=list(y=y,X=X)
```

```
Mcmc1=list(R=R)
```

```
# run analysis
```

```
out=runiregGibbs(Data=Data1,Mcmc=Mcmc1)
```

```
# print output
```

```
cat("Summary of beta and Sigma draws",fill=TRUE)
```

```
summary(out$betadraw,tvalues=beta)
```

```
summary(out$sigmasqdraw,tvalues=sigsq)
```

```
plot(out$betadraw)
```

# Your turn: Do it yourself

For an environmental health study we want to relate

$Y$  = amount of ammonia escaping in an industrial plant

$X$  = temperature

$X = c(27, 27, 25, 24, 22, 23, 24, 24, 23, 18, 18, 17, 18,$   
 $19, 18, 18, 19, 19, 20, 20)$

$Y = c(42, 37, 37, 28, 18, 18, 19, 20, 15, 14, 14,$   
 $13, 11, 12, 8, 7, 8, 8, 9, 15)$

# Do it yourself

The aim is to fit a simple regression

$$y=a+bx$$

to the data on the previous slide

- Open a new file in WinBUGS
- Write a simple regression model in WinBUGS code
- Type the data in your file, in a form that WinBUGS will read
- Type some initial values in your file, in a form that WinBUGS will read. OR let WinBUGS generate the initial values
- Run the model in WinBUGS

# Possible code for regression

```
model{  
  for( i in 1 : N ) {  
    Y[i] ~ dnorm(mu[i],tau)  
    mu[i] <- alpha + beta * x[i]  
  }  
  beta ~ dnorm(0.0, 1.0E-6)  
  alpha ~ dnorm(0.0,1.0E-6)  
  sigma1 ~ dunif(0,100)  
  tau <- (1/sigma*sigma)  
}
```

Data

```
list( N=20,  
  x=c(27,27,25,24,22,23,24,24,23,18,18,17,18,19,18,18,19,19,20,20),  
  Y=c(42, 37, 37, 28, 18, 18, 19, 20, 15, 14, 14, 13, 11, 12, 8, 7, 8, 8, 9, 15)  
)
```



# Recap

1. What is the difference between a likelihood, a prior, a posterior and an initial value?
2. What is a conjugate prior? Give an example
3. Describe the general concept of MCMC
4. How do you obtain a 95% credible interval for a parameter using MCMC?

# What about convergence?

- As with all simulation methods, we need to make sure that:
  - the simulations have converged to the right distributions
  - the whole distribution is being explored in the simulation
  - we have run the simulations long enough to obtain adequate estimates

# How do we do this?

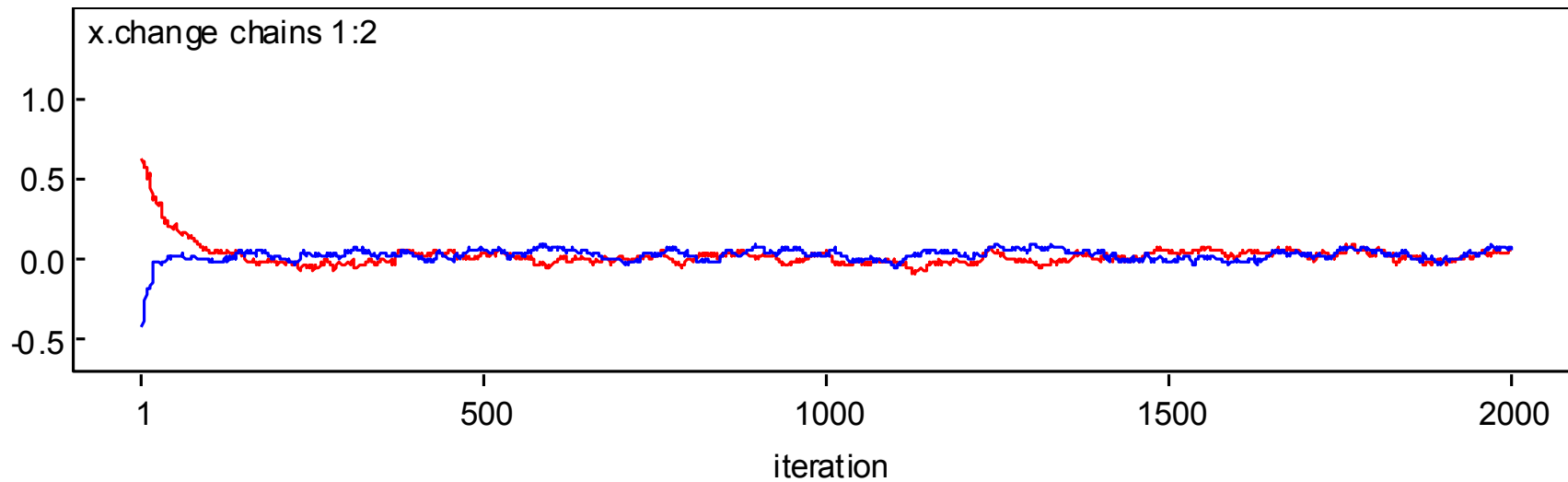
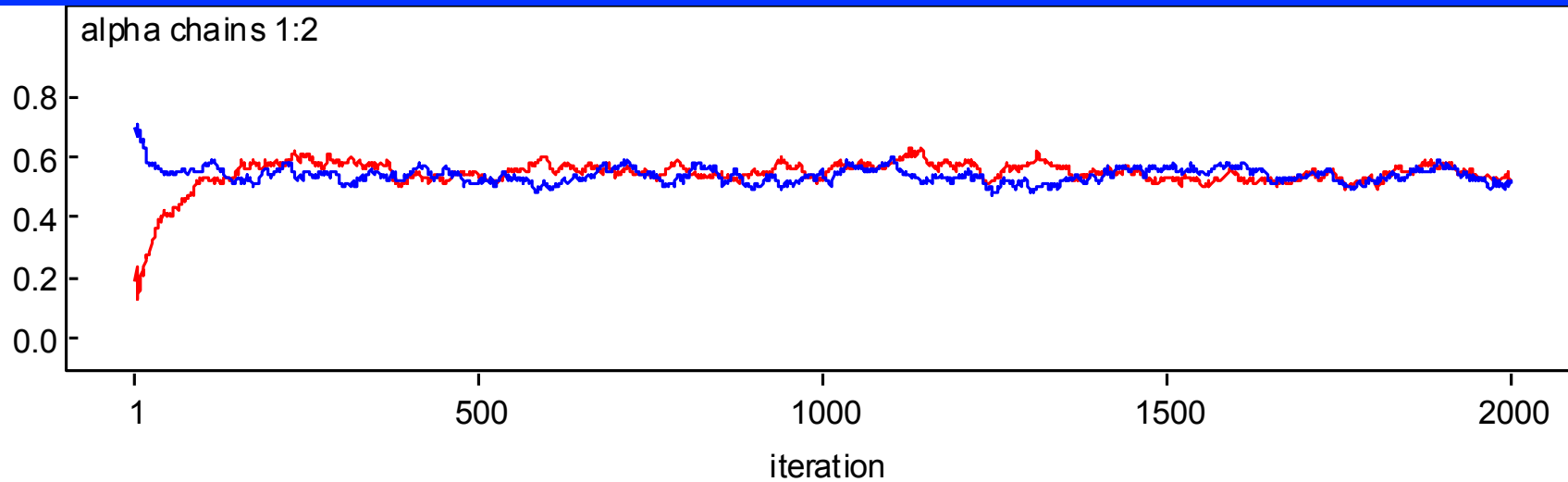
- Theoretical results
- Diagnostics:
  - Can you detect ‘burn-in’?
  - Do multiple chains show dependence on initial values?
  - Do the (time series) trace plots show that the chain is ‘stable’ around a mean value?
  - Are the posterior density plots smooth and well behaved?
  - Are the common diagnostic tests passed?

# Theoretical Results

$$\sup_{x \in C} |P^n(x, C) - P^\infty(C)| \leq M \rho^n_C$$

$$\int P(x, dy) V(y) \leq (1 - \beta) V(x) + I_C(x)$$

# Visual assessment of multiple chains



# Convergence: Geweke (1992)

- Look at a **single long run**
- Test for equal mean for  
“**early**” part (1st quarter) and  
“**late**” part (2nd half) of the chain
- Test statistic is  $Z \sim N(0,1)$  if the sample is all from the same distribution.
- (This is only a test of “non-convergence”)

# Convergence: Gelman & Rubin (1992)

- **Many long runs** from different starting points
- Convergence assessed via an **ANOVA** between and within the chains
- Monitor convergence by **R**: a conservative estimate of how much extra information about the variable that we could expect to gain by running the chains indefinitely
  - R tends to 1 as  $n$  tends to infinity
  - R is subject to sampling variation so monitor *R* and its upper 97.5% confidence limit
- Works best when posterior is approx. normal (may need to transform some variables, eg probs, variances)

# Convergence: Raftery & Lewis (1992)

- Look at a **single long run**
- Reduce to two-state Markov chain and use this theory
- Diagnostic estimates:
  - $n_0$  = length of burnin
  - $N$  = additional iterations needed to estimate a posterior quantile adequately
- Can give quite different estimates depending on starting values and required accuracy of estimation



# Convergence: Heidelberger & Welch (1983)

- Look at a **single long run**
- Hypothesis test based on Brownian bridge theory and spectral density estimation
- Iterative procedure:
  - test H0: **entire sample** of values for a given variable form a stationary process
  - if H0 rejected, **discard first 10%** and repeat test
  - continue discarding until H0 accepted or 50% samples are discarded (need to run chain for longer)
- Test has *very low power* to detect lack of convergence for small sample size.

# Convergence assessment in WinBUGS

- Trace plots
- Autocorrelation
- BGR diagnostic

# BGR diagnostic in WinBUGS

- Brooks-Gelman-Rubin convergence statistic
- Compare variation **between** & **within** multiple chains
- The width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio  $R$  (= pooled / within) is red
- Want convergence of  $R$  to 1
- Want convergence of both the pooled and within interval widths to stability

# Model evaluation

## posterior predictive checks

- Compare observed statistics with values predicted under the model
- Compare observed data with replicated data
  - if the model is adequate, replicated data generated under the model should look similar to the observed data

# Model Comparison

- Bayes factors, posterior odds, BIC, DIC
- Reversible jump MCMC  
Birth and death MCMC
- Model averaging

# Bayes factors

- Consider models  $M_1$  and  $M_2$  (not nec. nested)
- Choose a model based on its posterior probability given the data. This is proportional to the prior probability of the model multiplied by the likelihood of the model given the data.
- So we consider:

$$\frac{p(M_2|y)}{P(M_1|y)} = \left\{ \frac{p(M_2)}{p(M_1)} \right\} \times \left\{ \frac{p(y|M_2)}{P(y|M_1)} \right\}$$

# Bayes factors

$$p(M_2|y) / P(M_1|y) = \{p(M_2) / p(M_1)\} \times \{p(y|M_2) / P(y|M_1)\}$$

- The second term (the ratio of marginal likelihoods) is termed the Bayes factor  $B_{21}$  and is similar to a likelihood ratio, but  $p(y|M)$  is integrated over the parameters instead of maximised:
- 
- eg,  $p(y|M_1) = \int p(y|M_1, \theta_1) p(\theta_1) d\theta_1$
- $2\log(B_{21})$  gives same scale as usual deviance and LR statistics.

# Guidelines for Bayes Factors (arbitrary!)

$B_{21}$	$2\log(B_{21})$	Interpretation
$<1$	Negative	Support for $M_1$
1 to 3	0 to 2	Weak support for $M_2$
3-20	2-6	Support for $M_2$
20-150	6-10	Strong evidence for $M_2$
$>150$	$>10$	Very strong support for $M_2$



# Bayesian Information Criterion

## BIC

- Approximate the Bayes factor
- Assume the prior for  $\theta$  given a model is multivariate normal and that the prior is equivalent to a single extra observation
- $p$  is the number of parameters
- $n$  is the number of observations

$$\text{BIC} = \log P(y|\theta^*, M) - (p/2) \log n$$

Can rewrite as  $\text{BIC} = n \log(1-R^2) + k \log(n)$

# Discussion of BIC

- BIC penalises models which improve fit at the expense of more parameters (encourages parsimony)
- A problem is that the true dimensionality (number of parameters  $p$ ) of the model is often not known, and also that the number of parameters may increase with sample size  $n$ .
- Can approximate using the effective number of parameters (Speigelhalter et al, 1999)
- Alternatives are DIC (deviance information criterion, calculated in WinBUGS), conditional posterior predictive probabilities, etc.

# Model Averaging

- Instead of choosing a single model, a common practice is model averaging
- This is the practice of combining expected values obtained from different models (perhaps describing different combinations of variables) weighted by their corresponding posterior probabilities
- Adoption of this approach depends on the aim of the analysis and the trade-off between improved estimation and ease of interpretation

# Your turn: using R (*bayesm*) for linear modelling

- Install and load *bayesm*
- Read the documentation for *runiregGibbs*
- Run the example given in the documentation (see also next slide)
- Use this package to analyse the environmental health study data given in the previous example.

# bayesm code for linear modelling

```
# set number of iterations
```

```
  R = 10000
```

```
# simulate data
```

```
  X=cbind(rep(1,n),runif(n))
```

```
  beta=c(1,2)
```

```
  sigsq=.25 y=X%*%beta+rnorm(n,sd=sqrt(sigsq))
```

```
# set data
```

```
  Data1=list(y=y,X=X)
```

```
  Mcmc1=list(R=R)
```

```
# run analysis
```

```
  out=runiregGibbs(Data=Data1,Mcmc=Mcmc1)
```

```
# print output
```

```
  cat("Summary of beta and Sigma draws",fill=TRUE) summary(out  
$betadraw,tvalues=beta) summary(out$sigmasqdraw,tvalues=sigsq)
```

```
  plot(out$betadraw)
```

# code for env. health model

```
library(bayesm)
R = 10000
# read data from a .csv file, with columns Amm, Int, Temp; Int col = 1's
# wd<-"c://Work/Work13/courses/MISG2013"
# setwd(wd)
# seeds <- read.csv("seeds.csv")
# attach(seeds)
# alternative: directly enter Amm , Int and Temp in R:
Amm<- c(42,37,37,28,18,18,19,20,15,14,14,13,11,12,8,7,8,8,9,15)
Temp<- c(27,27,25,24,22,23,24,24,23,18,18,17,18,19,18,18,19,19,20,20 )
Int <- rep(1,20)
X=cbind(Int, Temp)
Data1=list(y=Amm,X=X)
Mcmc1=list(R=R)
out=runiregGibbs(Data=Data1,Mcmc=Mcmc1)
```

# code for env. health model (cont)

```
cat("Summary of beta and Sigma draws",fill=TRUE)
summary(out$betadraw)
summary(out$sigmasqdraw)
plot(out$betadraw)
b0 <- mean(out$betadraw[,1])
b1 <- mean(out$betadraw[,2])
plot(Temp, Amm)
lines(Temp, b0 + b1*Temp)
plot(Amm, Amm, type="l") # note, if you copy this into R, change the “
points(Amm, b0+b1*Temp)
```

- What could you do to improve the fit of this model? (Hint: consider a transformation.)

# Your turn: using R (MCMCPack) for linear regression

- Install and load *MCMCPack*
- Browse:  
*[http://mcmcpack.wustl.edu/files/  
MartinQuinnMCMCpackslides.pdf](http://mcmcpack.wustl.edu/files/MartinQuinnMCMCpackslides.pdf)*
- Read the documentation for *MCMCregress*
- Run the example given in the *MCMCregress* documentation (see also next slide)
- Use this package to analyse the environmental health example.



# code for example regression

```
library(MCMCPack)
```

```
line <- list(X = c(-2,-1,0,1,2), Y = c(1,3,3,3,5))
```

```
line
```

```
posterior <- MCMCregress(Y~X, data=line, verbose=FALSE)
```

```
posterior <- MCMCregress(Y~X, data=line, verbose=TRUE)
```

```
plot(posterior)
```

```
raftery.diag(posterior)
```

```
summary(posterior)
```

# code for env. health model

```
Amm<- c(42,37,37,28,18,18,19,20,15,14,14,13,11,12,8,7,8,8,9,15)
Temp<- c(27,27,25,24,22,23,24,24,23,18,18,17,18,19,18,18,19,19,20,20 )
Data1=list(y=Amm,X=Temp)
posterior <- MCMCregress(y~X, data=Data1, verbose=FALSE)
plot(posterior)
raftery.diag(posterior)
summary(posterior)
```

# Your turn: using R (MCMCpack) for linear logistic regression

- Read the documentation for *MCMClogit*
- Run the example given in the *MCMClogit* documentation (see also next slide)

How would you use this package to analyse the seeds example?

# MCMCpack example: logistic regression

```
library(MCMCpack)
data(birthwt)
?MCMClogit
?birthwt
summary(birthwt)
names(birthwt)
posterior <- MCMClogit(low~age+as.factor(race)+smoke,
  data=birthwt) plot(posterior)
summary(posterior)
```

# More modelling in R using bayesm

- Bivariate normal Gibbs sampler:

In your own time, read the documentation for [rbiNormGibbs](#) and run the example.

- Hierarchical linear model:

In your own time, read the documentation for [rhierLinearModel](#) and run the example.

Would this be applicable to the Rats example?  
If so, how?

# Your turn: model assessment and comparison in WinBugs

- Using WinBUGS, run the Seeds example with and without the interaction term.
  - What is the difference in goodness of fit of the models as measured by the DIC? (A smaller DIC indicates a better fit.)

Run some iterations. Choose the DIC option from the 'Inference' menu and set DIC. Run some more iterations. Return to the DIC box. Use the 'Total DIC'.

# Your turn: model assessment and comparison in R

- Read the documentation for the package *BayesFactor* in MCMCPack
- Using R, run the environmental health example with the following models and compute the BF for each:
  - i.  $y \sim X$
  - ii.  $\log(y) \sim X$
  - iii.  $y \sim X + X$
- Based on the Bayes Factor, what model would you choose for these data?

# Code for BF in the env. health example

```
Amm<- c(42,37,37,28,18,18,19,20,15,14,14,13,11,12,8,7,8,8,9,15)
Temp<- c(27,27,25,24,22,23,24,24,23,18,18,17,18,19,18,18,19,19,20,20 )

post1 <- MCMCregress(Amm~Temp, b0=1, B0=1e-6,
  marginal.likelihood="Chib95")
post2 <- MCMCregress(log(Amm)~Temp, b0=0, B0=1e-6,
  marginal.likelihood="Chib95")
Temp2 <- Temp**2
post3 <- MCMCregress(Amm~Temp+Temp2, b0=1, B0=1e-6,
  marginal.likelihood="Chib95")

raftery.diag(post1); raftery.diag(post2); rafter.diag(post3)

summary(post1) ; summary(post2); summary(post3)
BayesFactor(post1, post2, post3)
```



# References

- C.P. **Robert** and **G. Casella**  
*Monte Carlo Statistical Methods*, 1999  
*Méthodes de Monte-Carlo avec R*, 2011
- C.P. **Robert** and **J.M. Marin** *Bayesian Core: A practical approach to computational Bayesian analysis*, 2007
- **W. R. Gilks** , **S. Richardson** and **D. J. Spiegelhalter**,  
*Markov chain Monte Carlo in practice*, 1996
- **P. J. Green**, *Reversible Jump Markov-chain Monte Carlo computation and Bayesian model determination*, *Biometrika*, 1995