

**Lecture 1.**  
**Introduction to Bayesian Monte Carlo**  
**methods in WINBUGS**

## **Summary**

1. Probability as a means of representing uncertainty
2. Bayesian direct probability statements about parameters
3. Probability distributions
4. Monte Carlo simulation
5. Implementation in WinBUGS (and DoodleBUGS) - Demo
6. Directed graphs for representing probability models
7. Examples

How did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

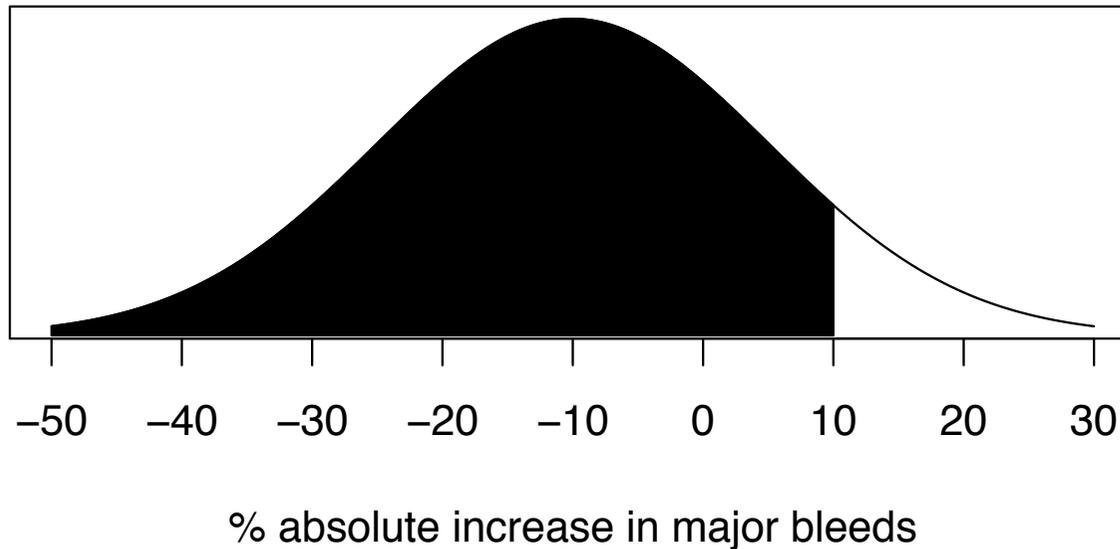
## P R O B L E M.

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

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

## Basic idea: Direct expression of uncertainty about unknown parameters

eg "There is an 89% probability that the absolute increase in major bleeds is less than 10 percent with low-dose PLT transfusions" (Tinmouth et al, Transfusion, 2004)



## **Why a direct probability distribution?**

1. Tells us what we want: what are plausible values for the parameter of interest?
2. No *P-values*: just calculate relevant tail areas
3. No (difficult to interpret) *confidence intervals*: just report, say, central area that contains 95% of distribution
4. Easy to make predictions (see later)
5. Fits naturally into decision analysis / cost-effectiveness analysis / project prioritisation
6. There is a procedure for adapting the distribution in the light of additional evidence: i.e. *Bayes theorem* allows us to learn from experience

## Inference on proportions

What is a reasonable form for a prior distribution for a proportion?

$\theta \sim \text{Beta}[a, b]$  represents a beta distribution with properties:

$$p(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}; \quad \theta \in (0, 1)$$

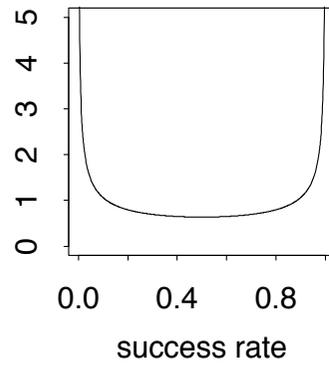
$$E(\theta|a, b) = \frac{a}{a+b}$$

$$V(\theta|a, b) = \frac{ab}{(a+b)^2(a+b+1)} :$$

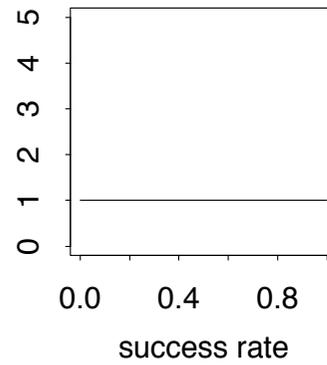
WinBUGS notation: `theta ~ dbeta(a,b)`

## Beta distribution

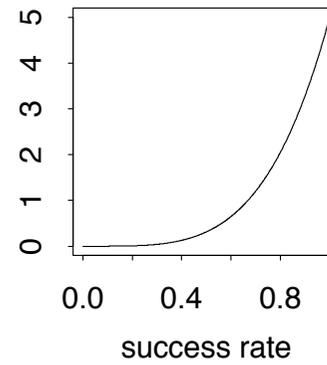
Beta(0.5,0.5)



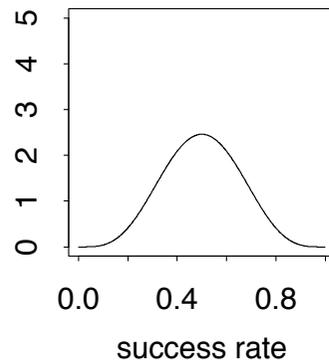
Beta(1,1)



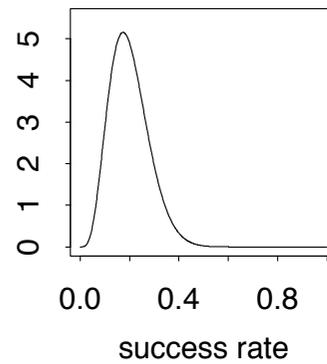
Beta(5,1)



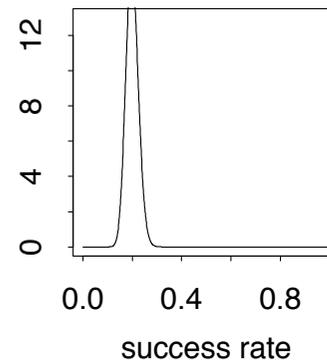
Beta(5,5)



Beta(5,20)

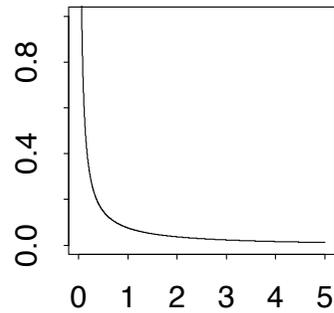


Beta(50,200)

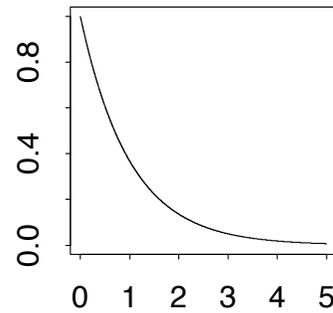


## Gamma distribution

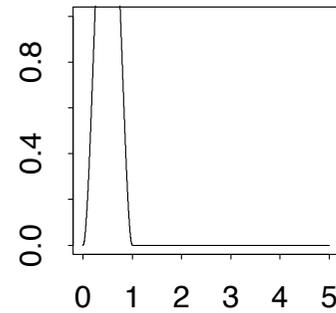
Gamma(0.1,0.1)



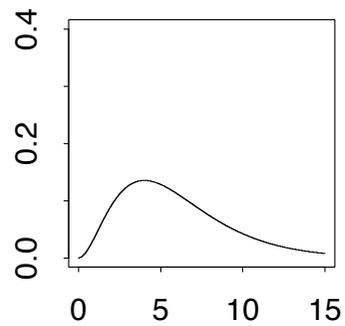
Gamma(1,1)



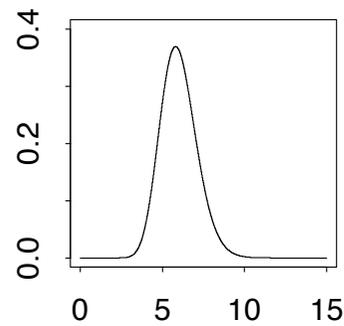
Gamma(3,3)



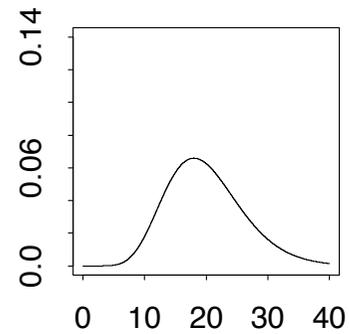
Gamma(3,0.5)



Gamma(30,5)



Gamma(10,0.5)



## The Gamma distribution

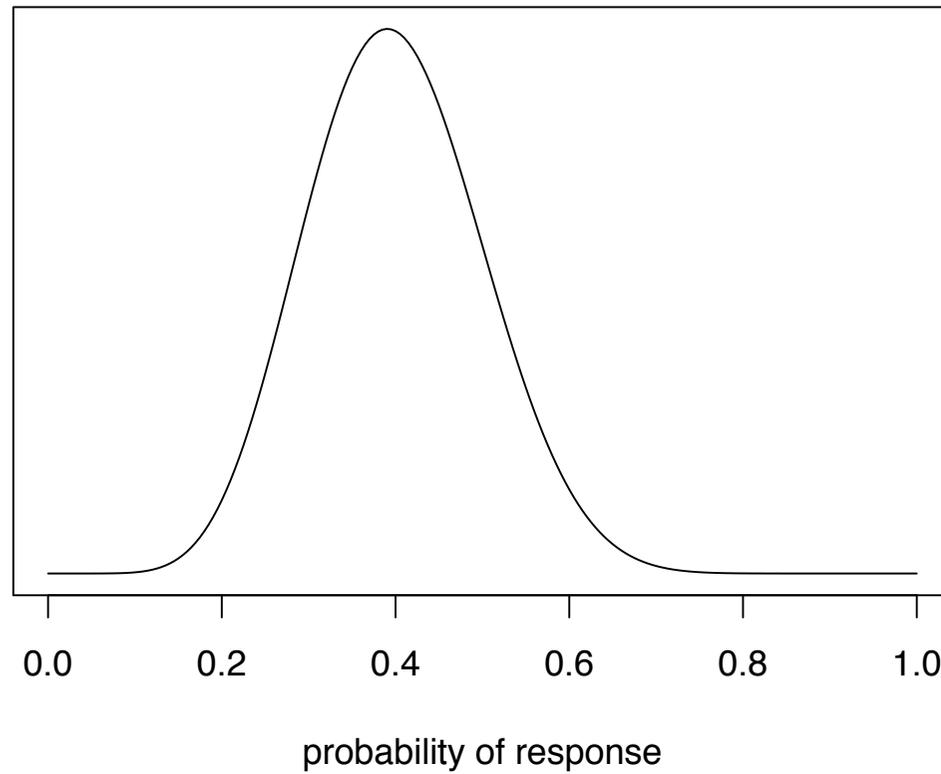
Flexible distribution for positive quantities. If  $Y \sim \text{Gamma}[a, b]$

$$\begin{aligned} p(y|a, b) &= \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}; & y \in (0, \infty) \\ E(Y|a, b) &= \frac{a}{b} \\ V(Y|a, b) &= \frac{a}{b^2}. \end{aligned}$$

- Gamma[1,  $b$ ] distribution is exponential with mean  $1/b$
- Gamma[ $\frac{v}{2}, \frac{1}{2}$ ] is a Chi-squared  $\chi_v^2$  distribution on  $v$  degrees of freedom
- $Y \sim \text{Gamma}[0.001, 0.001]$  means that  $p(y) \propto 1/y$ , or that  $\log Y \approx \text{Uniform}$
- Used as conjugate prior distribution for inverse variances (precisions)
- Used as sampling distribution for skewed positive valued quantities (alternative to log normal likelihood) — MLE of mean is sample mean
- WinBUGS notation:  $y \sim \text{dgamma}(a, b)$

## **Example: Drug**

- Consider a drug to be given for relief of chronic pain
- Experience with similar compounds has suggested that annual response rates between 0.2 and 0.6 could be feasible
- Interpret this as a distribution with mean = 0.4, standard deviation 0.1
- A Beta[9.2,13.8] distribution has these properties



Beta[9.2, 13.8] prior distribution supporting response rates between 0.2 and 0.6,

## **Making predictions**

Before observing a quantity  $Y$ , can provide its predictive distribution by integrating out unknown parameter

$$p(Y) = \int p(Y|\theta)p(\theta)d\theta.$$

Predictions are useful in e.g. cost-effectiveness models, design of studies, checking whether observed data is compatible with expectations, and so on.

If

$$\begin{aligned}\theta &\sim \text{Beta}[a, b] \\ Y_n &\sim \text{Binomial}(\theta, n),\end{aligned}$$

the exact predictive distribution for  $Y_n$  is known as the **Beta-Binomial**. It has the complex form

$$p(y_n) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \binom{n}{y_n} \frac{\Gamma(a+y_n)\Gamma(b+n-y_n)}{\Gamma(a+b+n)}.$$

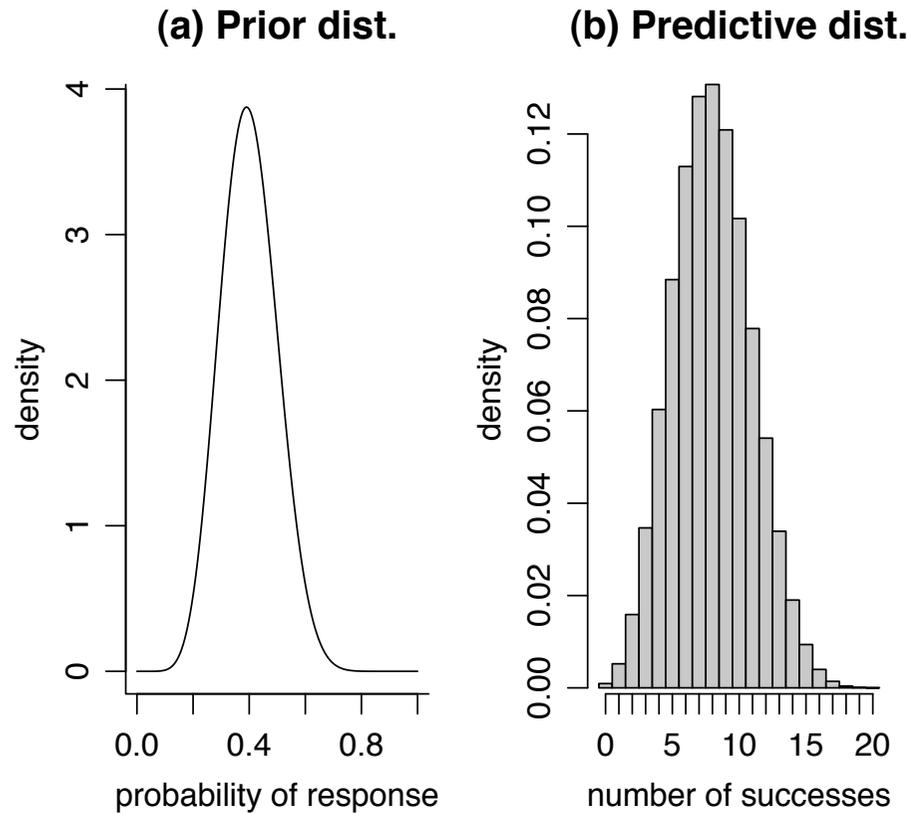
$$E(Y_n) = n \frac{a}{a+b}$$

If  $a = b = 1$  (Uniform distribution),  $p(y_n)$  is uniform over  $0, 1, \dots, n$ .

But in WinBUGS we can just write

```
theta ~ dbeta(a,b)
Y      ~ dbin(theta,n)
```

and the integration is automatically carried out and does not require algebraic cleverness.



(a) is the Beta prior distribution

(b) is the predictive Beta-Binomial distribution of the number of successes  $Y$  in the next 20 trials

From Beta-binomial distribution, can calculate  $P(Y_n \geq 15) = 0.015$ .

**Example: a Monte Carlo approach to estimating tail-areas of distributions**

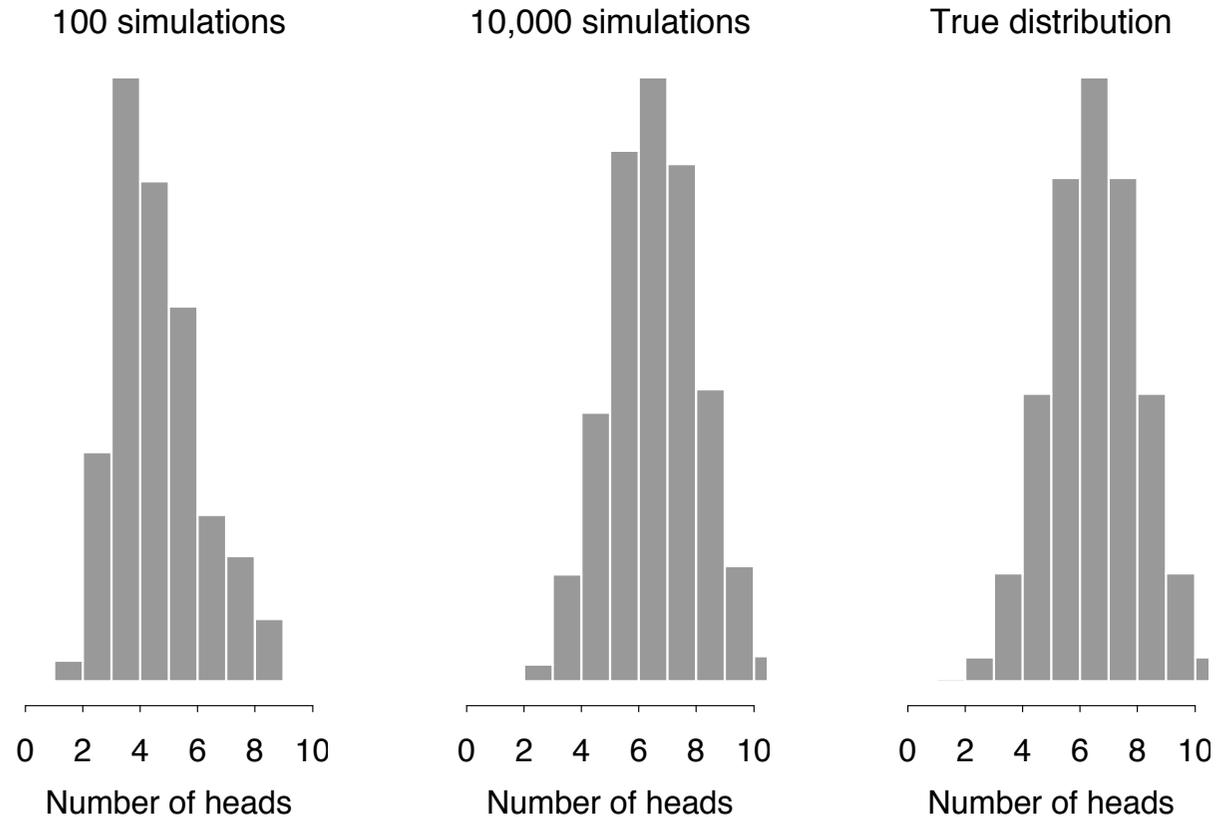
Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times.

An *algebraic* approach:

$$\begin{aligned}\Pr(\geq 8 \text{ heads}) &= \sum_{z=8}^{10} p\left(z \mid \pi = \frac{1}{2}, n = 10\right) \\ &= \binom{10}{8} \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 + \binom{10}{9} \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 + \binom{10}{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0 \\ &= 0.0547.\end{aligned}$$

A *physical* approach would be to repeatedly throw a set of 10 coins and count the proportion of throws that there were 8 or more heads.

A *simulation* approach uses a computer to toss the coins!



Proportion with 8 or more 'heads' in 10 tosses:

(a) After 100 'throws' (0.02); (b) after 10,000 throws (0.0577); (c) the true Binomial distribution (0.0547)

## **General Monte Carlo analysis - 'forward sampling'**

Used extensively in risk modelling - can think of as 'adding uncertainty' to a spreadsheet

- Suppose have logical function  $f$  containing uncertain parameters
- Can express our uncertainty as a prior distribution
- Simulate many values from this prior distribution
- Calculate  $f$  at the simulated values ('iterations')
- Obtain an empirical predictive distribution for  $f$
- Sometimes termed *probabilistic sensitivity analysis*
- Can do in Excel add-ons such as @RISK or Crystal Ball.

## The BUGS program

### Bayesian inference using Gibbs sampling

- Language for specifying complex Bayesian models
- Constructs object-oriented internal representation of the model
- Simulation from full conditionals using Gibbs sampling
- Current version (WinBUGS 1.4) runs in Windows
- 'Classic' BUGS available for UNIX but this is an old version

**WinBUGS is freely available from** <http://www.mrc-bsu.cam.ac.uk/bugs>

- Scripts enable WinBUGS 1.4 to run in batch mode or be called from other software
- Interfaces developed for R, Excel, Splus, SAS, Matlab
- OpenBUGS site <http://www.rni.helsinki.fi/openbugs> provides an open source version

## **Running WinBUGS for Monte Carlo analysis (no data)**

1. Open *Specification tool* from *Model* menu.
2. Program responses are shown on bottom-left of screen.
3. Highlight `model` by double-click. Click on *Check model*.
4. Click on *Compile*.
5. Click on *Gen Inits*.
6. Open *Update* from *Model* menu, and *Samples* from *Inference* menu.
7. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.
8. Type \* into *Sample Monitor*, and click *trace* to see sampled values.
9. Click on *Update* to generate samples.
10. Type \* into *Sample Monitor*, and click *stats* etc to see results on all monitored nodes.

## Using WinBUGS for Monte Carlo

The model for the 'coin' example is

$$Y \sim \text{Binomial}(0.5, 10)$$

and we want to know  $P(Y \geq 8)$ .

This model is represented in the BUGS language as

```
model{  
  Y      ~  dbin(0.5,10)  
  P8    <-  step(Y-7.5)  
}
```

P8 is a step function which will take on the value 1 if  $Y - 7.5$  is  $\geq 0$ , *i.e.* Y is 8 or more, 0 if 7 or less.

Running this simulation for 100, 10000 and 1000000 iterations, and then taking the empirical mean of P8, provided the previous estimated probabilities that Y will be 8 or more.

## Some aspects of the BUGS language

- `<-` represents logical dependence, *e.g.* `m <- a + b*x`
- `~` represents stochastic dependence, *e.g.* `r ~ dunif(a,b)`
- Can use arrays and loops

```
for (i in 1:n){  
  r[i] ~ dbin(p[i],n[i])  
  p[i] ~ dunif(0,1)  
}
```

- Some functions can appear on left-hand-side of an expression, *e.g.*

```
logit(p[i]) <- a + b*x[i]  
log(m[i]) <- c + d*y[i]
```

- `mean(p[])` to take mean of whole array, `mean(p[m:n])` to take mean of elements `m` to `n`. Also for `sum(p[])`.
- `dnorm(0,1)I(0,)` means the prior will be restricted to the range  $(0, \infty)$ .

## Functions in the BUGS language

- `p <- step(x-.7)` = 1 if  $x \geq 0.7$ , 0 otherwise. Hence monitoring `p` and recording its mean will give the probability that  $x \geq 0.7$ .
- `p <- equals(x,.7)` = 1 if  $x = 0.7$ , 0 otherwise.
- `tau <- 1/pow(s,2)` sets  $\tau = 1/s^2$ .
- `s <- 1/ sqrt(tau)` sets  $s = 1/\sqrt{\tau}$ .
- `p[i,k] <- inprod(pi[], Lambda[i,,k])` sets  $p_{ik} = \sum_j \pi_j \Lambda_{ijk}$ . `inprod2` may be faster.
- See 'Model Specification/Logical nodes' in manual for full syntax.

## Some common Distributions

### Expression Distribution Usage

---

dbin	binomial	$r \sim \text{dbin}(p,n)$
dnorm	normal	$x \sim \text{dnorm}(\mu,\tau)$
dpois	Poisson	$r \sim \text{dpois}(\lambda)$
dunif	uniform	$x \sim \text{dunif}(a,b)$
dgamma	gamma	$x \sim \text{dgamma}(a,b)$

NB. The normal is parameterised in terms of its mean and *precision* =  $1/\text{variance}$  =  $1/\text{sd}^2$ .

See 'Model Specification/The BUGS language: stochastic nodes/Distributions' in manual for full syntax.

**Functions cannot be used as arguments in distributions (you need to create new nodes).**

## Drug example: Monte Carlo predictions

Our prior distribution for proportion of responders in one year  $\theta$  was Beta[9.2, 13.8].

Consider situation *before* giving 20 patients the treatment. What is the chance if getting 15 or more responders?

$\theta \sim \text{Beta}[9.2, 13.8]$  prior distribution

$y \sim \text{Binomial}[\theta, 20]$  sampling distribution

$P_{\text{crit}} = P(y \geq 15)$  Probability of exceeding critical threshold

# In BUGS syntax:

```
model{
theta      ~ dbeta(9.2,13.8)      # prior distribution
y          ~ dbin(theta,20)      # sampling distribution
P.crit    <- step(y-14.5)        # =1 if y >= 15, 0 otherwise
}
```

## WinBUGS output and exact answers

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta	0.4008	0.09999	9.415E-4	0.2174	0.3981	0.6044	1	10000
y	8.058	2.917	0.03035	3.0	8.0	14.0	1	10000
P.crit	0.0151	0.122	0.001275	0.0	0.0	0.0	1	10000

Note that the mean of the 0-1 indicator P.crit provides the estimated tail-area probability.

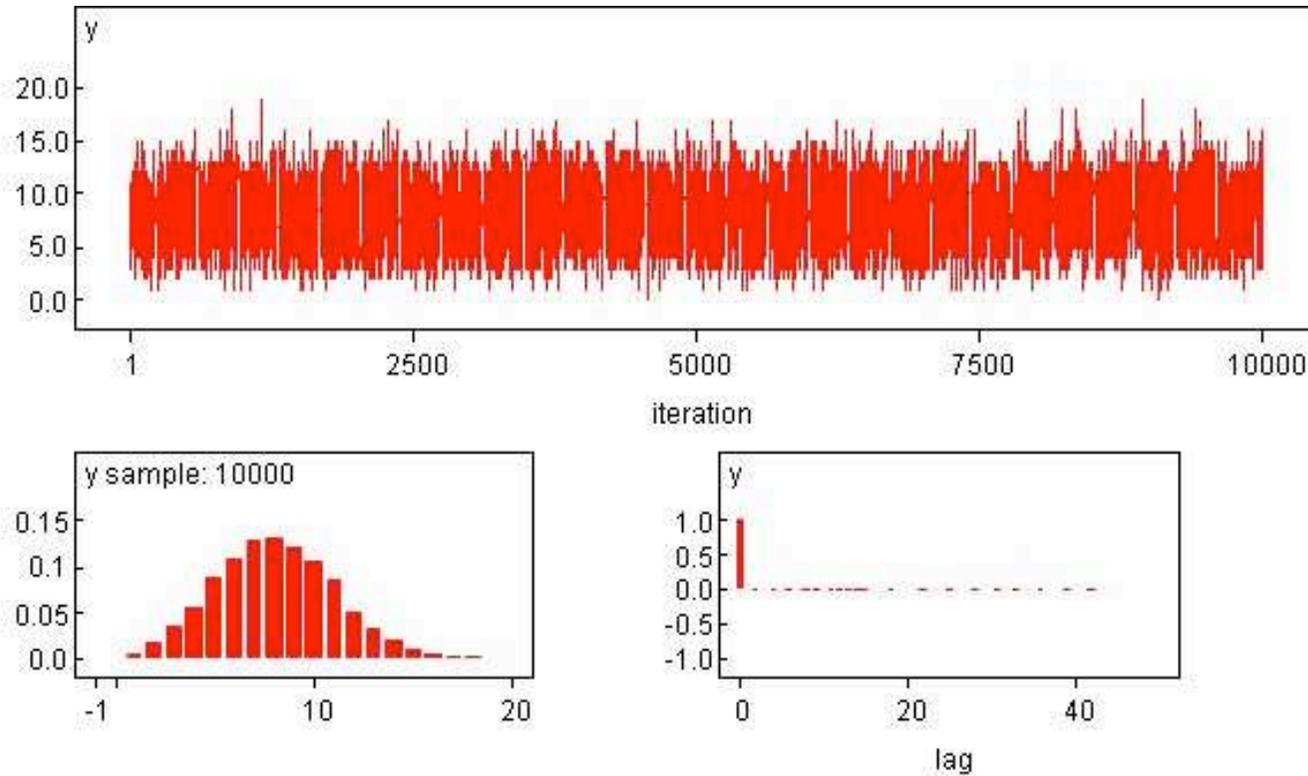
Exact answers from closed-form analysis:

- $\theta$ : mean 0.4 and standard deviation 0.1
- $y$ : mean 8 and standard deviation 2.93.
- Probability of at least 15: 0.015

These are independent samples, and so MC error =  $SD/\sqrt{\text{No.iterations}}$ .

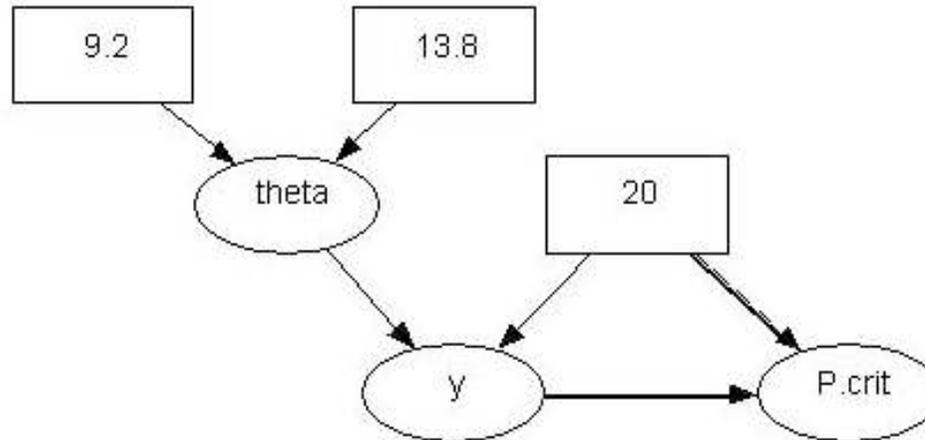
Can achieve arbitrary accuracy by running the simulation for longer.

## WinBUGS output



Independent samples, and so no auto-correlation and no concern with convergence.

## Graphical representation of models



- *Doodle* represents each quantity as a node in directed acyclic graph (DAG).
- Constants are placed in rectangles, random quantities in ovals
- Stochastic dependence is represented by a single arrow, and logical function as double arrow
- WinBUGS allows models to be specified graphically and run directly from the graphical interface
- Can write code from Doodles
- Good for explanation, but can be tricky to set up

## Script for running Drug Monte Carlo example

Run from Model/Script menu

```
display('log')          # set up log file
check('c:/bugscourse/drug-MC')      # check syntax of model
# data('c:/bugscourse/drug-data')   # load data file if there is one
compile(1)                # generate code for 1 simulations
# inits(1,'c:/bugscourse/drug-in1')  # load initial values if necessary
gen.inits()               # generate initial values for all unknown quantities
                          # not given initial values

set(theta)                # monitor the true response rate
set(y)                    # monitor the predicted number of successes
set(P.crit)               # monitor whether a critical number of successes occur
trace(*)                  # watch some simulated values (although slows down simulation)
update(10000)             # perform 10000 simulations
history(theta)            # Trace plot of samples for theta
stats(*)                  # Calculate summary statistics for all monitored quantities
density(theta)            # Plot distribution of theta
density(y)                # Plot distribution of y
```

**Example: Power — uncertainty in a power calculation**

- a randomised trial planned with  $n$  patients in each of two arms
- response with standard deviation  $\sigma = 1$
- aimed to have Type 1 error 5% and 80% power
- to detect a true difference of  $\theta = 0.5$  in mean response between the groups

Necessary sample size per group is

$$n = \frac{2\sigma^2}{\theta^2}(0.84 + 1.96)^2 = 63$$

Alternatively, for fixed  $n$ , the power is

$$\text{Power} = \Phi \left( \sqrt{\frac{n\theta^2}{2\sigma^2}} - 1.96 \right).$$

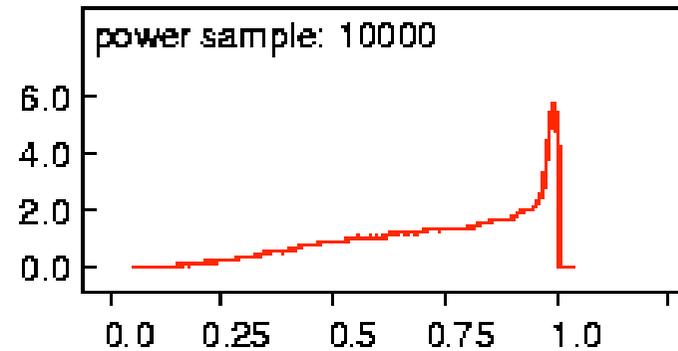
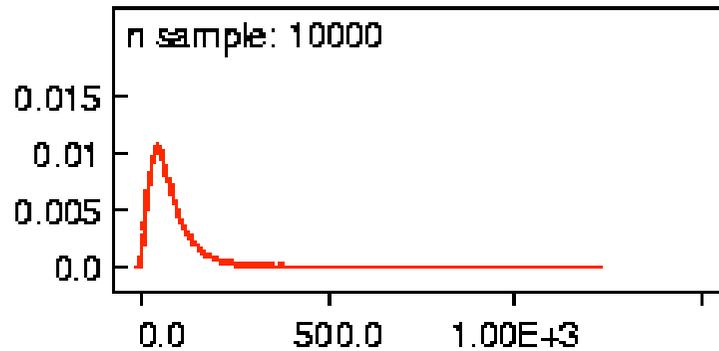
Suppose we wish to express uncertainty concerning both  $\theta$  and  $\sigma$ , e.g.

$$\theta \sim N[0.5, 0.1^2], \quad \sigma \sim N[1, 0.3^2].$$

1. Simulate values of  $\theta$  and  $\sigma$  from their prior distributions
2. Substitute them in the formulae
3. Obtain a predictive distribution over  $n$  or Power

```
prec.sigma <- 1/(0.3*0.3)      # transform sd to precision=1/sd2
prec.theta <- 1/(0.1*0.1)
sigma      ~ dnorm(1, prec.sigma)I(0,)
theta      ~ dnorm(.5, prec.theta)I(0,)
n          <- 2 * pow( (.84 +1.96) * sigma / theta , 2 )
power      <- phi( sqrt(63/2)* theta /sigma -1.96 )
prob70     <-step(power-.7)
```

	Median	95% interval
$n$	62.5	9.3 to 247.2
Power (%)	80	29 to 100



For  $n=63$ , the median power is 80%, and a trial of 63 patients per group could be seriously underpowered. There is a 37% chance that the power is less than 70%.

**Lecture 2.**  
**Introduction to conjugate Bayesian  
inference**

## **What are Bayesian methods?**

- Bayesian methods have been widely applied in many areas:
  - medicine / epidemiology
  - genetics
  - ecology
  - environmental sciences
  - social and political sciences
  - finance
  - archaeology
  - .....
- Motivations for adopting Bayesian approach vary:
  - natural and coherent way of thinking about science and learning
  - pragmatic choice that is suitable for the problem in hand

Spiegelhalter et al (2004) define a Bayesian approach as

‘the explicit use of external evidence in the design, monitoring, analysis, interpretation and reporting of a [scientific investigation]’

They argue that a Bayesian approach is:

- more flexible in adapting to each unique situation
- more efficient in using all available evidence
- more useful in providing relevant quantitative summaries

than traditional methods

## **Example**

A clinical trial is carried out to collect evidence about an unknown 'treatment effect'

### *Conventional analysis*

- p-value for  $H_0$ : treatment effect is zero
- Point estimate and CI as summaries of size of treatment effect

Aim is to learn what this trial tells us about the treatment effect

### *Bayesian analysis*

- Asks: 'how should this trial change our opinion about the treatment effect?'

The Bayesian analyst needs to explicitly state

- a reasonable opinion concerning the plausibility of different values of the treatment effect *excluding* the evidence from the trial (the **prior distribution**)
- the support for different values of the treatment effect based *solely* on data from the trial (the **likelihood**),

and to combine these two sources to produce

- a final opinion about the treatment effect (the **posterior distribution**)

The final combination is done using Bayes theorem, which essentially weights the likelihood from the trial with the relative plausibilities defined by the prior distribution

One can view the Bayesian approach as a formalisation of the process of learning from experience

Posterior distribution forms basis for all inference — can be summarised to provide

- point and interval estimates of treatment effect
- point and interval estimates of any function of the parameters
- probability that treatment effect exceeds a clinically relevant value
- prediction of treatment effect in a new patient
- prior information for future trials
- inputs for decision making
- ....

## Bayes theorem and its link with Bayesian inference

**Bayes' theorem** Provable from probability axioms

Let  $A$  and  $B$  be events, then

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

If  $A_i$  is a set of mutually exclusive and exhaustive events (*i.e.*  $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$ ), then

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)}.$$

### Example: use of Bayes theorem in diagnostic testing

- A new HIV test is claimed to have “95% sensitivity and 98% specificity”
- In a population with an HIV prevalence of 1/1000, what is the chance that patient testing positive actually has HIV?

Let  $A$  be the event that patient is truly HIV positive,  $\bar{A}$  be the event that they are truly HIV negative.

Let  $B$  be the event that they test positive.

We want  $p(A|B)$ .

“95% sensitivity” means that  $p(B|A) = .95$ .

“98% specificity” means that  $p(B|\bar{A}) = .02$ .

Now Bayes theorem says

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\bar{A})p(\bar{A})}.$$

$$\text{Hence } p(A|B) = \frac{.95 \times .001}{.95 \times .001 + .02 \times .999} = .045.$$

Thus over 95% of those testing positive will, in fact, not have HIV.

- Our intuition is poor when processing probabilistic evidence
- The vital issue is *how should this test result change our belief that patient is HIV positive?*
- The disease prevalence can be thought of as a '*prior*' probability ( $p = 0.001$ )
- Observing a positive result causes us to modify this probability to  $p = 0.045$ . This is our '*posterior*' probability that patient is HIV positive.
- Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established
- More controversial is the use of Bayes theorem in general statistical analyses, where *parameters* are the unknown quantities, and their prior distribution needs to be specified — this is **Bayesian inference**

## **Bayesian inference**

Makes fundamental distinction between

- Observable quantities  $x$ , i.e. the data
- Unknown quantities  $\theta$

$\theta$  can be statistical parameters, missing data, mismeasured data ...

→ parameters are treated as random variables

→ in the Bayesian framework, we make probability statements about model parameters

! in the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data

As with any statistical analysis, we start by positing a model which specifies

$$p(x | \theta)$$

This is the **likelihood**, which relates all variables into a '**full probability model**'

From a Bayesian point of view

- $\theta$  is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data  
→ need to specify a **prior distribution**  $p(\theta)$
- $x$  is known so we should condition on it  
→ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\theta | x) = \frac{p(\theta) p(x | \theta)}{\int p(\theta) p(x | \theta) d\theta} \propto p(\theta) p(x | \theta)$$

This is the **posterior distribution**

The prior distribution  $p(\theta)$ , expresses our uncertainty about  $\theta$  **before** seeing the data.

The posterior distribution  $p(\theta | x)$ , expresses our uncertainty about  $\theta$  **after** seeing the data.

## Inference on proportions using a continuous prior

Suppose we now observe  $r$  positive responses out of  $n$  patients.

Assuming patients are independent, with common unknown response rate  $\theta$ , leads to a binomial likelihood

$$p(r|n, \theta) = \binom{n}{r} \theta^r (1 - \theta)^{n-r} \propto \theta^r (1 - \theta)^{n-r}$$

$\theta$  needs to be given a continuous prior distribution.

Suppose that, before taking account of the evidence from our trial, we believe all values for  $\theta$  are equally likely (is this plausible?)  $\Rightarrow \theta \sim \text{Unif}(0, 1)$  i.e.  $p(\theta) = \frac{1}{1-0} = 1$

Posterior is then

$$p(\theta|r, n) \propto \theta^r (1 - \theta)^{(n-r)} \times 1$$

This has form of the *kernel* of a Beta( $r+1$ ,  $n-r+1$ ) distribution (see lect 1), where

$$\theta \sim \text{Beta}(a, b) \equiv \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1}$$

To represent external evidence that some response rates are more plausible than others, it is mathematically convenient to use a  $\text{Beta}(a, b)$  prior distribution for  $\theta$

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

Combining this with the binomial likelihood gives a posterior distribution

$$\begin{aligned} p(\theta | r, n) &\propto p(r | \theta, n)p(\theta) \\ &\propto \theta^r(1 - \theta)^{n-r}\theta^{a-1}(1 - \theta)^{b-1} \\ &= \theta^{r+a-1}(1 - \theta)^{n-r+b-1} \\ &\propto \text{Beta}(r + a, n - r + b) \end{aligned}$$

## Comments

- When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood
  - Occurs when prior and likelihood have the same ‘kernel’

- Recall from lecture 1 that a  $\text{Beta}(a, b)$  distribution has

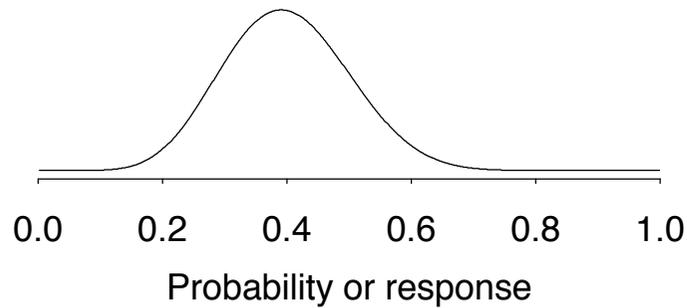
$$\begin{aligned}\text{mean} &= a/(a + b), \\ \text{variance} &= ab/[(a + b)^2(a + b + 1)]\end{aligned}$$

Hence posterior mean is  $E(\theta|r, n) = (r + a)/(n + a + b)$

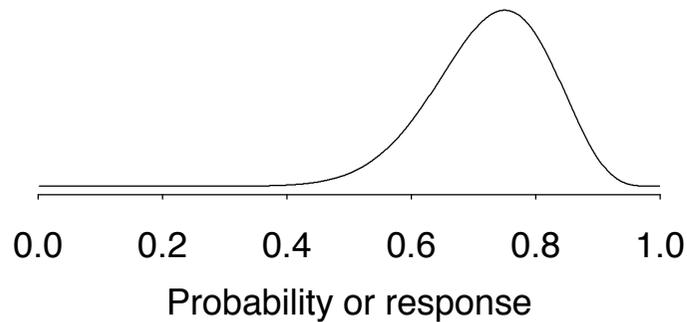
- $a$  and  $b$  are equivalent to observing a priori  $a - 1$  successes in  $a + b - 2$  trials  
→ can be elicited
- With fixed  $a$  and  $b$ , as  $r$  and  $n$  increase,  $E(\theta|r, n) \rightarrow r/n$  (the MLE), and the variance tends to zero
  - This is a general phenomenon: as  $n$  increases, posterior distribution gets more concentrated and the likelihood dominates the prior
- A  $\text{Beta}(1, 1)$  is equivalent to  $\text{Uniform}(0, 1)$

## **Example: Drug**

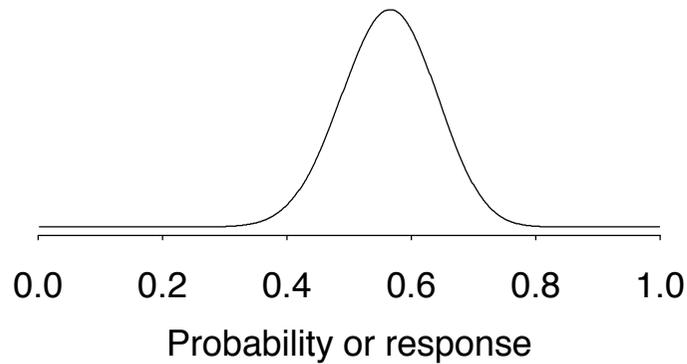
- Recall example from lecture 1, where we consider early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- We interpreted this as a distribution with mean = 0.4, standard deviation 0.1 and showed that a Beta(9.2,13.8) distribution has these properties
- Suppose we now treat  $n = 20$  volunteers with the compound and observe  $y = 15$  positive responses



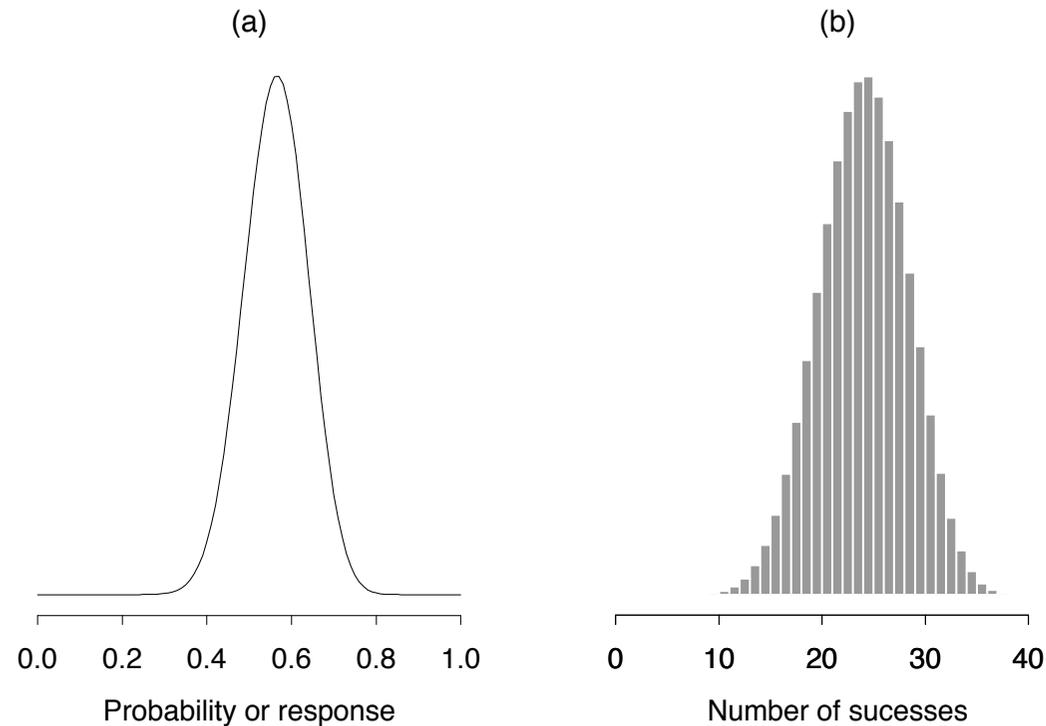
Beta(9.2, 13.8) prior distribution supporting response rates between 0.2 and 0.6



Likelihood arising from a Binomial observation of 15 successes out of 20 cases



Parameters of the Beta distribution are updated to  $(a+15, b+20-15) = (24.2, 18.8)$ :  
mean  $24.2/(24.2+18.8) = 0.56$



(a) Beta posterior distribution after having observed 15 successes in 20 trials

(b) predictive Beta-Binomial distribution of the number of successes  $\tilde{y}_{40}$  in the next 40 trials with mean 22.5 and standard deviation 4.3

Suppose we would consider continuing a development program if the drug managed to achieve at least a further 25 successes out of these 40 future trials

From Beta-binomial distribution, can calculate  $P(\tilde{y}_{40} \geq 25) = 0.329$

## **Drug (continued): learning about parameters from data using Markov chain Monte-Carlo (MCMC) methods in WinBUGS**

- Using MCMC (e.g. in WinBUGS), no need to explicitly specify posterior
- Can just specify the prior and likelihood separately
- WinBUGS contains algorithms to evaluate the posterior given (almost) arbitrary specification of prior and likelihood
  - posterior doesn't need to be closed form
  - but can (usually) recognise conjugacy when it exists

The drug model can be written

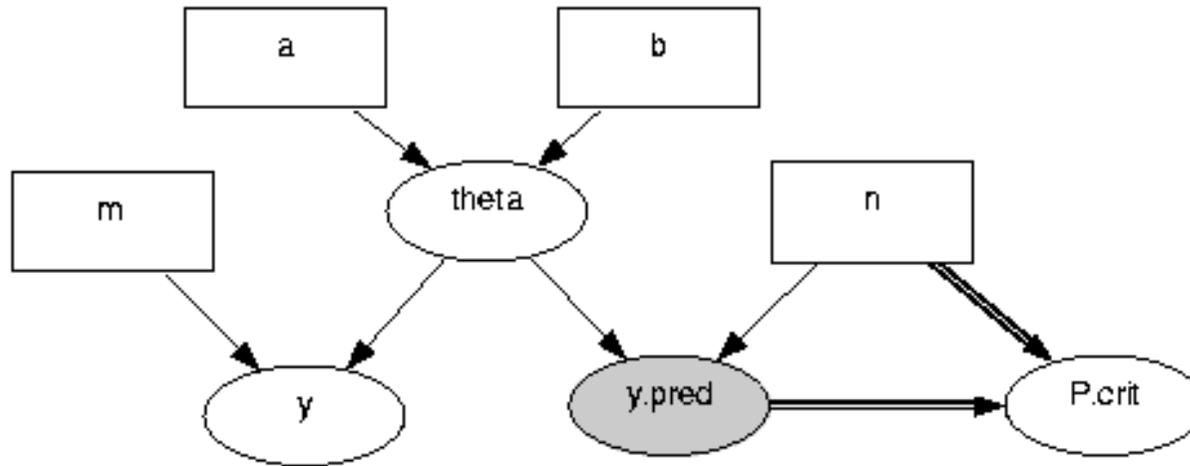
$\theta \sim \text{Beta}[a, b]$       prior distribution  
 $y \sim \text{Binomial}[\theta, m]$    sampling distribution  
 $y_{\text{pred}} \sim \text{Binomial}[\theta, n]$    predictive distribution  
 $P_{\text{crit}} = P(y_{\text{pred}} \geq n_{\text{crit}})$    Probability of exceeding critical threshold

# In BUGS syntax:

```
# Model description '  
model {  
  theta      ~ dbeta(a,b)          # prior distribution  
  y          ~ dbin(theta,m)      # sampling distribution  
  y.pred     ~ dbin(theta,n)      # predictive distribution  
  P.crit     <- step(y.pred-ncrit+0.5) # =1 if y.pred >= ncrit, 0 otherwise  
}
```

## Graphical representation of models

name: y.pred    type: stochastic    density: dbin  
proportion    theta    order    n    lower bound    upper bound



Note that adding data to a model is simply extending the graph.

## Data files

Data can be written after the model description, or held in a separate .txt or .odc file

```
list( a = 9.2,      # parameters of prior distribution
      b = 13.8,
      y = 15,      # number of successes
      m = 20,      # number of trials
      n = 40,      # future number of trials
      ncrit = 25) # critical value of future successes
```

Alternatively, in this simple example, we could have put all data and constants into model description:

```
model{
  theta      ~ dbeta(9.2,13.8)      # prior distribution
  y          ~ dbin(theta,20)      # sampling distribution
  y.pred     ~ dbin(theta,40)      # predictive distribution
  P.crit     <- step(y.pred-24.5)   # =1 if y.pred >= ncrit, 0 otherwise

  y         <- 15
}
```

## **The WinBUGS data formats**

WinBUGS accepts data files in:

1. Rectangular format (easy to cut and paste from spreadsheets)

```
n[] r[]
47  0
148 18
...
360 24
END
```

2. S-Plus format:

```
list(N=12,n = c(47,148,119,810,211,196,
               148,215,207,97,256,360),
      r = c(0,18,8,46,8,13,9,31,14,8,29,24))
```

Generally need a 'list' to give size of datasets etc.

## Initial values

- WinBUGS can automatically generate initial values for the MCMC analysis using *gen inits*
- Fine if have informative prior information
- If have fairly 'vague' priors, better to provide reasonable values in an initial-values list

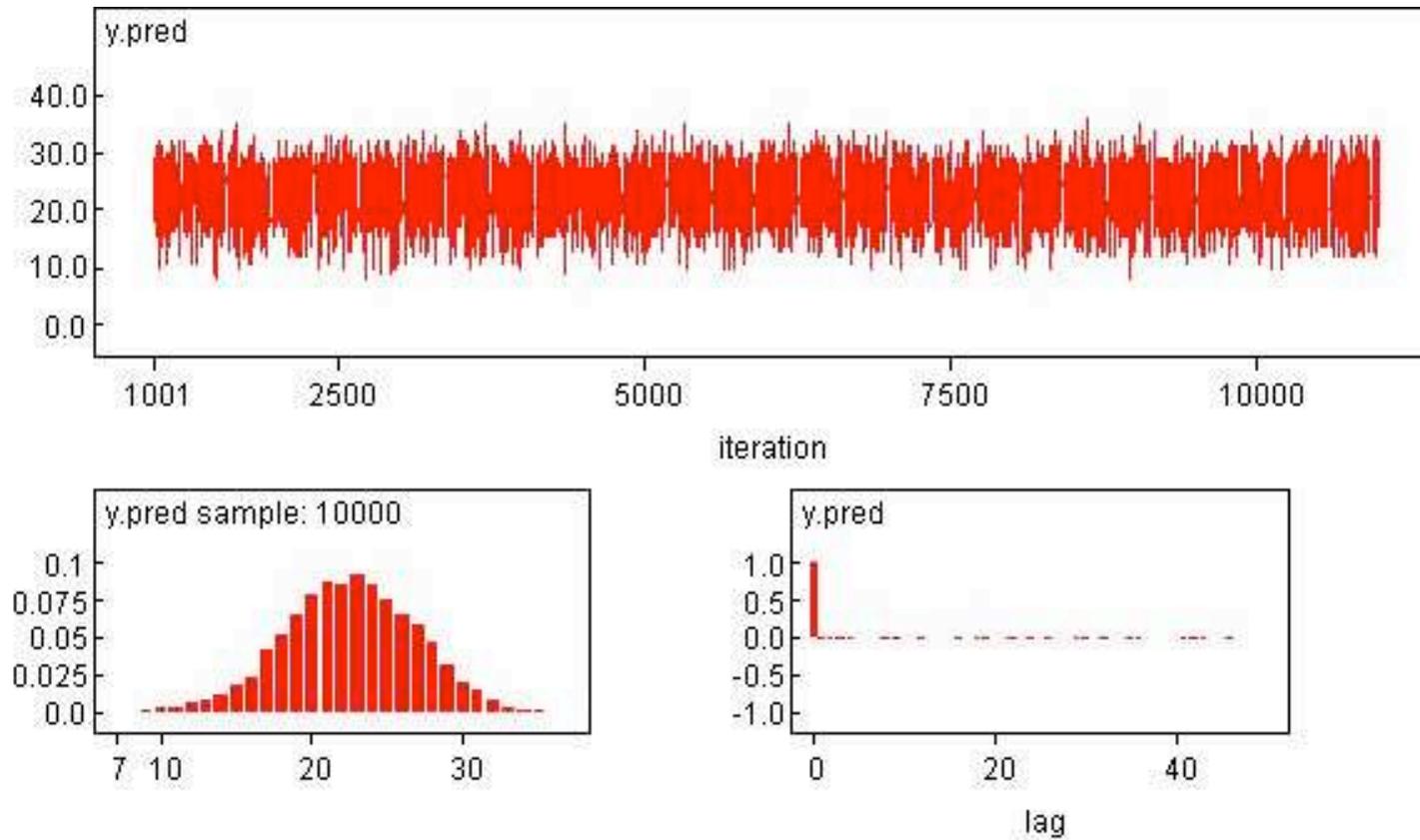
Initial values list can be after model description or in a separate file

```
list(theta=0.1)
```

## Running WinBUGS for MCMC analysis (single chain)

1. Open *Specification tool* from *Model* menu.
2. Program responses are shown on bottom-left of screen.
3. Highlight `model` by double-click. Click on *Check model*.
4. Highlight start of data. Click on *Load data*.
5. Click on *Compile*.
6. Highlight start of initial values. Click on *Load inits*.
7. Click on *Gen Inits* if more initial values needed.
8. Open *Update* from *Model* menu.
9. Click on *Update* to burn in.
10. Open *Samples* from *Inference* menu.
11. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.
12. Perform more updates.
13. Type `*` into *Sample Monitor*, and click *stats* etc to see results on all monitored nodes.

## WinBUGS output



## WinBUGS output and exact answers

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta	0.5633	0.07458	4.292E-4	0.4139	0.5647	0.7051	1001	30000
y.pred	22.52	4.278	0.02356	14.0	23.0	31.0	1001	30000
P.crit	0.3273	0.4692	0.002631	0.0	0.0	1.0	1001	30000

Exact answers from conjugate analysis

- $\theta$ : mean 0.563 and standard deviation 0.075
- $Y^{\text{pred}}$ : mean 22.51 and standard deviation 4.31.
- Probability of at least 25: 0.329

MCMC results are within Monte Carlo error of the true values

## **Bayesian inference using the Normal distribution**

### **Known variance, unknown mean**

Suppose we have a sample of Normal data  $x_i \sim N(\theta, \sigma^2)$  ( $i = 1, \dots, n$ ).

For now assume  $\sigma^2$  is known and  $\theta$  has a Normal prior  $\theta \sim N(\mu, \sigma^2/n_0)$

Same standard deviation  $\sigma$  is used in the likelihood and the prior. Prior variance is based on an 'implicit' sample size  $n_0$

Then straightforward to show that the posterior distribution is

$$\theta|x \sim N\left(\frac{n_0\mu + n\bar{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

- As  $n_0$  tends to 0, the prior variance becomes larger and the distribution becomes 'flatter', and in the limit the prior distribution becomes essentially uniform over  $-\infty, \infty$
- Posterior mean  $(n_0\mu + n\bar{x})/(n_0 + n)$  is a weighted average of the prior mean  $\mu$  and parameter estimate  $\bar{x}$ , weighted by their precisions (relative 'sample sizes'), and so is always a compromise between the two
- Posterior variance is based on an implicit sample size equivalent to the sum of the prior 'sample size'  $n_0$  and the sample size of the data  $n$
- As  $n \rightarrow \infty$ ,  $p(\theta|\mathbf{x}) \rightarrow N(\bar{x}, \sigma^2/n)$  which does not depend on the prior
- Compare with frequentist setting, the MLE is  $\hat{\theta} = \bar{x}$  with  $SE(\hat{\theta}) = \sigma/\sqrt{n}$ , and sampling distribution

$$p(\hat{\theta} | \theta) = p(\bar{x}|\theta) = N(\theta, \sigma^2/n)$$

## Example: THM concentrations

- Regional water companies in the UK are required to take routine measurements of trihalomethane (THM) concentrations in tap water samples for regulatory purposes
- Samples are tested throughout the year in each water supply zone
- Suppose we want to estimate the average THM concentration in a particular water zone,  $z$
- Two independent measurements,  $x_{z1}$  and  $x_{z2}$  are taken and their mean,  $\bar{x}_z$  is  $130 \mu\text{g}/\text{l}$
- Suppose we know that the assay measurement error has a standard deviation  $\sigma_{[e]} = 5 \mu\text{g}/\text{l}$
- What should we estimate the mean THM concentration to be in this water zone?

Let the mean THM conc. be denoted  $\theta_z$ .

A standard analysis would use the sample mean  $\bar{x}_z = 130 \mu\text{g}/\text{l}$  as an estimate of  $\theta_z$ , with standard error  $\sigma_{[e]}/\sqrt{n} = 5/\sqrt{2} = 3.5 \mu\text{g}/\text{l}$

A 95% confidence interval is  $\bar{x}_z \pm 1.96 \times \sigma_{[e]}/\sqrt{n}$ , i.e. 123.1 to 136.9  $\mu\text{g}/\text{l}$ .

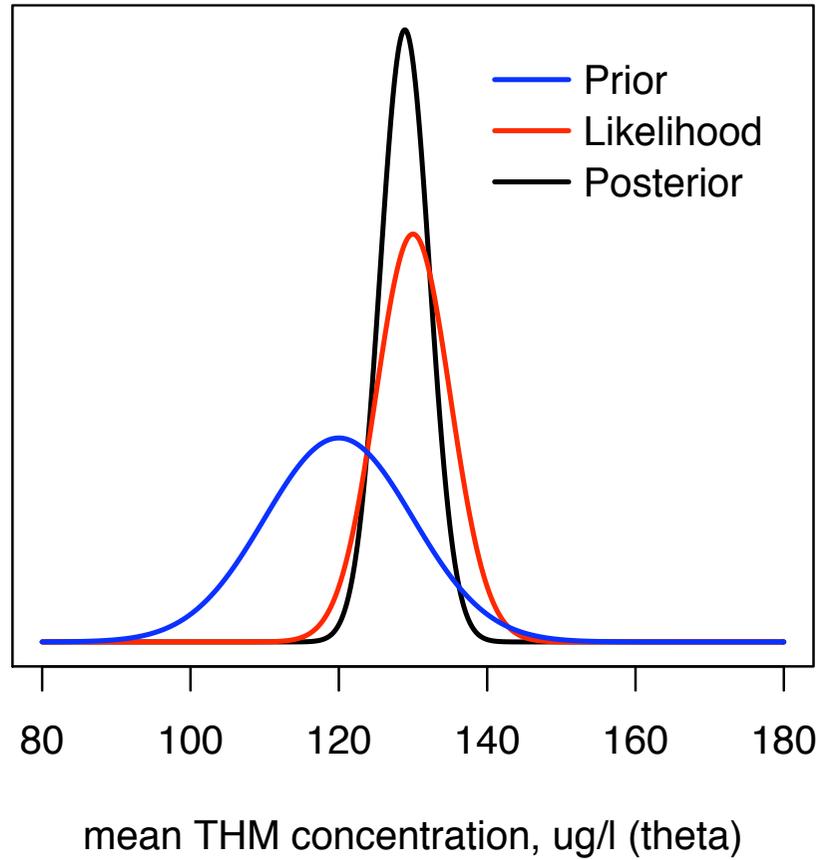
Suppose historical data on THM levels in other zones supplied from the same source showed that the mean THM concentration was  $120 \mu\text{g}/\text{l}$  with standard deviation  $10 \mu\text{g}/\text{l}$

- suggests  $\text{Normal}(120, 10^2)$  prior for  $\theta_z$
- if we express the prior standard deviation as  $\sigma_{[e]}/\sqrt{n_0}$ , we can solve to find  $n_0 = (\sigma_{[e]}/10)^2 = 0.25$
- so our prior can be written as  $\theta_z \sim \text{Normal}(120, \sigma_{[e]}^2/0.25)$

Posterior for  $\theta_z$  is then

$$\begin{aligned} p(\theta_z|\mathbf{x}) &= \text{Normal}\left(\frac{0.25 \times 120 + 2 \times 130}{0.25 + 2}, \frac{5^2}{0.25 + 2}\right) \\ &= \text{Normal}(128.9, 3.33^2) \end{aligned}$$

giving 95% interval for  $\theta_z$  of 122.4 to  $135.4 \mu\text{g}/\text{l}$



## Prediction

Denoting the posterior mean and variance as  $\mu_n = (n_0\mu + n\bar{x})/(n_0 + n)$  and  $\sigma_n^2 = \sigma^2/(n_0 + n)$ , the *predictive distribution* for a new observation  $\tilde{x}$  is

$$p(\tilde{x}|\mathbf{x}) = \int p(\tilde{x}|\mathbf{x}, \theta)p(\theta|\mathbf{x})d\theta$$

which generally simplifies to

$$p(\tilde{x}|\mathbf{x}) = \int p(\tilde{x}|\theta)p(\theta|\mathbf{x})d\theta$$

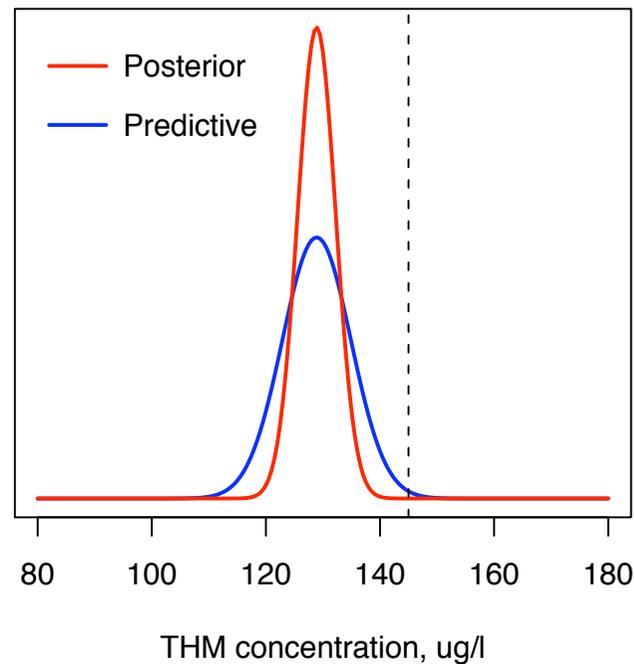
which can be shown to give

$$p(\tilde{x}|\mathbf{x}) \sim \text{N}(\mu_n, \sigma_n^2 + \sigma^2)$$

So the predictive distribution is centred around the posterior mean with variance equal to sum of the posterior variance and the sample variance of  $\tilde{x}$

### Example: THM concentration (continued)

- Suppose the water company will be fined if THM levels in the water supply exceed  $145\mu\text{g}/\text{l}$
- Predictive distribution for THM concentration in a future sample taken from the water zone is  $N(128.9, 3.33^2 + 5^2) = N(128.9, 36.1)$
- Probability that THM concentration in future sample exceeds  $145\mu\text{g}/\text{l}$  is  $1 - \Phi[(145 - 128.9)/\sqrt{(36.1)}] = 0.004$



## Bayesian inference using count data

Suppose we have an independent sample of counts  $x_1, \dots, x_n$  which can be assumed to follow a Poisson distribution with unknown mean  $\mu$ :

$$p(\mathbf{x}|\mu) = \prod_i \frac{\mu^{x_i} e^{-\mu}}{x_i!}$$

The kernel of the Poisson likelihood (as a function of  $\mu$ ) has the same form as that of a Gamma( $a, b$ ) prior for  $\mu$ :

$$p(\mu) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}$$

Note: A Gamma( $a, b$ ) density has mean  $a/b$  and variance  $a/b^2$

This implies the following posterior

$$\begin{aligned}
 p(\mu \mid \mathbf{x}) &\propto p(\mu) p(\mathbf{x} \mid \mu) \\
 &= \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu} \prod_{i=1}^n e^{-\mu} \frac{\mu^{x_i}}{x_i!} \\
 &\propto \mu^{a+n\bar{x}-1} e^{-(b+n)\mu} \\
 &= \text{Gamma}(a + n\bar{x}, b + n).
 \end{aligned}$$

The posterior is another (different) Gamma distribution.

The Gamma distribution is said to be the *conjugate* prior.

$$E(\mu \mid \mathbf{x}) = \frac{a + n\bar{x}}{b + n} = \bar{x} \left( \frac{n}{n + b} \right) + \frac{a}{b} \left( 1 - \frac{n}{n + b} \right)$$

So posterior mean is a compromise between the prior mean  $a/b$  and the MLE  $\bar{x}$

**Example: London bombings during WWII**

- Data below are the number of flying bomb hits on London during World War II in a 36 km<sup>2</sup> area of South London
- Area was partitioned into 0.25 km<sup>2</sup> grid squares and number of bombs falling in each grid was counted

Hits, $x$	0	1	2	3	4	7
Number of areas, $n$	229	211	93	35	7	1

Total hits,  $\sum_i n_i x_i = 537$

Total number of areas,  $\sum_i n_i = 576$

- If the hits are random, a Poisson distribution with constant hit rate  $\theta$  should fit the data
- Can think of  $n = 576$  observations from a Poisson distribution, with  $\bar{x} = 537/576 = 0.93$

The 'invariant' Jeffreys prior (see later) for the mean  $\theta$  of a Poisson distribution is  $p(\theta) \propto 1/\sqrt{\theta}$ , which is equivalent to an (improper) Gamma(0.5,0) distribution. Therefore

$$\begin{aligned} p(\theta|\mathbf{y}) &= \text{Gamma}(a + n\bar{x}, b + n) = \text{Gamma}(537.5, 576) \\ \mathbb{E}(\theta|\mathbf{y}) &= \frac{537.5}{576} = 0.933; \quad \text{Var}(\theta|\mathbf{y}) = \frac{537.5}{576^2} = 0.0016 \end{aligned}$$

Note that these are almost exactly the same as the MLE and the square of the SE(MLE)

## Summary

For all these examples, we see that

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of the prior s.d. and the s.e.(MLE)

*'A Bayesian is one who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule'* (Senn, 1997)

As  $n \rightarrow \infty$ ,

- the posterior mean  $\rightarrow$  the MLE
- the posterior s.d.  $\rightarrow$  the s.e.(MLE)
- the posterior does not depend on the prior.

These observations are generally true, when the MLE exists and is unique

## Choosing prior distributions

When the posterior is in the same family as the prior then we have what is known as *conjugacy*. This has the advantage that prior parameters can usually be interpreted as a *prior sample*. Examples include:

Likelihood	Parameter	Prior	Posterior
Normal	mean	Normal	Normal
Normal	precision	Gamma	Gamma
Binomial	success prob.	Beta	Beta
Poisson	rate or mean	Gamma	Gamma

- Conjugate prior distributions are mathematically convenient, but do not exist for all likelihoods, and can be restrictive
- Computations for non-conjugate priors are harder, but possible using MCMC (see next lecture)

## Calling WinBUGS from other software

- Scripts enable WinBUGS 1.4 to be called from other software
- Interfaces developed for R, Splus, SAS, Matlab
- See [www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml](http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml)
- Andrew Gelman's `bugs` function for R is most developed - reads in data, writes script, monitors output etc. Now packaged as `R2WinBUGS`.
- OpenBUGS site <http://mathstat.helsinki.fi/openbugs/> provides an open source version, including `BRugs` package which works from within R

## **Further reading**

Berry (1996) (Introductory text on Bayesian methods, with medical slant)

Lee (2004) (Good intro to Bayesian inference; more mathematical than Berry; 3rd edition contains WinBUGS examples)

Bernardo and Smith (1994) (Advanced text on Bayesian theory)

**Lecture 3.**  
**Introduction to MCMC**

## Why is computation important?

- Bayesian inference centres around the posterior distribution

$$p(\boldsymbol{\theta}|x) \propto p(x|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

where  $\boldsymbol{\theta}$  is typically a large vector of parameters  $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_k\}$

- $p(x|\boldsymbol{\theta})$  and  $p(\boldsymbol{\theta})$  will often be available in closed form, but  $p(\boldsymbol{\theta}|x)$  is usually not analytically tractable, and we want to
  - obtain marginal posterior  $p(\theta_i|x) = \int \int \dots \int p(\boldsymbol{\theta}|x) d\boldsymbol{\theta}_{(-i)}$  where  $\boldsymbol{\theta}_{(-i)}$  denotes the vector of  $\theta$ 's excluding  $\theta_i$
  - calculate properties of  $p(\theta_i|x)$ , such as mean ( $= \int \theta_i p(\theta_i|x) d\theta_i$ ), tail areas ( $= \int_T^\infty p(\theta_i|x) d\theta_i$ ) etc.

→ numerical integration becomes vital

## Monte Carlo integration

We have already seen that Monte Carlo methods can be used to simulate values from prior distributions and from **closed form** posterior distributions

If we had algorithms for sampling from arbitrary (typically high-dimensional) posterior distributions, we could use Monte Carlo methods for Bayesian estimation:

- Suppose we can draw samples from the joint posterior distribution for  $\theta$ , *i.e.*

$$(\theta_1^{(1)}, \dots, \theta_k^{(1)}), (\theta_1^{(2)}, \dots, \theta_k^{(2)}), \dots, (\theta_1^{(N)}, \dots, \theta_k^{(N)}) \sim p(\theta|x)$$

- Then

- $\theta_1^{(1)}, \dots, \theta_1^{(N)}$  are a sample from the marginal posterior  $p(\theta_1|x)$

- $E(g(\theta_1)) = \int g(\theta_1)p(\theta_1|x)d\theta_1 \approx \frac{1}{N} \sum_{i=1}^N g(\theta_1^{(i)})$

→ this is Monte Carlo integration

→ theorems exist which prove convergence in limit as  $N \rightarrow \infty$  even if the sample is dependent (crucial to the success of MCMC)

## How do we sample from the posterior?

- We want samples from joint posterior distribution  $p(\boldsymbol{\theta}|x)$
- *Independent* sampling from  $p(\boldsymbol{\theta}|x)$  may be difficult
- **BUT** *dependent* sampling from a *Markov chain* with  $p(\boldsymbol{\theta}|x)$  as its stationary (equilibrium) distribution is easier
- A sequence of random variables  $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots$  forms a Markov chain if  $\theta^{(i+1)} \sim p(\theta|\theta^{(i)})$  i.e. conditional on the value of  $\theta^{(i)}$ ,  $\theta^{(i+1)}$  is independent of  $\theta^{(i-1)}, \dots, \theta^{(0)}$
- Several standard ‘recipes’ available for designing Markov chains with required stationary distribution  $p(\boldsymbol{\theta}|x)$ 
  - Metropolis *et al.* (1953); generalised by Hastings (1970)
  - **Gibbs Sampling** (see Geman and Geman (1984), Gelfand and Smith (1990), Casella and George (1992)) is a special case of the Metropolis-Hastings algorithm which generates a Markov chain by sampling from **full conditional distributions**
  - See Gilks, Richardson and Spiegelhalter (1996) for a full introduction and many worked examples.

## **Gibbs sampling**

Let our vector of unknowns  $\theta$  consist of  $k$  sub-components  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$

1) Choose starting values  $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$

2) Sample  $\theta_1^{(1)}$  from  $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, x)$

Sample  $\theta_2^{(1)}$  from  $p(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, x)$

.....

Sample  $\theta_k^{(1)}$  from  $p(\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)}, x)$

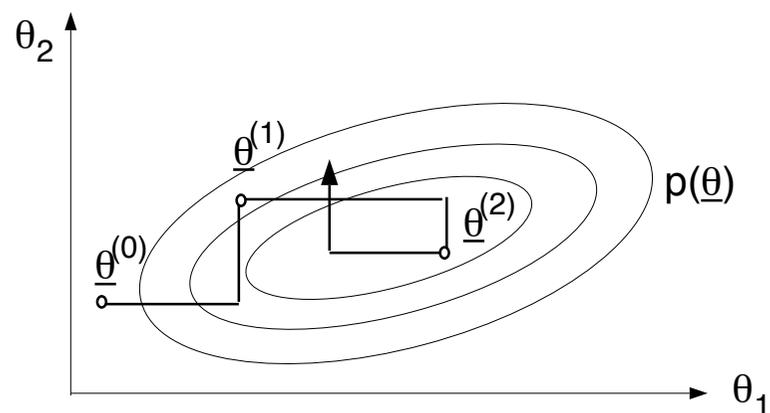
3) Repeat step 2 many 1000s of times

– eventually obtain sample from  $p(\theta | x)$

The conditional distributions are called ‘full conditionals’ as they condition on all other parameters

## Gibbs sampling ctd.

Example with  $k = 2$



- Sample  $\theta_1^{(1)}$  from  $p(\theta_1|\theta_2^{(0)}, x)$
- Sample  $\theta_2^{(1)}$  from  $p(\theta_2|\theta_1^{(1)}, x)$
- Sample  $\theta_1^{(2)}$  from  $p(\theta_1|\theta_2^{(1)}, x)$
- .....

$\theta^{(n)}$  forms a Markov chain with (*eventually*) a stationary distribution  $p(\theta|x)$ .

## Using MCMC methods

There are two main issues to consider

- Convergence (how quickly does the distribution of  $\theta^{(t)}$  approach  $p(\theta|x)$ ?)
- Efficiency (how well are functionals of  $p(\theta|x)$  estimated from  $\{\theta^{(t)}\}$ ?)

## Checking convergence

This is the users responsibility!

- Note: Convergence is to target **distribution** (the required posterior), not to a single value.
- Once convergence reached, samples should look like a random scatter about a stable mean value

## **Convergence diagnosis**

- How do we know we have reached convergence?
- *i.e.* How do we know the number of 'burn-in' iterations?
- Many 'convergence diagnostics' exist, but none foolproof
- CODA and BOA software contain large number of diagnostics

### *Gelman-Rubin-Brooks diagnostic*

- A number of runs
- Widely differing starting points
- Convergence assessed by quantifying whether sequences are much further apart than expected based on their internal variability
- Diagnostic uses components of variance of the multiple sequences

**Example: A dose-response model**

Consider the following response rates for different doses of a drug

dose $x_i$	No. subjects $n_i$	No. responses $r_i$
1.69	59	6
1.72	60	13
1.75	62	18
1.78	56	28
1.81	63	52
1.83	59	53
1.86	62	61
1.88	60	60

Fit a logistic curve with 'centred' covariate  $(x_i - \bar{x})$ :

$$\begin{aligned}
 r_i &\sim \text{Bin}(p_i, n_i) \\
 \text{logit } p_i &= \alpha + \beta(x_i - \bar{x}) \\
 \alpha &\sim \text{N}(0, 10000) \\
 \beta &\sim \text{N}(0, 10000)
 \end{aligned}$$

## Checking convergence with multiple runs

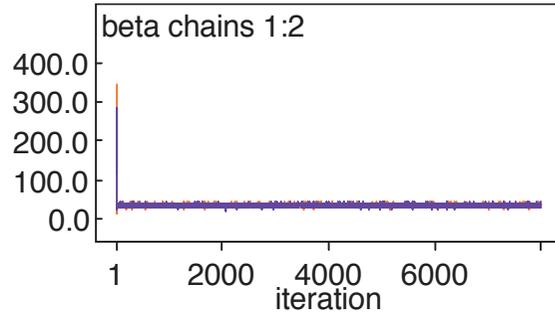
- Set up multiple initial value lists, *e.g.*  
`list(alpha=-100, beta=100)`  
`list(alpha=100, beta=-100)`
- Before clicking *compile*, set *num of chains* to 2
- Load both sets of initial values
- Monitor from the start of sampling
- Assess how much burn-in needed using the *bgr* statistic

### Using the *bgr* statistic

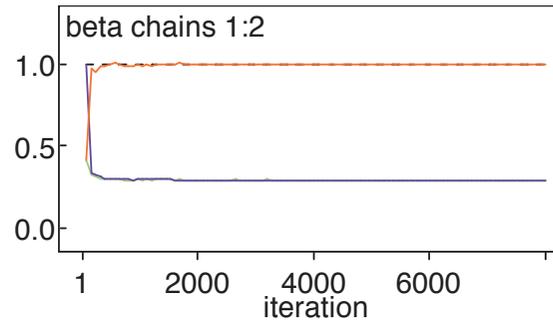
- *Green*: width of 80% intervals of pooled chains: should be stable
- *Blue*: average width of 80% intervals for chains: should be stable
- *Red*: ratio of pooled/within: should be near 1
- Double-click on plot, then *ctrl* + right click gives statistics

## Output for 'centred' analysis

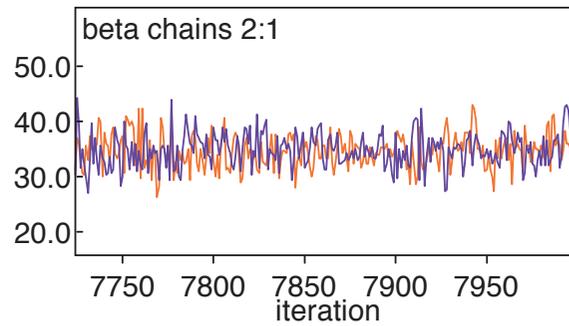
history



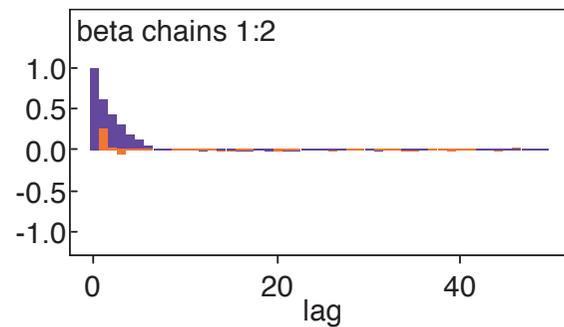
bgr diagnostic



trace



autocorrelation



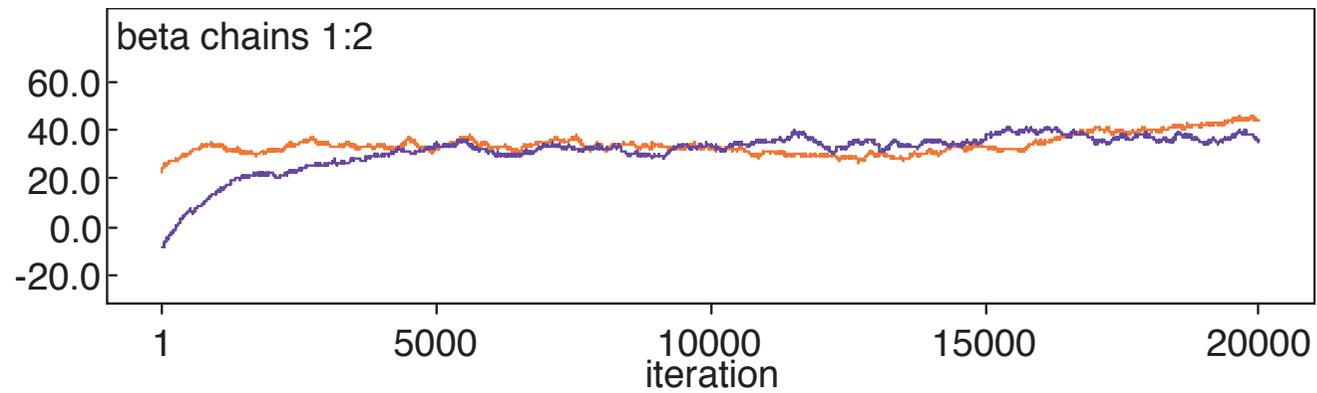
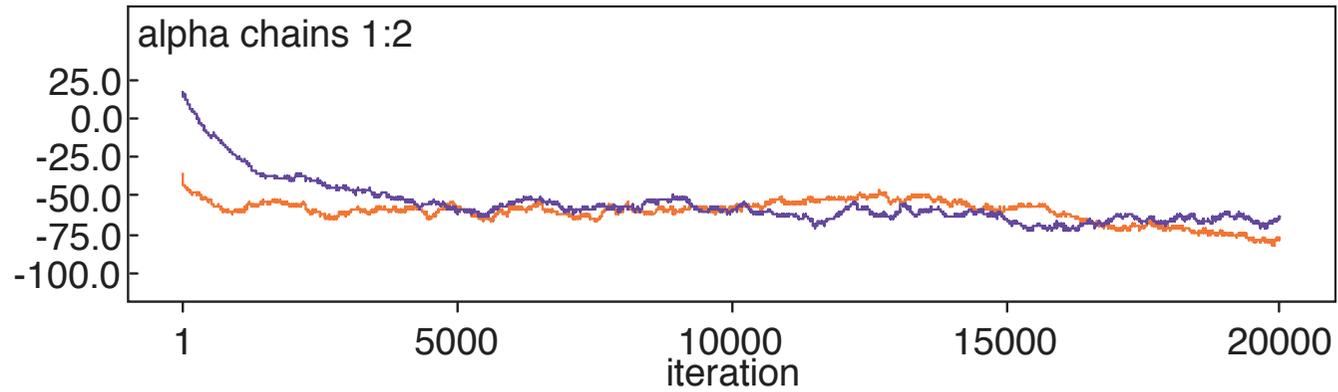
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	0.7489	0.139	0.00138	0.4816	0.7468	1.026	1001	14000
beta	34.6	2.929	0.02639	29.11	34.53	40.51	1001	14000

## **Problems with convergence**

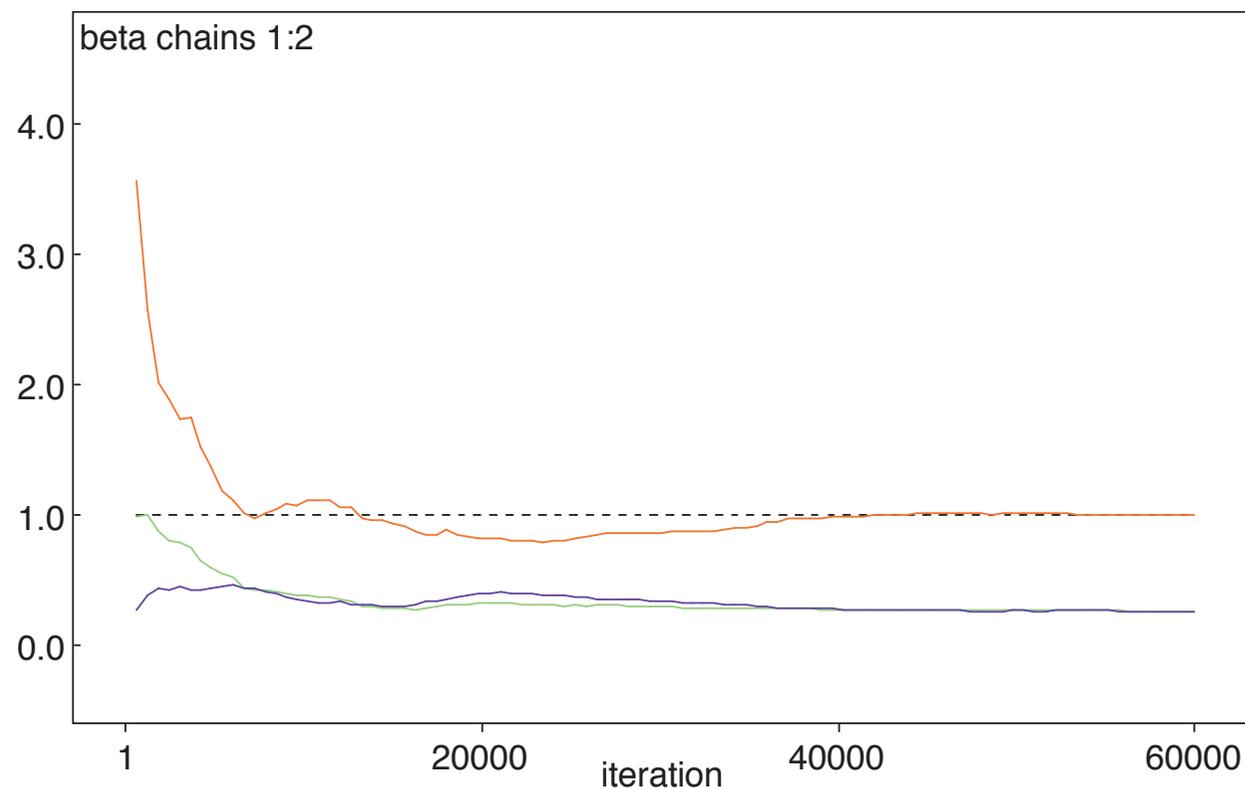
Fit a logistic curve with 'un-centred' covariate  $x$ :

$$\begin{aligned} r_i &\sim \text{Bin}(p_i, n_i) \\ \text{logit } p_i &= \alpha + \beta x_i \\ \alpha &\sim \text{N}(0, 10000) \\ \beta &\sim \text{N}(0, 10000) \end{aligned}$$

## History plots for 'un-centred' analysis



## bgr output for 'un-centred' analysis

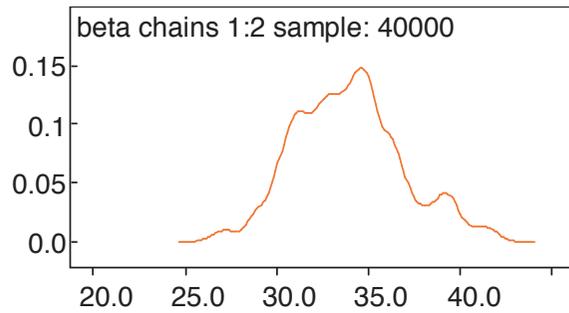


Drop first 40,000 iterations as burn-in

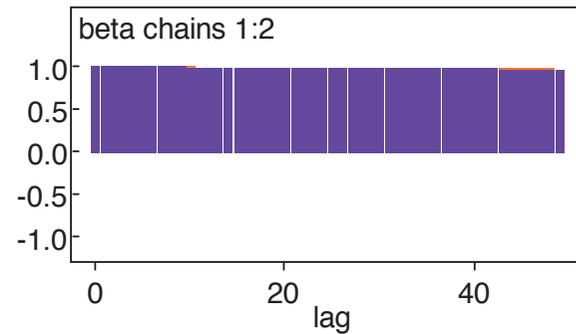
<b>node</b>	<b>mean</b>	<b>sd</b>	<b>MC error</b>	<b>2.5%</b>	<b>median</b>	<b>97.5%</b>	<b>start</b>	<b>sample</b>
beta	33.97	2.955	0.1734	28.7	33.89	40.3	40001	40000

## Output for 'un-centred' analysis

posterior density

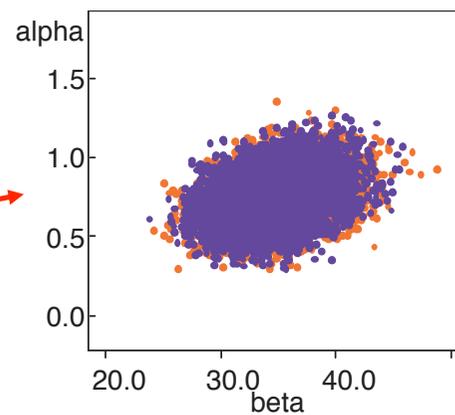


autocorrelation

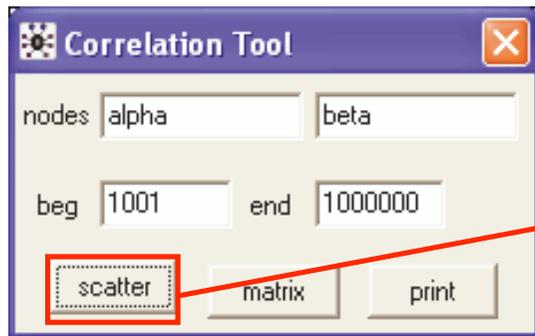
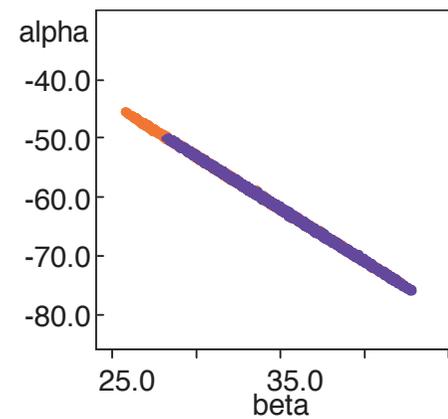


## bivariate posteriors

centred



un-centred



## How many iterations after convergence?

- After convergence, further iterations are needed to obtain samples for posterior inference.
- More iterations = more accurate posterior estimates.
- Efficiency of sample mean of  $\theta$  as estimate of theoretical posterior expectation  $E(\theta)$  usually assessed by calculating Monte Carlo standard error (MC error)
- MC error = standard error of posterior sample mean as estimate of theoretical expectation for given parameter
- MC error depends on
  - true variance of posterior distribution
  - posterior sample size (number of MCMC iterations)
  - autocorrelation in MCMC sample
- Rule of thumb: want MC error  $< 1 - 5\%$  of posterior SD

## **Inference using posterior samples from MCMC runs**

A powerful feature of the Bayesian approach is that all inference is based on the joint posterior distribution

⇒ can address wide range of substantive questions by appropriate summaries of the posterior

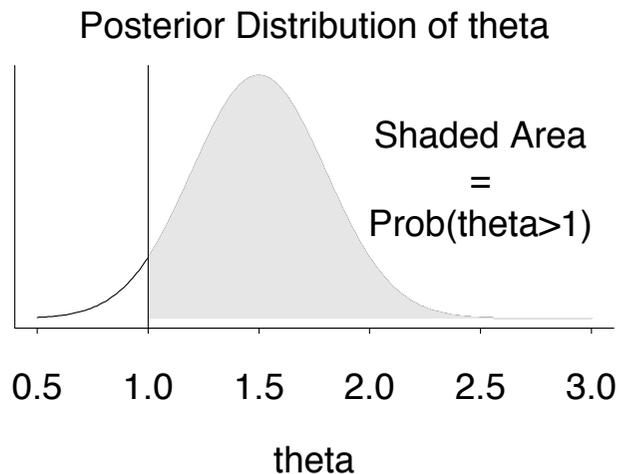
- Typically report either mean or median of the posterior samples for each parameter of interest as a point estimate
- 2.5% and 97.5% percentiles of the posterior samples for each parameter give a 95% posterior credible interval (interval within which the parameter lies with probability 0.95)

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta	34.60	2.929	0.0239	29.11	34.53	40.51	1001	14000

So point estimate of beta would be 34.60, with 95% credible interval (29.11, 40.51)

## Probability statements about parameters

- Classical inference cannot provide probability statements about parameters (e.g. p-value is not  $\Pr(H_0 \text{ true})$ , but probability of observing data as or more extreme than we obtained, given that  $H_0$  is true)
- In Bayesian inference, it is simple to calculate e.g.  $\Pr(\theta > 1)$ :
  - = Area under posterior distribution curve to the right of 1
  - = Proportion of values in posterior sample of  $\theta$  which are  $> 1$



- In WinBUGS use the step function:  
`p.theta <- step(theta - 1)`
- For discrete parameters, may also be interested in  $\Pr(\delta = \delta_0)$ :  
`p.delta <- equals(delta, delta0)`
- Posterior means of `p.theta` and `p.delta` give the required probabilities

## Complex functions of parameters

- Classical inference about a function of the parameters  $g(\theta)$  requires construction of a specific estimator of  $g(\theta)$ . Obtaining appropriate error can be difficult.
- Easy using MCMC: just calculate required function  $g(\theta)$  as a logical node at each iteration and summarise posterior samples of  $g(\theta)$

In dose-response example, suppose we want to estimate the  $ED_{95}$ : that is the dose that will provide 95% of maximum efficacy.

$$\begin{aligned}\text{logit } 0.95 &= \alpha + \beta(ED_{95} - \bar{x}) \\ ED_{95} &= (\text{logit } 0.95 - \alpha) / \beta + \bar{x}\end{aligned}$$

Simply add into model

```
ED95 <- (logit(0.95) - alpha)/beta + mean(x[])
```

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
ED95	1.857	0.007716	8.514E-5	1.843	1.857	1.874	1001	10000

## How to rank if you must

- Recent trend in UK towards ranking 'institutional' performance e.g. schools, hospitals
- Might also want to rank treatments, answer 'which is the best' etc
- Rank of a point estimate is a highly unreliable summary statistic

⇒ Would like measure of uncertainty about rank

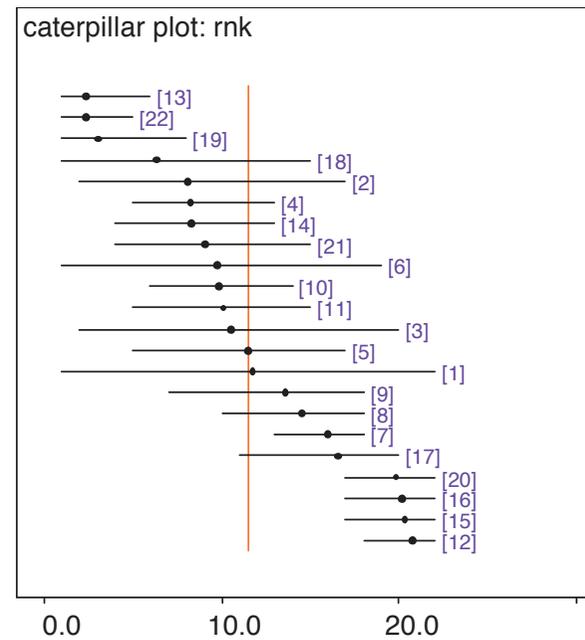
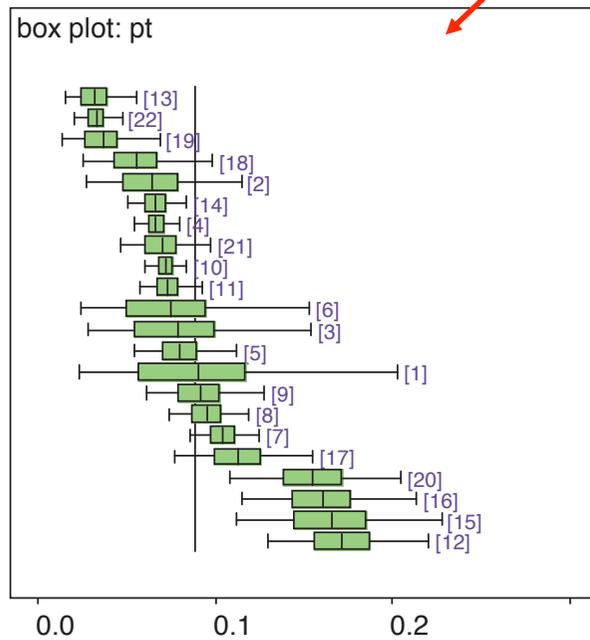
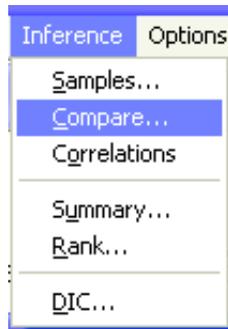
- Bayesian methods provide *posterior interval estimates* for ranks
- WinBUGS contains 'built-in' options for ranks:
  - Rank option of Inference menu monitors the rank of the elements of a specified vector
  - `rank(x[], i)` returns the rank of the  $i^{th}$  element of  $x$
  - `equals(rank(x[],i),1)` =1 if  $i^{th}$  element is ranked lowest, 0 otherwise. Mean is probability that  $i^{th}$  element is 'best' (if counting adverse events)
  - `ranked(x[], i)` returns the value of the  $i^{th}$ -ranked element of  $x$

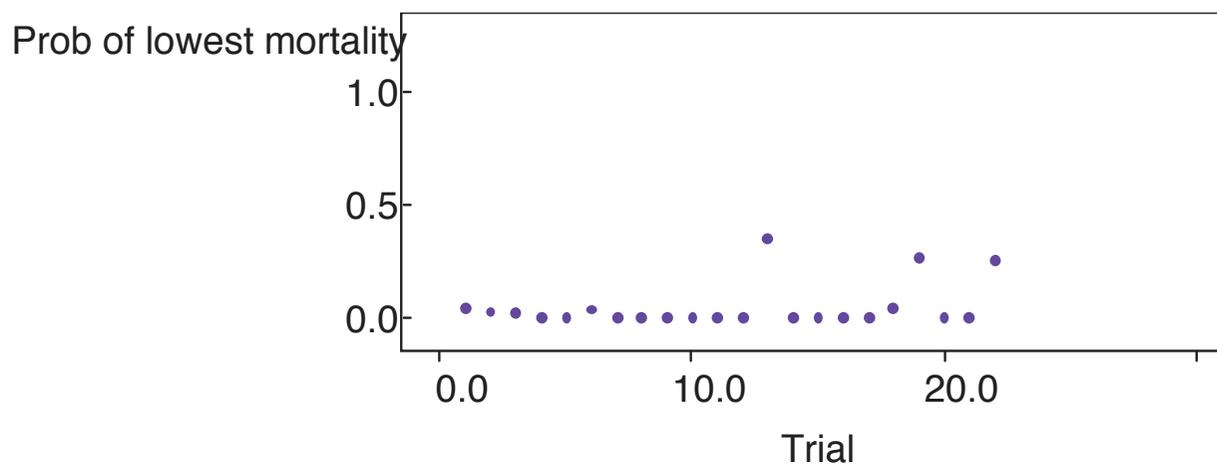
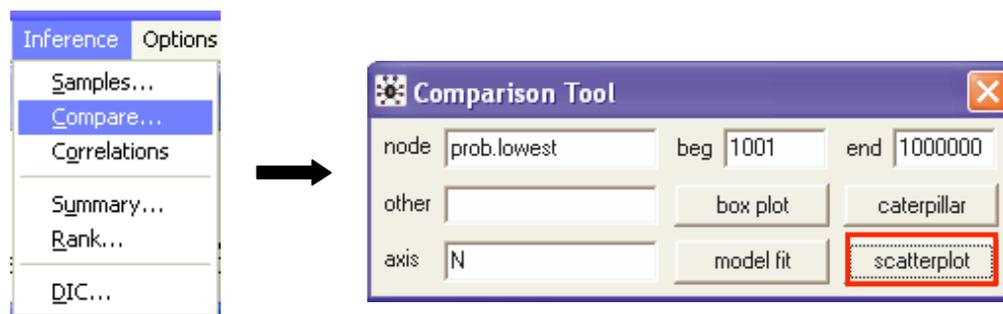
## Example of ranking: 'Blocker' trials

- 22 trials of beta-blockers used in WinBUGS manual to illustrate random-effects meta-analysis.
- Consider just treatment arms: which trial has the lowest mortality rate?
- Assume independent 'Jeffreys'  $\text{beta}[0.5, 0.5]$  prior for each response rate.

```
for( i in 1 : Num) {  
  rt[i] ~ dbin(pt[i],nt[i])  
  pt[i] ~ dbeta(0.5,0.5)           # Jeffreys prior  
  rnk[i] <- rank(pt[], i)         # rank of i'th trial  
  prob.lowest[i] <- equals(rnk[i],1) # prob that i'th trial lowest  
  N[i]<-i                          # used for indexing plot  
}
```

## Mortality rates and ranks





Ranking methods may be useful when

- comparing alternative treatments
- comparing subsets
- comparing response-rates, cost-effectiveness or any summary measure

## **Further reading**

Gelfand and Smith (1990) (key reference to use of Gibbs sampling for Bayesian calculations)

Casella and George (1992) (Explanation of Gibbs sampling)

Brooks (1998) (tutorial paper on MCMC)

Spiegelhalter et al (1996) (Comprehensive coverage of practical aspects of MCMC)

**Lecture 4.**  
**Bayesian linear regression models**

## **Bayesian regression models**

Standard (and non standard) regression models can be easily formulated within a Bayesian framework.

- Specify probability distribution (likelihood) for the data
- Specify form of relationship between response and explanatory variables
- Specify prior distributions for regression coefficients and any other unknown (nuisance) parameters

Some advantages of a Bayesian formulation in regression modelling include:

- Easy to include parameter restrictions and other relevant prior knowledge
- Easily extended to non-linear regression
- Easily 'robustified'
- Easy to make inference about functions of regression parameters and/or predictions
- Easily extended to handle missing data and covariate measurement error

## Linear regression

Consider a simple linear regression with univariate Normal outcome  $y_i$  and a vector of covariates  $x_{1i}, \dots, x_{pi}$ ,  $i = 1, \dots, n$

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki} + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

An equivalent Bayesian formulation would typically specify

$$y_i \sim \text{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki}$$
$$(\beta_0, \beta_1, \dots, \beta_p, \sigma^2) \sim \text{Prior distributions}$$

A typical choice of 'vague' prior distribution (see later for more details) that will give numerical results similar to OLS or MLE is:

$$\beta_k \sim N(0, 100000) \quad k = 0, \dots, p$$
$$1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$$

## Example: New York Crime data

- 23 Precincts in New York City
- Response = THEFT: seasonally adjusted changes in larcenies (thefts) from a 27-week base period in 1966 to a 58-week experimental period in 1966-1967
- Predictors = MAN: % change in police manpower; DIST: district indicator (1 Downtown, 2 Mid-town, 3 Up-town)
- Model specification:

$$\begin{aligned} \text{THEFT}_i &\sim \text{Normal}(\mu_i, \sigma^2) \quad i = 1, \dots, 21 \\ \mu_i &= \alpha + \beta \times \text{MAN}_i + \langle \text{effect of DIST} \rangle \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \\ \alpha &\sim \text{N}(0, 100000) \\ \beta &\sim \text{N}(0, 100000) \\ \text{Prior} &\text{ on coefficients for DIST effect} \end{aligned}$$

## Specifying categorical covariates using the BUGS language

$\text{DIST}_i$  is a 3-level categorical explanatory variable

Two alternative ways of specifying model in BUGS language

1. Create usual 'design matrix' in data file:

```
MAN[] THEFT[] DIST2[] DIST3[]
-15.76  3.19      0      0 # district 1
  0.98 -3.45      0      0
  3.71  0.04      0      0
.....
-9.56   3.68      0      0
-2.06   8.63      1      0 # district 2
-0.76  10.82      1      0
-6.30  -0.50      1      0
.....
-2.82  -2.02      1      0
-16.19  0.94      0      1 # district 3
-11.00  4.42      0      1
.....
-10.77  1.58      0      1
END
```

BUGS model code is then

```
for (i in 1:N) {  
  THEFT[i] ~ dnorm(mu[i], tau)  
  mu[i] <- alpha + beta*MAN[i] + delta2*DIST2[i] + delta3*DIST3[i]  
}  
alpha ~ dnorm(0, 0.00001)  
beta ~ dnorm(0, 0.00001)  
delta2 ~ dnorm(0, 0.00001)  
delta3 ~ dnorm(0, 0.00001)  
tau ~ dgamma(0.001, 0.001)  
sigma2 <- 1/tau
```

Note: BUGS parameterises normal in terms of mean and **precision** (1/variance)!!

Initial values file would be something like

```
list(alpha = 1, beta = -2, delta2 = -2, delta3 = 4, tau = 2)
```

2. Alternatively, input explanatory variable as single vector coded by its level:

MAN []	THEFT []	DIST []
-15.76	3.19	1
0.98	-3.45	1
3.71	0.04	1
.....		
-9.56	3.68	1
-2.06	8.63	2
-0.76	10.82	2
-6.30	-0.50	2
.....		
-2.82	-2.02	2
-16.19	0.94	3
-11.00	4.42	3
.....		
-10.77	1.58	3

END

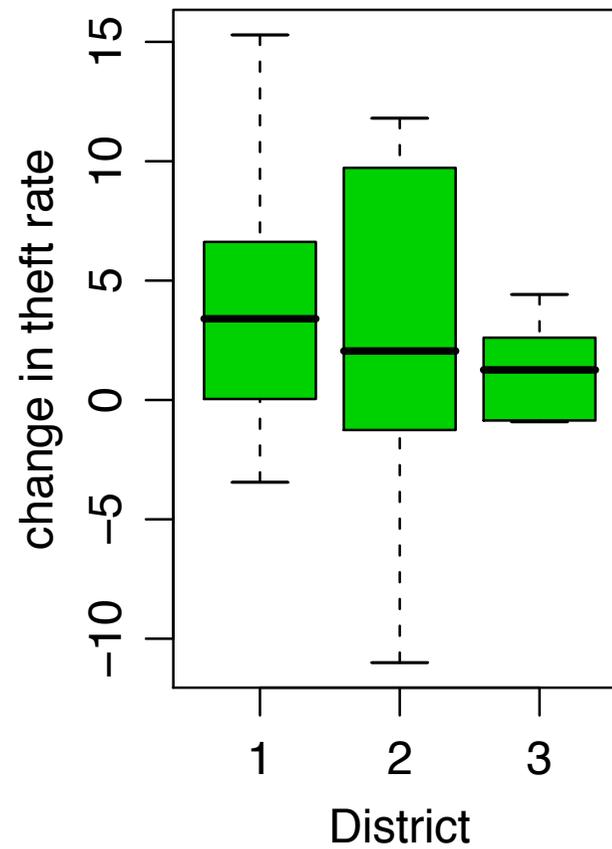
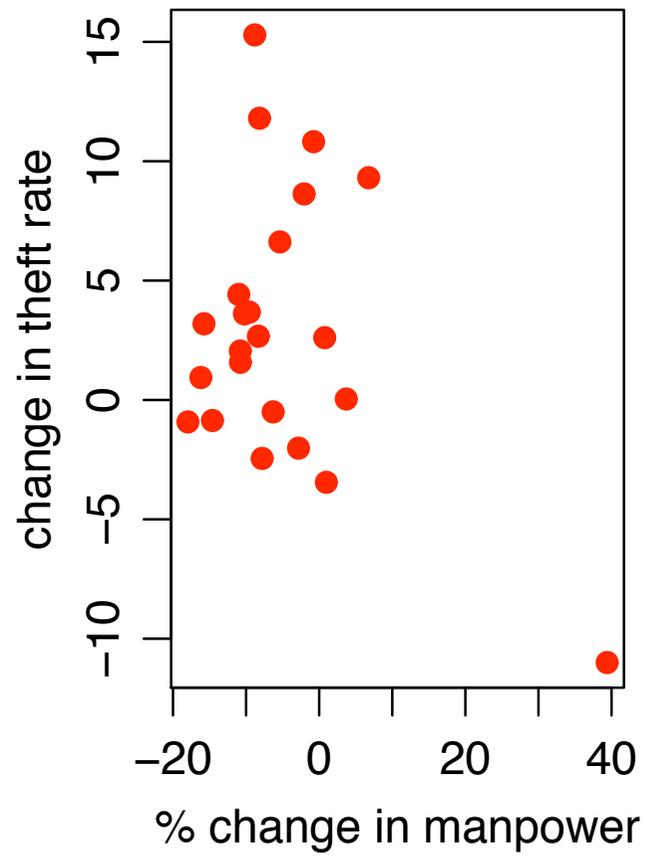
Then use 'double indexing' feature of BUGS language

```
for (i in 1:23) {  
  THEFT[i] ~ dnorm(mu[i], tau)  
  mu[i] <- alpha + beta*MAN[i] + delta[DIST[i]]  
}  
alpha ~ dnorm(0, 0.00001)  
beta ~ dnorm(0, 0.00001)  
delta[1] <- 0 # set coefficient for reference category to zero  
delta[2] ~ dnorm(0, 0.00001)  
delta[3] ~ dnorm(0, 0.00001)  
tau ~ dgamma(0.001, 0.001)  
sigma2 <- 1/tau
```

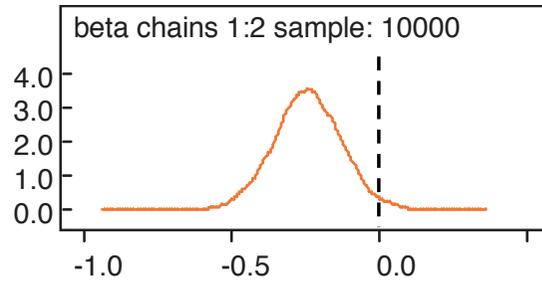
In initial values file, need to specify initial values for delta[2] and delta[3] but not delta[1]. Use following syntax:

```
list(alpha = 1, beta = -2, delta = c(NA, -2, 4), tau = 2)
```

Raw data



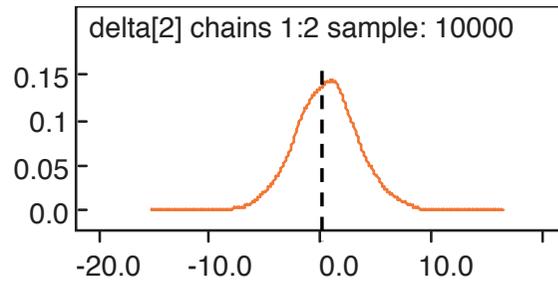
### Change in theft rate per 1% increase in police manpower



Posterior mean -0.24

95% interval (-0.47, -0.01)

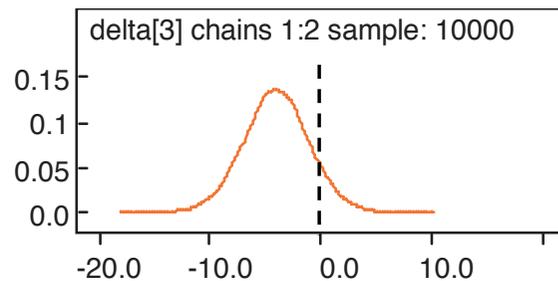
### Change in theft rate in Midtown relative to Downtown



Posterior mean 0.6

95% interval (-5.1, 6.6)

### Change in theft rate in Uptown relative to Downtown

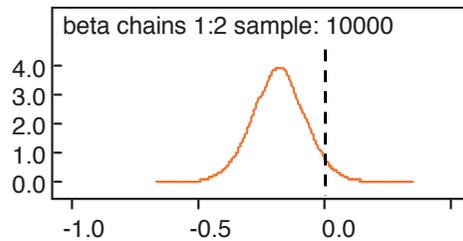


Posterior mean -4.0

95% interval (-10.0, 2.1)

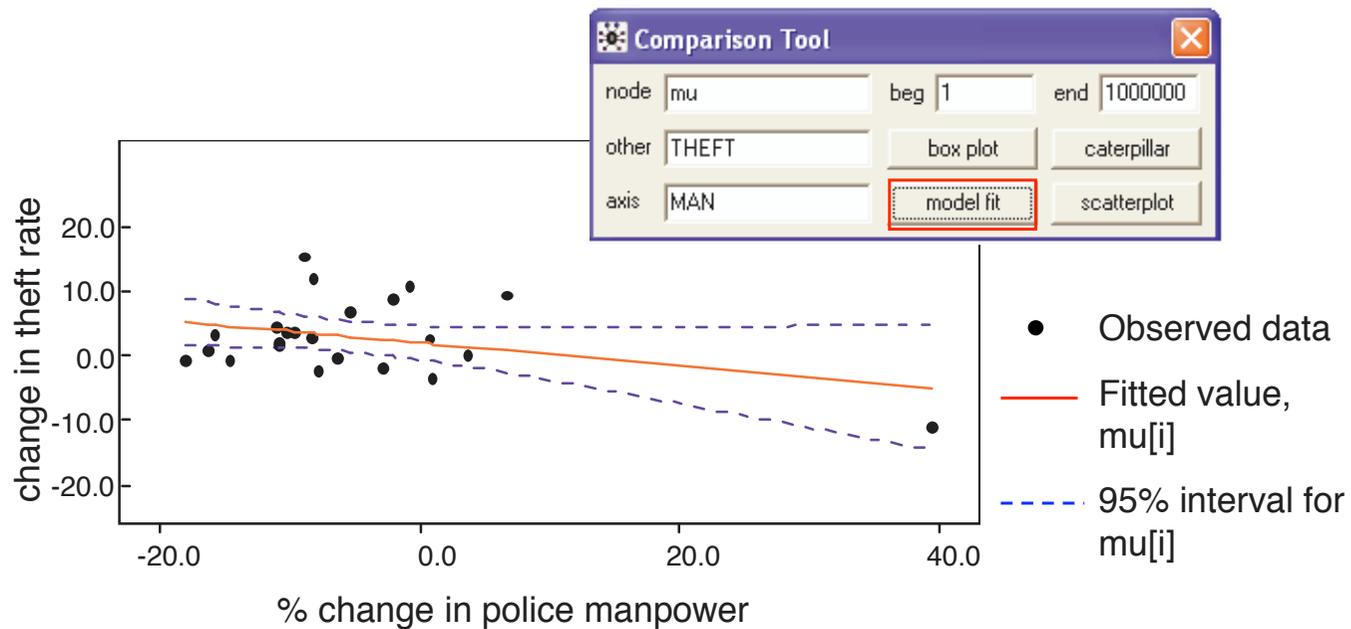
- 95% intervals for DIST effect both include zero  
→ drop DIST from model (see later for Bayesian model comparison criteria)

### Change in theft rate per 1% increase in police manpower



Posterior mean -0.18

95% interval (-0.39, 0.04)

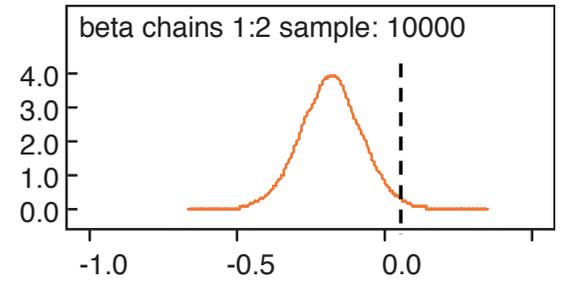
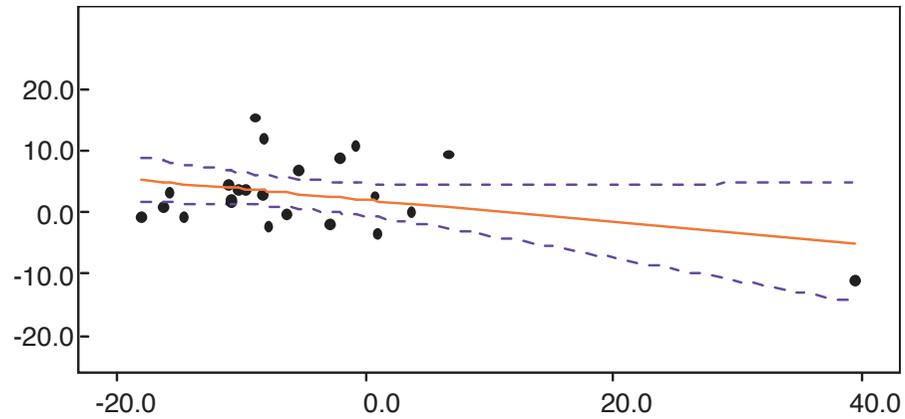


- Influential point corresponds to 20<sup>th</sup> Precinct
- During 2nd period, manpower assigned to this Precinct was experimentally increased by about 40%
- No experimental increases in any other Precinct

→ Robustify model assuming t-distributed errors

```
for (i in 1:23) {  
  THEFT[i] ~ dt(mu[i], tau, 4)  # robust likelihood (t on 4 df)  
  mu[i] <- alpha + beta*MAN[i]  
}  
alpha ~ dnorm(0, 0.00001)  
beta ~ dnorm(0, 0.00001)  
tau ~ dgamma(0.001, 0.001)  
sigma2 <- 1/tau  
  
dummy <- DIST[1]  # ensures all variables in data file appear in model code
```

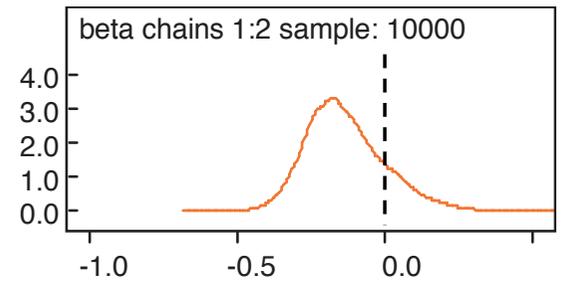
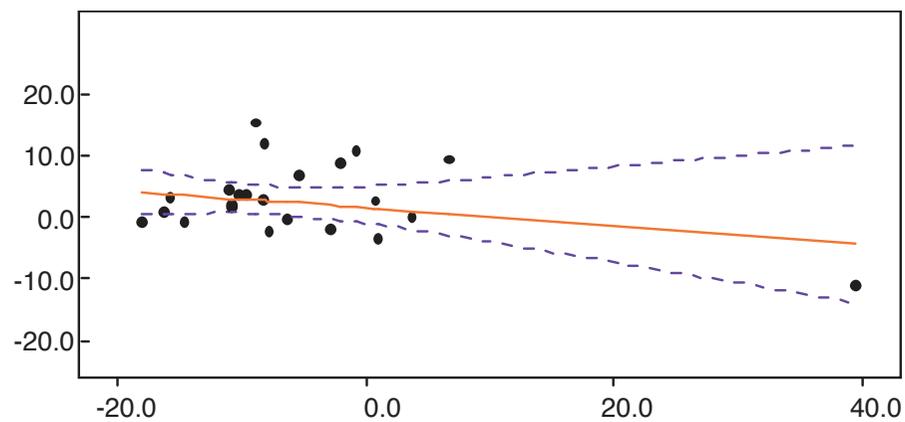
### Model with Normal errors



Posterior mean -0.18

95% interval (-0.39, 0.04)

### Model with Student t errors

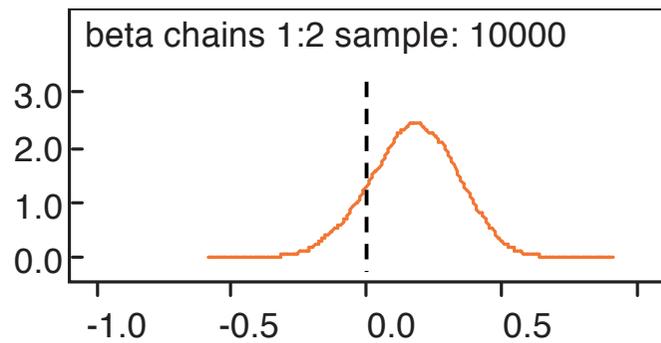
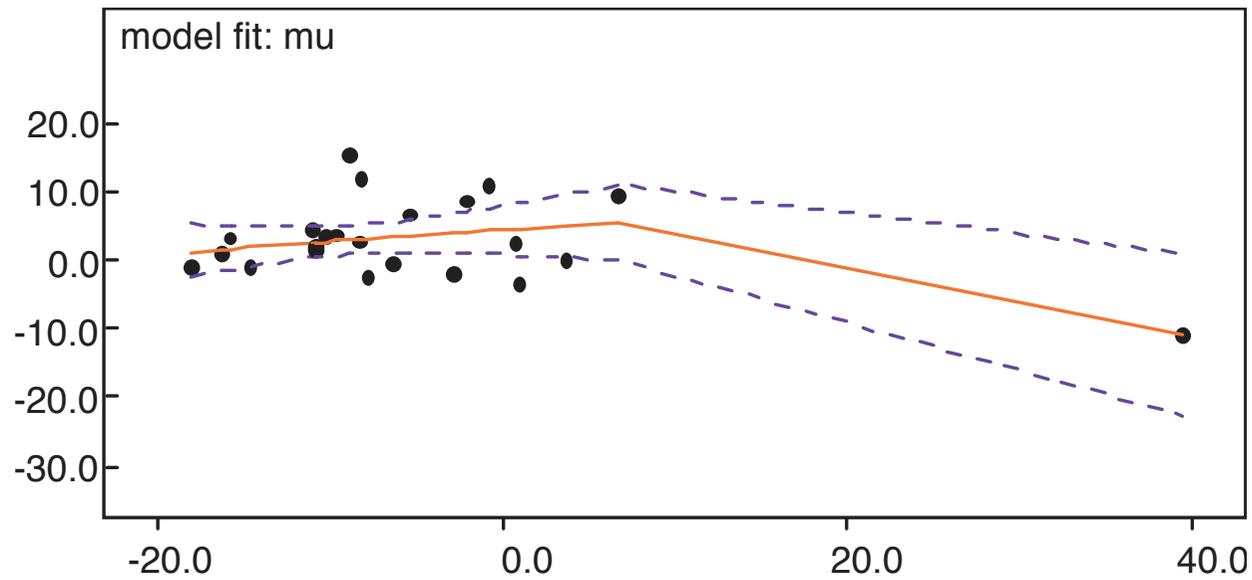


Posterior mean -0.13

95% interval (-0.36, 0.18)

- Precinct 20 still quite influential
- Add additional covariate corresponding to a binary indicator for Precinct 20
  - Equivalent to fitting separate (saturated) model to Precinct 20

```
for(i in 1:23) {  
  THEFT[i] ~ dt(mu[i], tau, 4) # robust likelihood (t on 4 df)  
  mu[i] <- alpha + beta*MAN[i] + delta*PREC20[i] # separate term for precinct 20  
}  
alpha ~ dnorm(0, 0.000001)  
beta ~ dnorm(0, 0.000001)  
delta ~ dnorm(0, 0.000001)  
tau ~ dgamma(0.001, 0.001)  
sigma2 <- 1/tau # residual error variance  
  
dummy <- DIST[1] # ensures all variables in data file appear in model code  
  
# Create indicator variable for precinct 20  
# (alternatively, could add this variable to data file)  
for(i in 1:13) { PREC20[i] <- 0 }  
PREC20[14] <- 1  
for(i in 15:23) { PREC20[i] <- 0 }
```



Posterior mean 0.18  
95% interval (-0.16, 0.49)

## Specifying prior distributions

Why did we choose a  $N(0, 100000)$  prior for each regression coefficient and a  $\text{Gamma}(0.001, 0.001)$  prior for the inverse of the error variance?

Choice of prior is, in principle, subjective

- it might be elicited from experts (see Spiegelhalter et al (2004), sections 5.2, 5.3)
- it might be more convincing to be based on historical data, *e.g.* a previous study
  - assumed relevance is still a subjective judgement (see Spiegelhalter et al (2004), section 5.4)
- there has been a long and complex search for various 'non-informative', 'reference' or 'objective' priors (Kass and Wasserman, 1996)

## **‘Non-informative’ priors**

- Better to refer to as ‘vague’, ‘diffuse’ or ‘minimally informative’ priors
- Prior is vague with respect to the likelihood
  - prior mass is diffusely spread over range of parameter values that are plausible, i.e. supported by the data (likelihood)

## Uniform priors (Bayes 1763; Laplace, 1776)

Set  $p(\theta) \propto 1$

- This is improper ( $\int p(\theta)d\theta \neq 1$ )
- The posterior will still usually be proper
- Inference is based on the likelihood  $p(x | \theta)$
- It is not really objective, since a flat prior  $p(\theta) \propto 1$  on  $\theta$  does not correspond to a flat prior on  $\phi = g(\theta)$ , but to  $p(\phi) \propto \left| \frac{d\theta}{d\phi} \right|$  where  $\left| \frac{d\theta}{d\phi} \right|$  is the Jacobian
  - Note: Jacobian ensures area under curve (probability) in a specified interval  $(\theta_1, \theta_2)$  is preserved under the transformation  $\rightarrow$  same area in interval  $(\phi_1 = g(\theta_1), \phi_2 = g(\theta_2))$

Proper approximations to  $\text{Uniform}(-\infty, \infty)$  prior:

- $p(\theta) = \text{Uniform}(a, b)$  where  $a$  and  $b$  specify an appropriately wide range, e.g.  $\text{Uniform}(-1000, 1000)$
- $p(\theta) = N(0, V)$  where  $V$  is an appropriately large value for the variance, e.g.  $N(0, 100000)$
- Recall that WinBUGS parameterises Normal in terms of mean and precision, so vague normal prior will be, e.g.  $\text{theta} \sim \text{dnorm}(0, 0.00001)$

## Jeffreys' invariance priors

Consider 1-to-1 transformation of  $\theta$  :  $\phi = g(\theta)$ , e.g.  $\phi = 1 + \theta^3$

Transformation of variables:  $p(\theta)$  is equivalent to  $p(\phi) = p(\theta = g^{-1}(\phi)) \left| \frac{d\theta}{d\phi} \right|$

Jeffreys proposed defining a non-informative prior for  $\theta$  as  $p(\theta) \propto I(\theta)^{1/2}$  where  $I(\theta)$  is Fisher information for  $\theta$

$$I(\theta) = -\mathbb{E}_{X|\theta} \left[ \frac{\partial^2 \log p(X|\theta)}{\partial \theta^2} \right] = \mathbb{E}_{X|\theta} \left[ \left( \frac{\partial \log p(X|\theta)}{\partial \theta} \right)^2 \right]$$

- Fisher Information measures curvature of log likelihood
- High curvature occurs wherever small changes in parameter values are associated with large changes in the likelihood
  - Jeffreys' prior gives more weight to these parameter values
  - data provide strong information about parameter values in this region
  - ensures data dominate prior everywhere
- Jeffreys' prior is invariant to reparameterisation because

$$I(\phi)^{1/2} = I(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right|$$

## Examples of Jeffreys' priors

- Normal case: unknown mean  $m$ , known variance  $v$   
Sample  $x_1, \dots, x_n$  from  $N(m, v)$

$$\log p(x|m) = -\sum \frac{(x_i - m)^2}{2v} + C \quad \Rightarrow I(m) = n/v$$

So Jeffreys' prior for  $m$  is  $\propto 1$ , i.e. the Uniform distribution

- Normal case: known mean  $m$ , unknown variance  $v$ , with  $s = \sum (x_i - m)^2$

$$\log p(x|v) = -n/2 \log v - \frac{s}{2v} \quad \Rightarrow I(v) = \frac{n}{2v^2}$$

So Jeffreys' prior for  $v$  is  $\propto v^{-1}$

This improper distribution is approximated by a  $\text{Gamma}(\epsilon, \epsilon)$  distribution with  $\epsilon \rightarrow 0$

Note:  $p(v) \propto v^{-1}$  is equivalent to a uniform prior on  $\log v$

## Some recommendations

### Distinguish

- *primary* parameters of interest in which one may want minimal influence of priors
- *secondary* structure used for smoothing *etc.* in which informative priors may be more acceptable

Prior best placed on interpretable parameters

Great caution needed in complex models that an apparently innocuous uniform prior is not introducing substantial information

*'There is no such thing as a 'noninformative' prior. Even improper priors give information: all possible values are equally likely'* (Fisher, 1996)

Location parameters (e.g. means, regression coefficients)

- Uniform prior on a wide range, or a Normal prior with a large variance can be used, e.g.

$$\theta \sim \text{Unif}(-100, 100)$$

$$\text{theta} \sim \text{dunif}(-100, 100)$$

$$\theta \sim \text{Normal}(0, 100000)$$

$$\text{theta} \sim \text{dnorm}(0, 0.00001)$$

Prior will be locally uniform over the region supported by the likelihood

- ! remember that WinBUGS parameterises the Normal in terms of mean and *precision* so a vague Normal prior will have a *small* precision
- ! ‘wide’ range and ‘small’ precision depend on the scale of measurement of  $\theta$

### Scale parameters

- Sample variance  $\sigma^2$ : standard 'reference' (Jeffreys') prior

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \propto \text{Gamma}(0,0)$$
$$p(\log(\sigma)) \propto \text{Uniform}(-\infty, \infty)$$

- Note that Jeffreys' prior on the inverse variance (precision),  $\tau = \sigma^{-2}$  is also

$$p(\tau) \propto \frac{1}{\tau} \propto \text{Gamma}(0, 0)$$

which may be approximated by a 'just proper' prior

$$\tau \sim \text{Gamma}(\epsilon, \epsilon)$$

This is also the conjugate prior and so is widely used as a 'vague' proper prior for the precision of a Normal likelihood

In BUGS language: `tau ~ dgamma(0.001, 0.001)`

or alternatively `tau <- 1/exp(logsigma2); logsigma2 ~ dunif(-100, 100)`

**Sensitivity analysis** plays a crucial role in assessing the impact of particular prior distributions, whether elicited, derived from evidence, or reference, on the conclusions of an analysis.

## **Informative priors**

- An informative prior expresses specific, definite information about a variable
- Example: a prior distribution for the temperature at noon tomorrow
  - A reasonable approach is to make the prior a normal distribution with mean equal to today's noontime temperature, with variance equal to the day-to-day variance of atmospheric temperature
- Posterior from one problem (today's temperature) becomes the prior for another problem (tomorrow's temperature)
- Priors elicited from experts can be used to take account of domain-specific knowledge, judgement, experience
- Priors can also be used to impose constraints on variables (e.g. based on physical or assumed properties) and bound variables to plausible ranges

## **Example: Trade union density**

(Western and Jackman, 1994)

- Example of regression analysis in comparative research
- What explains cross-national variation in union density?
- Union density is defined as the percentage of the work force who belongs to a trade union
- Competing theories:
  - Wallerstein: union density depends on the size of the civilian labour force (LabF)
  - Stephens: union density depends on industrial concentration (IndC)
  - Note: These two predictors correlate at -0.92.

- Data:  $n = 20$  countries with a continuous history of democracy since World War II
- Variables: Union density (Uden), (log) labour force size (LabF), industrial concentration (IndC), left wing government (LeftG), measured in late 1970s
- Fit linear regression model to compare theories

$$\text{Uden}_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = b_0 + b_1(\text{LeftG}_i - \overline{\text{LeftG}}) + b_2(\text{LabF}_i - \overline{\text{LabF}}) + b_3(\text{IndC}_i - \overline{\text{IndC}})$$

Vague priors:

$$1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$$

$$b_0 \sim N(0, 100000)$$

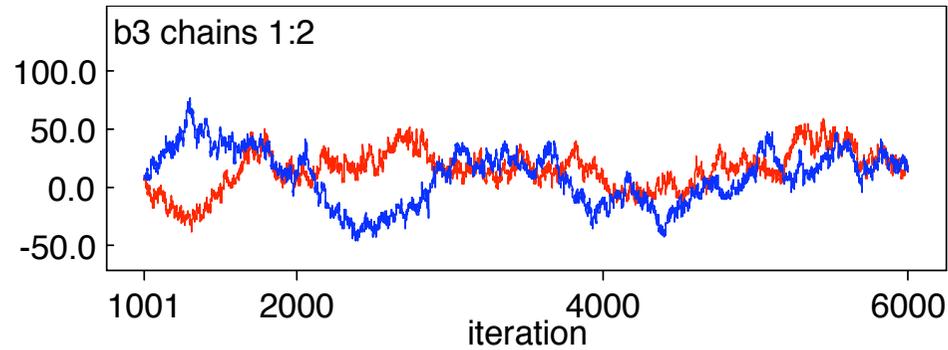
$$b_1 \sim N(0, 100000)$$

$$b_2 \sim N(0, 100000)$$

$$b_3 \sim N(0, 100000)$$

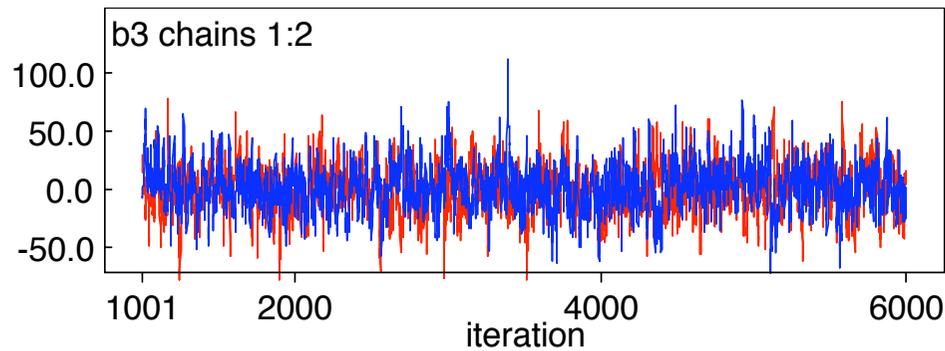
Trace plots, posterior estimates and MC error for regression coefficients

**Without centering covariates**



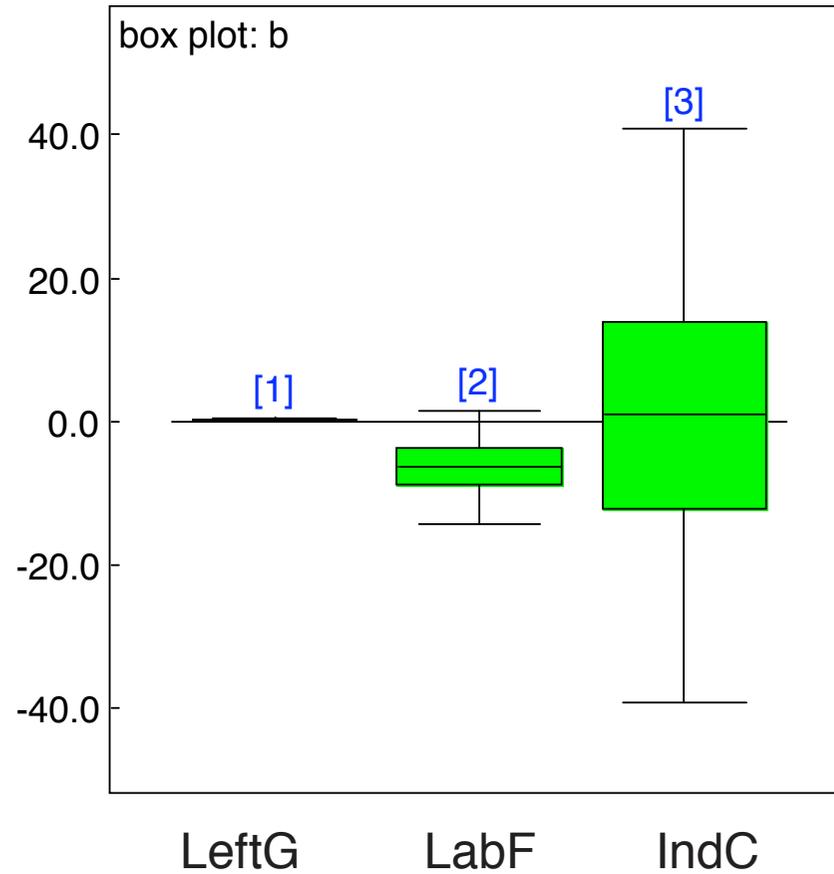
	mean	sd	MC error
b0	61.7	62.8	5.19
b1	0.27	0.08	.002
b2	-4.14	4.18	0.34
b3	12.1	20.6	1.67

**With centered covariates**



	mean	sd	MC error
b0	54.0	2.48	0.02
b1	0.27	0.08	.001
b2	-6.33	3.96	0.13
b3	0.98	20.2	0.67

## Posterior distribution of regression coefficients



*Motivation for Bayesian approach with informative priors*

- Because of small sample size and multicollinear variables, not able to adjudicate between theories
- Data tend to favour Wallerstein (union density depends on labour force size), but neither coefficient estimated very precisely
- Other historical data are available that could provide further relevant information
- Incorporation of prior information provides additional structure to the data, which helps to uniquely identify the two coefficients

Wallerstein informative prior

- Believes in negative labour force effect
- Comparison of Sweden and Norway in 1950:
  - doubling of labour force corresponds to 3.5-4% drop in union density
  - on log scale, labour force effect size  $\approx -3.5/\log(2) \approx -5$
- Confidence in direction of effect represented by prior SD giving 95% interval that excludes 0

$$b_2 \sim N(-5, 2.5^2)$$

- Vague prior assumed for IndC effect,  $b_3 \sim N(0, 100000)$

Stephens informative prior

- Believes in positive industrial concentration effect
- Decline in industrial concentration in UK in 1980s:
  - drop of 0.3 in industrial concentration corresponds to about 3% drop in union density
  - industrial concentration effect size  $\approx 3/0.3 = 10$
- Confidence in direction of effect represented by prior SD giving 95% interval that excludes 0

$$b_3 \sim N(10, 5^2)$$

- Vague prior assumed for IndC effect,  $b_3 \sim N(0, 100000)$

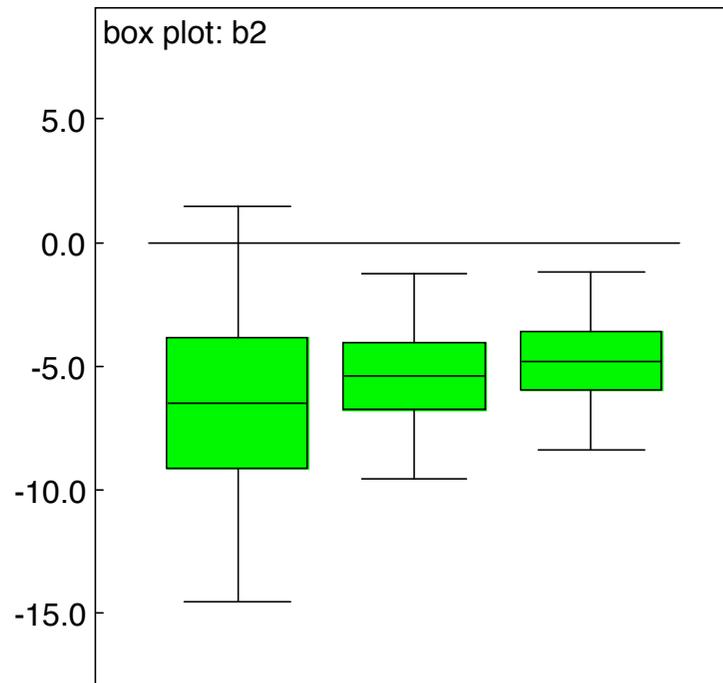
Both Wallerstein and Stephens priors

- Both believe left-wing governments assist union growth
- Assuming 1 year of left-wing government increases union density by about 1% translates to effect size of 0.3
- Confidence in direction of effect represented by prior SD giving 95% interval that excludes 0

$$b_1 \sim N(0.3, 0.15^2)$$

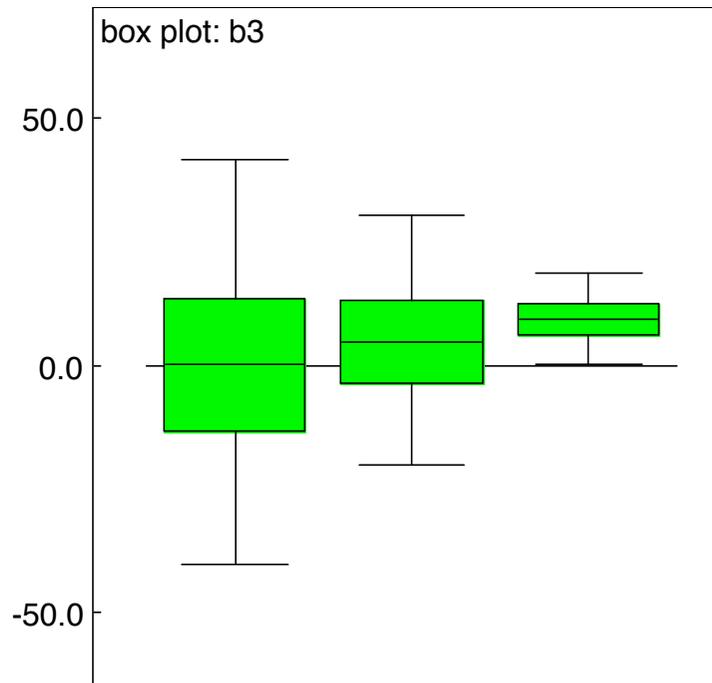
- Vague prior  $b_0 \sim N(0, 100000)$  assumed for intercept

### Effect of Labour Force Size (Wallerstein hypothesis)



**b2 (LabF):** Vague Info Vague  
**b3 (IndC):** Vague Vague Info

### Effect of Industrial Concentration (Stephens hypothesis)



**b2 (LabF):** Vague Info Vague  
**b3 (IndC):** Vague Vague Info

Comments

- Effects of LabF and IndC estimated more precisely
- Both sets of prior beliefs support inference that labour-force size decreases union density
- Only Stephens prior supports conclusion that industrial concentration increases union density
- Choice of prior is subjective – if no consensus, can we be satisfied that data have been interpreted fairly?
  - Sensitivity to priors (e.g. repeat analysis using priors with increasing variance) — see Practical exercises
  - Sensitivity to data (e.g. residuals, influence diagnostics) — see later lecture

## Multivariate responses

- In many applications, it is common to collect data on a number of different outcomes measured on the same units, e.g.
    - sample survey, where respondents asked several different questions
    - experiment with several different outcomes measured on each unit
  - May wish to fit regression model to each response
    - May have different covariates in each regression
    - But errors may be correlated
    - Might also wish to impose cross-equation parameter restrictions
- Seemingly Unrelated Regressions (SUR) (Zellner, 1962)

Bayesian approach to SUR models → model vector of response for each unit as multivariate normal (could also have robust version using multivariate t)

Possible to extend Bayesian SUR models to binary, categorical, count responses using multivariate latent variable approach.

## **Example: Analysis of compositional data**

- Compositional data are vectors of proportions  $\mathbf{p}_i = (p_{i1}, \dots, p_{iJ})$  representing relative contributions of each of  $J$  categories to the whole, e.g.
  - Proportion of income spent on different categories of expenditure
  - Proportion of electorate voting for different political parties
  - Relative abundance of different species in a habitat
  - Chemical composition of a rock or soil sample
  - Proportion of deaths from different causes in a population
- Regression models for compositional data must satisfy two constraints

$$0 \leq p_{ij} \leq 1$$
$$\sum_j p_{ij} = 1$$

- Two main modelling strategies, treating vector of proportions as the data (sufficient statistics)
  - Model  $\mathbf{p}_i$  using a Dirichlet likelihood (multivariate generalisation of a beta distribution)
    - \* assumes the ratios of “compositions” (i.e. proportions) are independent
  - Multivariate logistic normal model (Aitchison, 1986) — apply *additive log ratio* (alr) transformation,  $y_{ij} = \log(p_{ij}/p_{iJ})$  and model  $\mathbf{y}_i$  as multivariate normal
    - \* allows dependence between ratios of proportions
    - \* this can be thought of as a type of SUR model
- To allow for sampling variability in the observed counts (including zero counts), model counts (rather than proportions) as multinomial (see later)

## Multivariate logistic normal model

Define

$$y_{ij} = \log(p_{ij}/p_{iJ})$$

the log ratios of proportions in each category relative to a reference category  $J$

Note that  $p_{iJ} = 1 - \sum_{j \neq J} p_{ij}$ , so

$$p_{ij} = \frac{\exp y_{ij}}{1 + \sum_{j \neq J} \exp y_{ij}}$$

Since  $y_{ij}$  are unconstrained, can model vector  $\mathbf{y}_i = \{y_{ij}, j \neq J\}$  as multivariate normal

## **Example: British General Election 1992**

- Data originally analysed by Katz and King (1999), and formulated as BUGS example by Simon Jackman
- Data consist of vote proportions for Conservative ( $j=1$ ), Labour ( $j=2$ ) and Lib-Dem ( $j=3$ ) parties from 1992 General Election for each of 521 constituencies
- Additive log ratio transformation applied to proportions, taking Lib-Dem vote as reference category
- Covariates include lagged values of the log ratios from previous election, and indicators of the incumbency status of each party's candidate

*BUGS model code*

```
for(i in 1:521){
  y[i,1:2] ~ dnorm(mu[i,], prec[ , ])
  for(j in 1:2){
    mu[i,j] <- beta[j,1]*x[i,1] + beta[j,2]*x[i,2] + beta[j,3]*x[i,3] +
              beta[j,4]*x[i,4] + beta[j,5]*x[i,5] + beta[j,6]*x[i,6]
  }
}

## priors for elements of precision matrix
prec[1:2,1:2] ~ dwish(R[,],k)
R[1,1] <- .01; R[1,2] <- 0; R[2,1] <- 0; R[2,2] <- .01; k <- 2

# convert precision to covariance matrix
sigma[1:2,1:2] <- inverse(prec[ , ])
rho <- sigma[1,2]/sqrt(sigma[1,1]*sigma[2,2]) # correlation

## Priors for regression coefficients
for(j in 1:2){
  for(k in 1:6) {
    beta[j,k] ~ dnorm(0, 0.000001)
  }
}
```

## Priors on precision matrix of multivariate normal

The multivariate generalisation of the Gamma (or  $\chi^2$ ) distribution is the Wishart distribution, which arises in classical statistics as the distribution of the sum-of-squares-and-products matrix in multivariate normal sampling.

The Wishart distribution  $W_p(k, R)$  for a symmetric positive definite  $p \times p$  matrix  $\Omega$  has joint density function proportional to

$$|R|^{k/2} |\Omega|^{(k-p-1)/2} \exp(-(1/2)\text{tr}(R\Omega))$$

in terms of two parameters: a real scalar  $k > p - 1$  and a symmetric positive definite matrix  $R$ . The expectation of this distribution is

$$E[\Omega] = kR^{-1}$$

When the dimension  $p$  is 1, i.e. reverting back to univariate case, it is easy to show that the Wishart distribution becomes the more familiar:

$$W_1(k, R) \equiv \text{Gamma}(k/2, R/2) \equiv (\chi_k^2)/R$$

If we use the Wishart distribution as a prior distribution for a precision matrix  $\Omega$  in sampling from  $N_p(\mu, \Omega^{-1})$ , we find, generalising the univariate case above, that we get the same form for the posterior for  $\Omega$  – another Wishart distribution.

In view of the result above for the expectation of the Wishart distribution, we usually set  $(1/k)R$  to be a prior guess at the unknown true variance matrix. A common choice is to take  $k = p$ .

## Note

- In BUGS language, you must specify the dimension of vectors or arrays on the left hand side of multivariate distributions
- In above example, each row (observation) of the  $521 \times 2$  matrix  $\mathbf{y}$  is a vector of length 2, hence `y[i,1:2] ~ dmnorm.....`
- Likewise, `prec` is a  $2 \times 2$  matrix, hence `prec[1:2, 1:2] ~ dwish.....`
- You **cannot** specify the dimension to be a parameter (even if the value of the parameter is specified elsewhere in the code or data file), e.g.

```
J <- 2
prec[1:J, 1:J] ~ dwish(R[,], k)
```

will give an error at compilation

- You do not need to specify the dimension of vectors or arrays on the right hand side of distribution statements, e.g. dimension of `R[,]` is not specified above (the dimension is implicit from dimension of left hand side)

## **Interpretation of model parameters**

- Interpretation of parameter estimates on a multivariate log-odds scale is difficult
- Using the inverse alr transformation, easy to recover estimates of expected proportions or predicted counts in different categories
- Effect of covariates can be examined by calculating difference or ratio of expected proportions for different values of the covariate
  - Using MCMC, easy to obtain uncertainty intervals for such contrasts
- Example: effect of incumbency on expected proportion of votes for each party
  - For party  $j$ , calculate expected alr-transformed proportion for two values of incumbency: (1) party  $j$ 's candidate is incumbent; (2) open seat (no candidate is incumbent)
  - Hold values of all other covariates constant, e.g. at their means
  - Use inverse alr transformation to obtain expected proportions under each incumbency value, and take difference

*BUGS code for calculating incumbency effects*

```
for(j in 1:2){
  # value of mu with Conservative incumbent and average values of other variables
  mu.con[j] <- beta[j,1]*mean(x[,1]) + beta[j,2]*mean(x[,2]) +
              beta[j,3]*mean(x[,3]) + beta[j,4]*1

  # value of mu with Labour incumbent and average values of other variables
  mu.lab[j] <- beta[j,1]*mean(x[,1]) + beta[j,2]*mean(x[,2]) +
              beta[j,3]*mean(x[,3]) + beta[j,5]*1

  # value of mu with open seat and average values of other variables
  mu.open[j] <- beta[j,1]*mean(x[,1]) + beta[j,2]*mean(x[,2]) + beta[j,3]*mean(x[,3])

  # expected proportions
  exp.mu.con[j] <- exp(mu.con[j]); p.con[j] <- exp.mu.con[j]/(1 + sum(exp.mu.con[]))

  exp.mu.lab[j] <- exp(mu.lab[j]); p.lab[j] <- exp.mu.lab[j]/(1 + sum(exp.mu.lab[]))

  exp.mu.open[j] <- exp(mu.open[j]); p.open[j] <- exp.mu.open[j]/(1 + sum(exp.mu.open[]))
}

# difference in expected proportions due to incumbency
incumbency.con <- p.con[1] - p.open[1]
incumbency.lab <- p.lab[2] - p.open[2]
```

Results

	Posterior mean	95% CI
Expected vote (Cons), area 1	44.8%	(44.2%, 45.4%)
Expected vote (Lab), area 1	46.1%	(45.3%, 46.8%)
Expected vote (LibDem), area 1	9.1%	(8.7%, 9.5%)
incumbency advantage (Cons)	-0.06%	(-0.75%, 0.70%)
incumbency advantage (Lab)	-0.30%	(-1.6%, 1.0%)
incumbency advantage (LibDem)	8.6%	(5.1%, 12.3%)
$\rho$ (correlation between log ratio for Cons:LibDem and Lab:LibDem)	0.87	(0.85, 0.89)

Note: results for incumbency advantage agree with those in Tomz et al (2002) but not with Katz and King, who analysed data from 10 consecutive elections and used empirical Bayes shrinkage priors on the  $\beta$  coefficients across years

## **Multivariate t likelihood**

- Multivariate logistic normal for compositional data relies on assumption that the log ratios are approximately multivariate normal
- Katz and King (1999) argue that this assumption is not appropriate for British election data
  - majority of constituencies tend to be more clustered, and a minority more widely dispersed, than the multivariate normal implies
- K&K propose replacing multivariate normal by a heavier-tailed multivariate student t distribution
- Multivariate t has 3 parameters:  $p$ -dimensional mean vector,  $p \times p$  inverse scale (precision) matrix and a scalar degrees of freedom parameter
- A Wishart prior can be used for the inverse scale matrix
- Degrees of freedom parameter can either be fixed, or assigned a prior distribution
- Note: as degrees of freedom  $\rightarrow \infty$ , t  $\rightarrow$  Normal

*BUGS code for multivariate t likelihood*

- Only need to change 2 lines of code

1. Likelihood:

```
## y[i,1:2] ~ dmnorm(mu[i,1:2], prec[ , ])  
y[i,1:2] ~ dmt(mu[i,1:2], prec[ , ], nu)
```

2. Specify either fixed value for degrees of freedom, `nu`, or a suitable prior

```
## nu <- 4  
nu ~ dunif(2, 250)
```

## Results

	Normal	Student t
Expected vote (Cons), area 1	44.8% (44.2%, 45.4%)	44.7% (44.2%, 45.2%)
Expected vote (Lab), area 1	46.1% (45.3%, 46.8%)	46.4% (45.7%, 47.0%)
Expected vote (LibDem), area 1	9.1% (8.7%, 9.5%)	8.8% (8.6%, 9.3%)
incumbency advantage (Cons)	-0.06% (-0.75%, 0.70%)	0.00% (-0.65%, 0.60%)
incumbency advantage (Lab)	-0.30% (-1.6%, 1.0%)	-0.50% (-1.7%, 0.70%)
incumbency advantage (LibDem)	8.6% (5.1%, 12.3%)	2.3% (-2.0%, 8.3%)
$\rho$	0.87 (0.85, 0.89)	0.87 (0.85, 0.90)
$\nu$	–	4.5 (3.4, 5.9)
DIC	-480	-1220
$p_D$	14.5	11.8