# Bayesian data analysis (Level 11, 10 credits)

Natalia Bochkina, JCMB 4612; N.Bochkina@ed.ac.uk

Introduction to Bayesian data analysis: practical experience of applying Bayesian analyses to a range of statistical models. The statistical analyses will be conducted using the widely used computer package WinBUGS.

Pre-requisite: **Bayesian Theory**

Course structure: 12 lectures, 5 (2hr) computer labs.

Assessment: 100% coursework

- 4 assignments. Each counts towards 5% of coursework (20% total).
- 2 data analysis reports, each report counts for 40% of coursework.

**Reports without correctly completed cover sheets will not be accepted.**

Recommended books:

- Bayesian Data Analysis (3rd edition). Gelman, Carlin, Stern, Dunson, Vehtari and Rubin. CRC Press
- The BUGS Book: A Practical Introduction to Bayesian Analysis. Lunn, Jackson, Best, Thomas and Spiegelhalter. CRC Press

# 1. Introduction to Bayesian inference

# What are Bayesian methods?

- Bayesian methods have been widely applied in many areas of sciences and medicine
- 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

## Bayesians and frequentists differ in their concept of probability

- In Bayesian statistics, probability is used as a fundamental measure of uncertainty

  e.g. probability that it will rain tomorrow

  e.g. probability that an unknown quantity lies within a specified range

  e.g. probability that the percentage of individuals with a certain opinion is less than 20%

  → subjective probability

  - ! subjective probabilities should be "coherent", i.e. obey the law of probability
  - ! they should be constructed with scientific judgement to reflect current state of knowledge

- Frequentists think of probabilities as frequencies of events observed in a long run of repeated experiments

- Where both concepts are applicable, they are unlikely to give different answers; the main distinction is that Bayesians are prepared to use 'the same framework' to describe *all* their uncertainties.

## Bayesian inference

Distinguish fundamentally:

- Observable quantities $\mathrm{Y}$, i.e. the data
- Unknown quantities $\theta$

These can be statistical parameters, missing data, mismeasured data ...

→ parameters are treated as random variables

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

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

## Components of Bayesian inference

- The prior distribution $p(\theta)$

  – Expresses uncertainty or information available at the start of the study about unknown quantities (variables) by means of a probability distribution

- The likelihood $p(\mathrm{Y}|\theta)$

  – The basic sampling model needed in all probability-based inference, Bayesian or otherwise

  – Relate all variables into a 'full probability model' that summarises current knowledge on the random phenomenon

  – 'How are the data influenced by the parameters?' and hence 'How does observing the data change what we think about the parameters?'

- The posterior distribution $p(\theta|\mathrm{Y})$

  – After observation of some variables (the data), use Bayes theorem to obtain conditional probability distributions for unobserved quantities of interest

  – Expresses our uncertainty about $\theta$ after seeing the data

  – Bayes theorem tells us how to calculate this

# Bayesian inference

Suppose we observe data $\boldsymbol{y} = (y_1, \ldots, y_n)$ which we model as a realisation of random variable $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \mid \theta \sim f(\boldsymbol{Y} \mid \theta)$, $\theta \in \Theta$.

1. Before using any information from data $\boldsymbol{y}$, we assume there is a distribution over $\Theta$ called the prior distribution with p.d.f $p(\theta)$.
2. The parametric family of distributions with p.d.f. $f(\boldsymbol{y} \mid \theta)$ can be viewed as a conditional distribution of data $\boldsymbol{y}$ given $\theta$.
3. Can update our knowledge about $\theta$ using observed data $\boldsymbol{y}$ from $p(\theta)$ to the conditional distribution of $\theta$ given observed data $\boldsymbol{y}$, called posterior distribution of $\theta$, using Bayes theorem

$$p(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)p(\theta)}{\int_\Theta f(\boldsymbol{y}|\theta)p(\theta)d\theta},$$

which is $\propto f(\boldsymbol{y}|\theta)p(\theta)$ as a function of $\theta$. Thus,

posterior p.d.f. $\propto$ likelihood × prior p.d.f.

# Bayesian approach: summary

1. "Build" a probabilistic model for future data: $\boldsymbol{Y} \mid \theta \sim f(\boldsymbol{Y} \mid \theta), \theta \in \Theta$
2. Choose a prior distribution for $\theta$, $p(\theta)$ on $\Theta$
3. Observe data $\boldsymbol{y}$ - a realisation of $\boldsymbol{Y}$
4. Suppose we need to make a decision $\delta(\boldsymbol{y})$ based on observed data (e.g. estimate $\theta$, test a hypothesis about $\theta$, find an interval estimate - confidence interval):
    - choose a loss function $Q(\delta, \theta)$
    - find the decision that minimises the posterior expected loss function, i.e. given observed data:

$$\hat{\delta}_{Q,p}(\boldsymbol{y}) = \arg\min \mathbb{E}(Q(\delta, \theta) \mid \boldsymbol{y}) = \int Q(\delta, \theta) p(\theta \mid \boldsymbol{y}) d\theta.$$

5. If the choice of prior distribution was not fully informed, check sensitivity of the inference to the choice of prior.

Inference is based on posterior distribution

# Bayesian inference with binary data

**Example: Inference on proportions using discrete prior**

Assume treatment may have response rate $\theta$ of .2, .4, .6 or .8., each of equal prior probability. If we observe a single positive response ($x = 1$), how is our belief revised?

Likelihood, $p(x \mid \theta) = \theta^x(1-\theta)^{(1-x)}$

| $\theta$ | Prior<br>$p(\theta)$ | Likelihood<br>$p(x = 1 \mid \theta) = \theta$ | Likelihood × prior<br>$p(x = 1 \vert \theta)p(\theta)$ | Posterior<br>$p(\theta \vert x = 1) = \frac{p(x=1 \vert \theta)p(\theta)}{\sum_j p(x=1 \vert \theta_j)p(\theta_j)}$ |
|---|---|---|---|---|
| .2 | .25 | .2 | .05 | .10 |
| .4 | .25 | .4 | .10 | .20 |
| .6 | .25 | .6 | .15 | .30 |
| .8 | .25 | .8 | .20 | .40 |
| $\sum_j$ | 1.0 | | .50 | 1.0 |

Note: a single positive response makes it four times as likely that the true response rate is 80% rather than 20%.

## *Example: Surgical — Intervals on Ranks*

- Recent trend in UK towards ranking 'institutional' performance e.g. schools, hospitals

- Rank of a point estimate is a highly unreliable summary statistic

⇒ Would like measure of uncertainty about rank

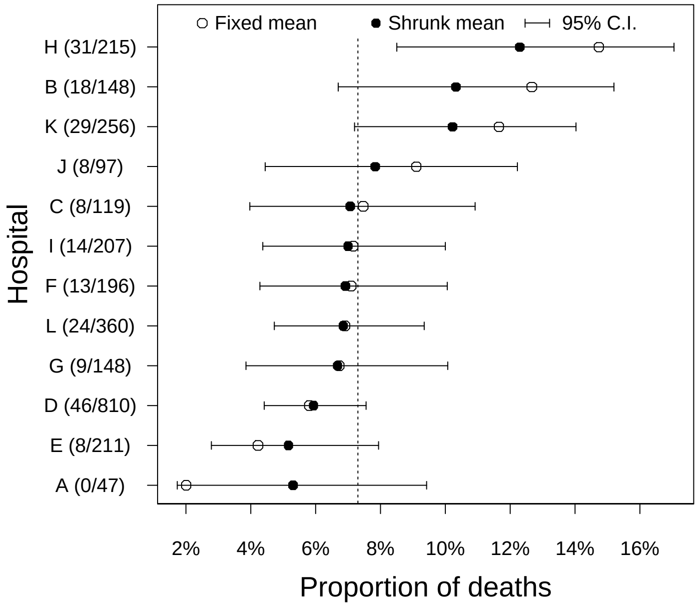
- Bayesian methods provide *posterior interval estimates* for ranks, e.g.

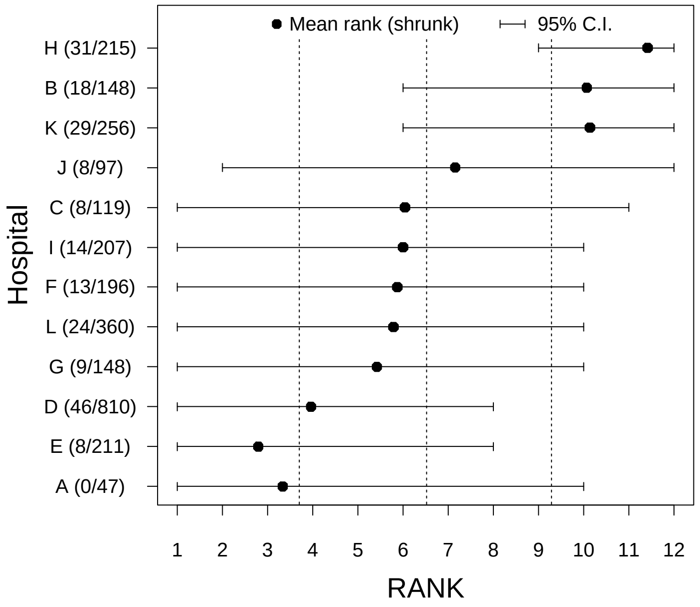
  - Surgical mortality rates in 12 hospitals carrying out cardiac surgery

  - Fit conjugate beta prior × binomial likelihood independently to each hospital to estimate posterior distribution of mortality rate, $\pi_i$

  - Using MCMC, rank sampled values of $(\pi_1, ...., \pi_{12})$ at each iteration → sample from posterior distribution of ranks for each hospital

Example 1: Neonatal cardiac surgery mortality in 12 European hospitals





## Example: Small area disease counts — hierarchical models for count data

*Aim*: to estimate relative risk of disease in small areas and look for evidence of geographical variation in risk that may indicate presence of environmental risk factor

*Data*: Counts of cases of childhood leukaemia and population in 873 electoral wards in London:
$y_i$ : observed number of leukaemias in area $i$,
$E_i$ : expected number of leukaemias in area $i$, adjusted for age, sex

*Parameters*
$\theta_i$ : underlying relative risk of leukaemia in area $i$

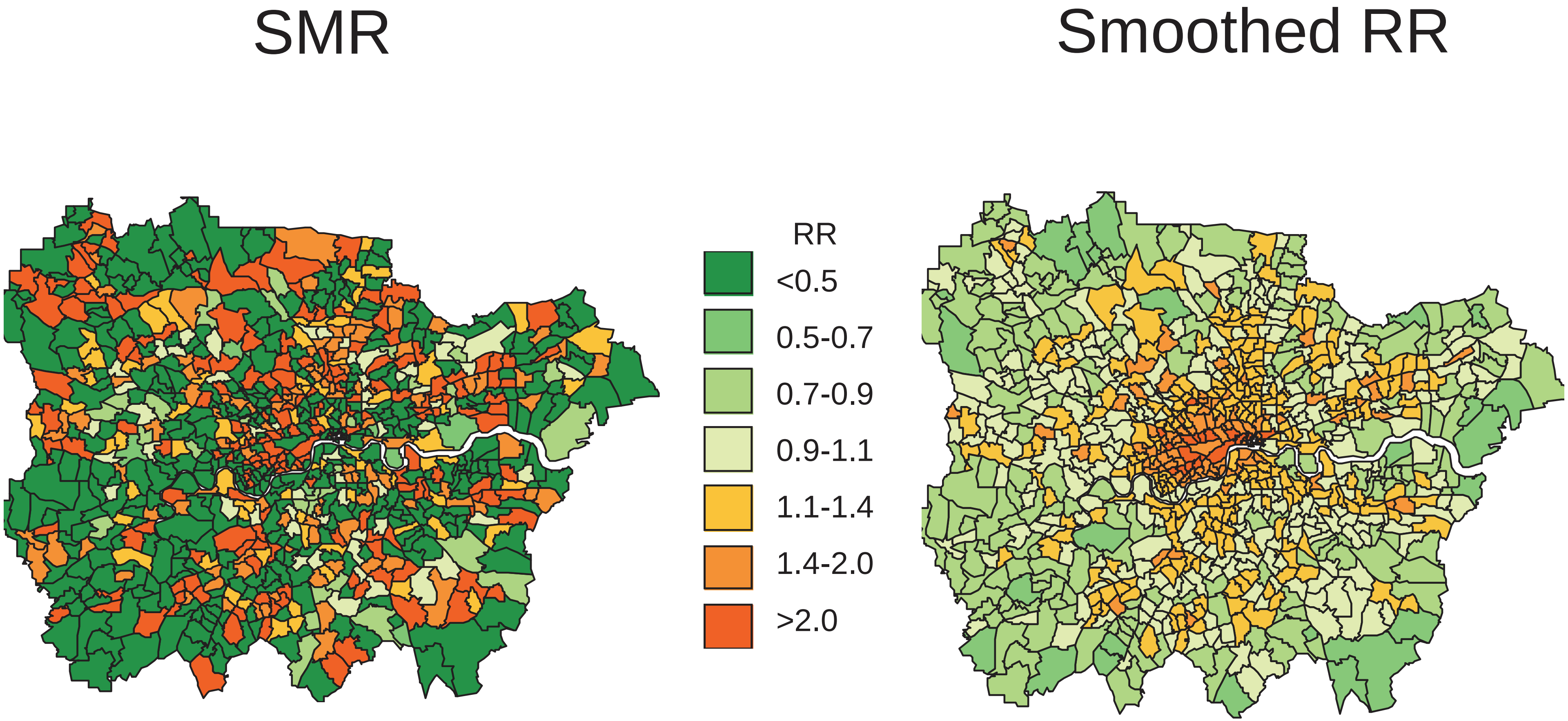
Likelihood (sampling variability within area):

$$y_i \sim \text{Poisson}(E_i\theta_i)$$

Prior:

$$\text{log}\theta_i \sim \text{N}(\mu,\sigma^2)$$

Priors on μ, σ


SMR
Smoothed RR
RR
<0.5
0.5-0.7
0.7-0.9
0.9-1.1
1.1-1.4
1.4-2.0
>2.0

## Summarising posterior distributions

Posterior distribution $p(\theta|\mathrm{Y})$ forms basis for **all** inference — can be summarised in different ways as appropriate to the inferential goal, e.g.

- mean, standard deviations, medians etc
- probability of exceeding certain thresholds, $p(\theta > \theta_0|\mathrm{Y})$
- credibility intervals

! Note that these have a direct probabilistic interpretation:
e.g. probability of 0.95 e.g that θ lies between 1.1 and 2

They are different from classical confidence intervals:

95% of the "95% confidence" intervals would contain the true value of θ in a long run of repeated experiments.

If $\theta$ is multi-dimensional, need techniques for integrating out nuisance parameters (see later).

# Bayesian inference: decision theory

| **Type of decision $\delta$** | **Loss** | **Optimal decision** |
|---|---|---|
| Estimation $\delta = \hat{\theta}$ | $Q(\theta, \delta) = \lvert\lvert\theta - \delta\rvert\rvert^2$ | $\hat{\delta}(\boldsymbol{y}) = \mathbb{E}(\theta \mid \boldsymbol{y})$ |
| Estimation $\delta = \hat{\theta}$ | $Q(\theta, \delta) = \lvert\theta - \delta\rvert$ | $\hat{\delta}(\boldsymbol{y}) = \text{Median}(\theta \mid \boldsymbol{y})$ |
| Estimation $\delta = \hat{\theta}$ | $Q(\theta, \delta) = I(\delta \neq \theta)$ | $\hat{\delta}(\boldsymbol{y}) = \text{MAP of } p(\theta \mid \boldsymbol{y})$ |
| Hypothesis testing:<br>$\delta = 1$ if accept $\Theta_1$,<br>$\delta = 0$ if accept $\Theta_0$ | $Q(\theta, \delta) = c_0$ if $\theta \in \Theta_0$ & $\delta = 1$<br>$Q(\theta, \delta) = c_1$ if $\theta \in \Theta_1$ & $\delta = 0$<br>$Q(\theta, \delta) = 0$ otherwise | $\hat{\delta}(\boldsymbol{y}) = 1$ if<br>$\frac{\mathbb{P}(\theta \in \Theta_1 \mid \boldsymbol{y})}{\mathbb{P}(\theta \in \Theta_0 \mid \boldsymbol{y})} > \frac{c_0}{c_1}$,<br>& $\hat{\delta}(\boldsymbol{y}) = 0$ o/w |
| $(1-\alpha)100\%$ upper CI:<br>$(-\infty, \delta] \cap \Theta$ | $Q_{1-\alpha}(\theta, \delta) = \alpha(\theta - \delta)I(\theta \geq \delta)$<br>$+(1-\alpha)(\delta - \theta)I(\theta < \delta)$ | $\hat{\delta} : P(\theta < \hat{\delta} \mid \boldsymbol{y}) = 1 - \alpha$ |
| $(1-\alpha)100\%$ CI:<br>two-sided<br>$\delta = [a, b]$ | $Q(\theta, \delta) = Q_{\alpha/2}(\theta, a)$<br>$+Q_{1-\alpha/2}(\theta, b)$ | $\hat{\delta} = [\hat{a}, \hat{b})$: equitailed<br>$P(\theta < \hat{b} \mid \boldsymbol{y}) = 1 - \alpha/2$<br>& $P(\theta < \hat{a} \mid \boldsymbol{y}) = \alpha/2$ |
| $(1-\alpha)100\%$ CI:<br>$\delta = [a, b]$ | $Q_k(\theta, \delta) = I(\theta \notin [a, b])$<br>$+k(b - a)$,<br>$k = k_\alpha \in (0, \max_{\theta \in \Theta}(p(\theta \mid \boldsymbol{y})))$ | $\hat{\delta} = [\hat{a}, \hat{b}]$ - HPDR<br>(high posterior<br>density region) |

*Relating an estimator or a decision to a particular loss function can help with interpretation.* Verification: exercise.

# Topics in this course

Types of models

- Linear and generalised linear models (fixed effects)
- Hierarchical Bayesian models: linear and generalised models with random effects
- Models with missing data
- Mixture models

Bayesian analysis techniques

- Choice of prior (with justification)
- Model checks and model comparison
- Graphical representation of Bayesian models
- Prior sensitivity analysis (can be viewed as a part of model checks)
- Posterior predictive approach

# Choice of prior distribution

- the 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 at 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)

# Noninformative (objective, ignorance, vague) prior

## Uniform priors

(Bayes 1763; Laplace, 1776)

Set $p(\theta) \propto 1$

Inference is based on the likelihood $p(\mathrm{y} \mid \theta)$

This is improper (\int $p(\theta)d\theta \neq 1$)

The posterior will still usually be proper

It is not really objective unless the range of θ is the whole real line, since a flat prior $p(\theta) \propto 1$ on $\theta$ does not correspond to a flat prior on φ = g(θ), but to p(φ) $\propto \frac{d\theta}{d\phi}$

# Noninformative: Jeffreys (invariance) prior

## Definition

Prior distribution with probability density $p(\theta) \propto \sqrt{\det(I(\theta))}$, $\theta \in \Theta$ is called *Jeffreys prior*.

For one-dimensional $\theta$, Jeffreys prior is $p(\theta) \propto \sqrt{I(\theta)}$. Here $I(\theta)$ is the Fisher information (matrix) for $\theta$:

$$I(\theta) = \mathbb{E}\left(\frac{\partial \log f(\boldsymbol{Y} \mid \theta)}{\partial \theta}\right)^2,$$

which in regular cases equals to $I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(\boldsymbol{Y}|\theta)}{\partial \theta^2}\right)$.

**Invariant to reparametrisation**. Say we want a prior for $\eta = \eta(\theta)$, e.g. $\eta = \theta^2$.

- Derive Jeffreys prior for $\eta = \eta(\theta)$ using the definition, $p_J(\eta)$.
- Use Jeffreys prior for $\theta$ and the rule for the change of variables: $p(\eta) = p_J(\theta)\left|\frac{d\theta}{d\eta}\right|$. These two distributions coincide.

# Noninformative: Jeffreys (invariance) prior

Has appealing property of giving more weight to values of $\theta$where the amount of information is larger.

**Definition**

Prior distribution with probability density $p(\theta) \propto \sqrt{\det(I(\theta))}$, $\theta \in \Theta$ is called *Jeffreys prior*.

For one-dimensional $\theta$, Jeffreys prior is $p(\theta) \propto \sqrt{I(\theta)}$. Here $I(\theta)$ is the Fisher information (matrix) for $\theta$:

$$I(\theta) = \mathbb{E}\left(\frac{\partial \log f(\boldsymbol{Y} \mid \theta)}{\partial\theta}\right)^2,$$

which in regular cases equals to $I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(\boldsymbol{Y}|\theta)}{\partial\theta^2}\right)$.

**Invariant to reparametrisation**. Say we want a prior for $\eta = \eta(\theta)$, e.g. $\eta = \theta^2$.

- Derive Jeffreys prior for $\eta = \eta(\theta)$ using the definition, $p_J(\eta)$.
- Use Jeffreys prior for $\theta$ and the rule for the change of variables: $p(\eta) = p_J(\theta)\left|\frac{d\theta}{d\eta}\right|$. These two distributions coincide.

## Examples of Jeffreys' priors

- Poisson case with parameter $\theta$

  $$\log p(x|\theta) = -\theta + x\log\theta + C \qquad \Rightarrow I(\theta) = 1/\theta$$

  So Jeffreys' prior is $\propto \theta^{-1/2}$
  This improper distribution is approximated by a Gamma distribution with $\alpha = 1/2$ and $\beta \to 0$

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

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

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

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

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

  So Jeffreys' prior on variance is $\propto \theta^{-1}$
  This improper distribution is approximated by a Gamma$(\epsilon, \epsilon)$ distribution with $\epsilon \to 0$
  Note: $p(\theta) \propto \theta^{-1}$ is equivalent to a uniform prior on $\log\theta$

Some inconsistencies associated with Jeffreys' priors have been discussed:

For example, applying this rule to the normal case with both mean and variance parameters unknown does not lead to the same prior as applying separately the rule for the mean and the variance and assuming a priori independence between these parameters.

*Various 'non-informative' priors for the binomial parameter*

Consider $r$ successes from $n$ trials: $r \sim \text{Binom}(n, \theta)$, then

$$\log p(r|\theta) = r \log \theta + (n-r) \log(1-\theta) + C$$

and $I(\theta) = \frac{n}{\theta(1-\theta)}$.

Thus Jeffreys' prior is

$$p(\theta) \;\propto\; (\theta(1-\theta))^{-\frac{1}{2}},$$

which is a Beta(1/2, 1/2) distribution .

The Bayes-Laplace Uniform density is a Beta(1,1) distribution.

A prior density that is uniform for logit$\theta$ is $\propto$ to $(\theta(1-\theta))^{-1}$, which is the improper Beta(0,0) distribution.

In practise, there will not be much difference between these alternatives, but the improper Beta(0,0) prior distribution leads to an improper posterior if $r = 0$ (or $n = 0$)!

*Location parameters*

*e.g.* means, regression coefficients:

$$\begin{aligned}\alpha &\sim \text{Unif}(-100, 100)\\ \alpha &\sim \text{Normal}(0, 100000)\end{aligned}$$

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

*Scale parameters*

- Sample variance $\sigma^2$: standard ‘reference’ prior

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

- Variance of random effects $\phi^2$: standard ‘reference’ prior will give an **improper** posterior distribution since $\phi^2 = 0$, is supported by non-negligible likelihood
  - A number of alternatives have been suggested — see later

**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.

# Conjugate prior distributions

## Definition

Suppose a family of prior p.d.f.'s for $\theta$ has the form $\{p(\theta;\alpha) : \alpha \in A\}$ where the index $\alpha$ has finite dimension. The family is **conjugate** with respect to the p.d.f. $f(\boldsymbol{y}|\theta)$ if for any $\boldsymbol{y}$, a prior p.d.f. in the family leads to a posterior p.d.f. also in the family.

**Remark.** If $\{p(\theta;\alpha) : \alpha \in A\}$ is a family of conjugate priors with respect to the p.d.f. $f(\boldsymbol{y}|\theta)$, then the following family is also conjugate:

$$\left\{ \sum_{j=1}^{N} \pi_j p(\theta;\alpha_j) : \quad \alpha_j \in A \,\&\, \pi_j \in (0,1)\ \forall j = 1,\ldots,N,\ \sum_{j=1}^{N} \pi_j = 1 \right\}$$

which is true for any $N \geq 1$. This is called a mixture prior.

# Conjugate prior distributions

**Definition**

Suppose a family of prior p.d.f.'s for $\theta$ has the form $\{p(\theta;\alpha) : \alpha \in A\}$ where the index $\alpha$ has finite dimension. The family is **conjugate** with respect to the p.d.f. $f(\boldsymbol{y}|\theta)$ if for any $\boldsymbol{y}$, a prior p.d.f. in the family leads to a posterior p.d.f. also in the family.

**Remark.** If $\{p(\theta;\alpha) : \alpha \in A\}$ is a family of conjugate priors with respect to the p.d.f. $f(\boldsymbol{y}|\theta)$, then the following family is also conjugate:

$$\left\{ \sum_{j=1}^{N} \pi_j p(\theta;\alpha_j) : \quad \alpha_j \in A \,\&\, \pi_j \in (0,1)\ \forall j = 1,\ldots,N,\ \sum_{j=1}^{N} \pi_j = 1 \right\}$$

which is true for any $N \geq 1$. This is called a mixture prior.

# Conjugate priors

When the posterior is in the same family as the prior then we have what is known as *conjugacy*.

| Likelihood | Target Parameter | Prior | Posterior |
|---|---|---|---|
| Binomial | Proportion | Beta | Beta |
| Poisson | Mean/Rate | Gamma | Gamma |
| Normal | Mean | Normal | Normal |
| Normal | Precision | Gamma | Gamma |
| MVN | Mean vector | MVN | MVN |
| MVN | Precision matrix | Wishart | Wishart |

- Conjugate priors have the advantage that
  - Computation is easy!
  - Prior parameters can usually be interpreted as a *prior sample* → aids elicitation
- Unfortunately conjugate priors do not exist for all likelihoods, and could be restrictive.
- Non-conjugate priors can also be used, but computations usually harder
  ⇒ need numerical or simulation-based computational methods (see later)

# Data reduction and sufficient statistics

**Sufficient statistic:**

$s$ is sufficient for $\theta$ iff $f(\boldsymbol{y}|\theta)$ can be factored as $f(\boldsymbol{y}|s)f(s|\theta)$.

In this case, for any $p(\theta)$,

$$p(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|s)f(s|\theta)p(\theta)}{\int_\Theta f(\boldsymbol{y}|s)f(s|\theta)p(\theta)d\theta} = \frac{f(s|\theta)p(\theta)}{\int_\Theta f(s|\theta)p(\theta)d\theta} = p(\theta|s),$$

i.e. the posterior distribution given $\boldsymbol{y}$ is the same as the posterior distribution given $s$.

This is used in practice for reduction of high dimensional data without loss of information about $\theta$.

# Predictive inference

Predictive distributions are **distributions for observable random variables**, not involving unknown parameters.

- Prior predictive p.d.f. of $\boldsymbol{y}$ is

  $$f(\boldsymbol{y}) = \int_{\Theta} f(\boldsymbol{y}|\theta)p(\theta)d\theta$$

  (also called marginal distribution of the data, usually in the frequentist approach).
  Use:
  - check whether the model (likelihood + prior) gives unreasonable predictions.
  - sample size calculations, e.g. in clinical trials.
- Posterior predictive distribution of a **future** observation $z$, given data $\boldsymbol{y}$ (and the likelihood and the prior) has p.d.f.

  $$f(z|\boldsymbol{y}) = \int_{\Theta} f(z|\theta)p(\theta|\boldsymbol{y})d\theta.$$

  Thus the p.d.f. of $z$ given $\theta$ is averaged over the posterior distribution of $\theta$.

## **Prediction** (=posterior prediction)

Say, we wish to predict the outcome of a new observation $x^{\sim}$, given what we have already observed.

For discrete $\theta$ we have

$$p(\tilde{x}|x) = \sum_{\theta_j} p(\tilde{x}, \theta_j|x)$$

which is generally equal to

$$p(\tilde{x}|x) = \sum_{\theta_j} p(\tilde{x}|\theta_j) P(\theta_j|x)$$

where the $P(\theta_j|x)$ can be thought of as 'posterior weights'.

In example, predictive probability of treatment outcome for a new patient is:

$$\begin{aligned} p(\tilde{x} = 0|x = 1) &= \sum_{\theta_j} (1 - \theta_j) p(\theta_j|x = 1) \\ &= (0.8) \times 0.1 + (0.6) \times 0.2 + (0.4) \times 0.3 + (0.2) \times 0.4 = 0.4 \\ p(\tilde{x} = 1|x = 1) &= \sum_{\theta_j} \theta_j p(\theta_j|x = 1) \\ &= 0.2 \times 0.1 + 0.4 \times 0.2 + 0.6 \times 0.3 + 0.8 \times 0.4 = 0.6 \end{aligned}$$

## 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

$$p(r|n,\theta) \;=\; \left(\begin{array}{c} n \\ r \end{array}\right)\theta^r(1-\theta)^{n-r} \;\propto\; \theta^r(1-\theta)^{n-r}$$

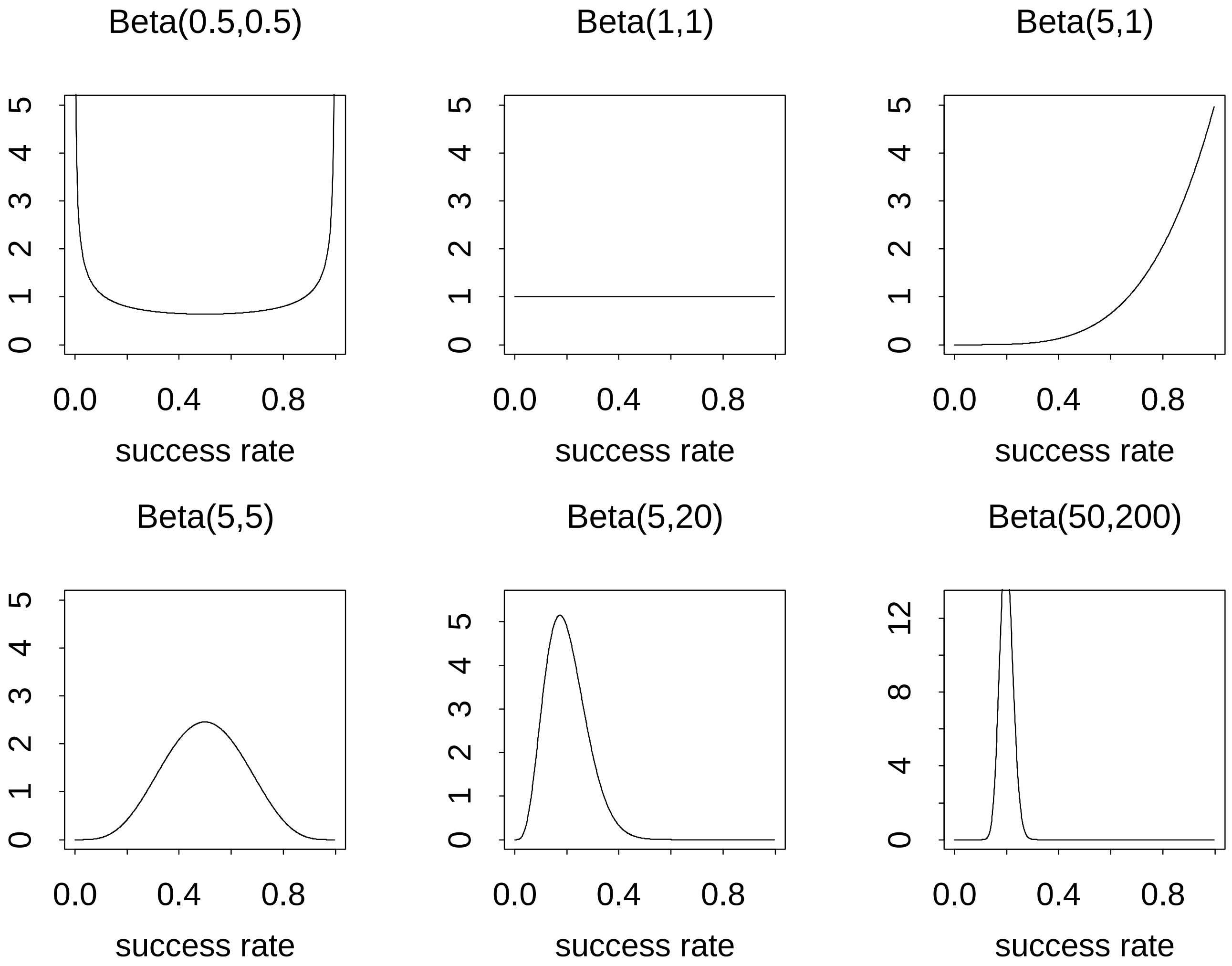
Also, more reasonable to consider the response rate, $\theta$ to be a be a continuous parameter $\Rightarrow$ needs to be given a continuous prior distribution.

Mathematically convenient to use a Beta$(a,b)$ prior distribution for $\theta$, which has the form

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

Combining this with the binomial likelihood gives a posterior distribution

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



A Beta($a$, $b$) distribution has

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

## Comments

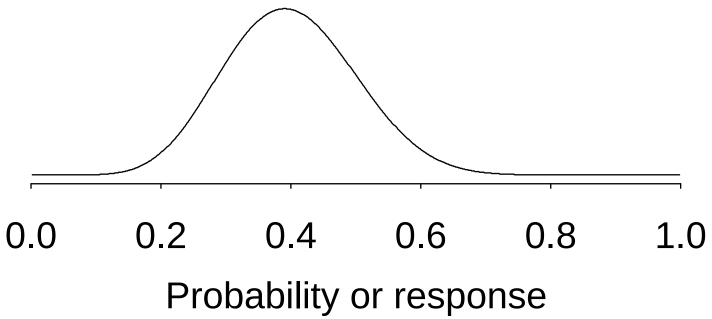
- the prior is **conjugate** to the likelihood

- Prior mean is $a/(a+b)$
  Data 'mean' (MLE) is $r/n$
  Posterior mean is $(r+a)/(n+a+b)$
  - can interpret prior information as being equivalent to having observed $a-1$ successes in $a+b-2$ prior trials

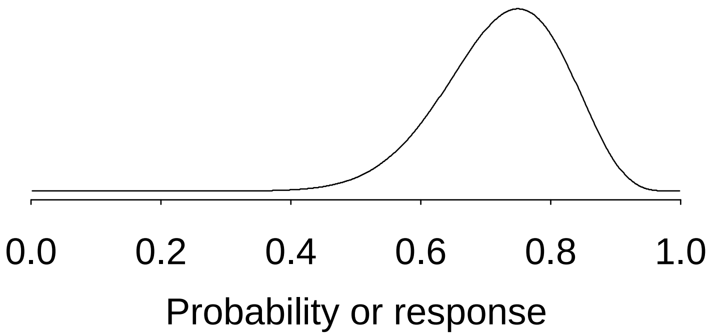
- With fixed $a$ and $b$, as $r$ and $n$ increase, $E(\theta|r,n) \to r/n$ (the MLE), and the variance tends to zero
  - This is a general phenomenon: as $n$ increases, posterior distribution gets more concentrated and the likelihood dominates the prior

## Example: Drug

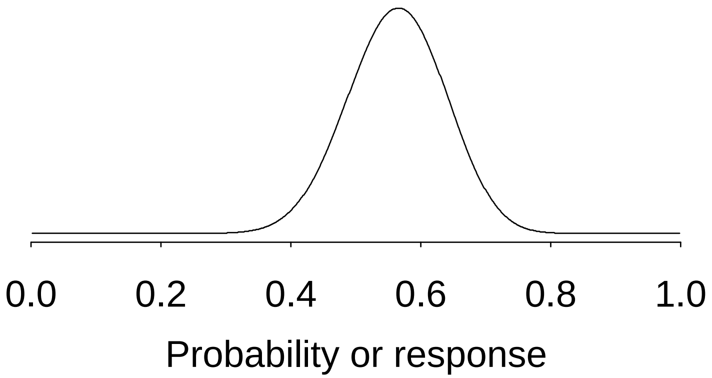
- Consider early investigation of a new drug
- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible
- Interpret this as a distribution with mean = 0.4, standard deviation 0.1
- A Beta(9.2,13.8) distribution has these properties
- Suppose we treat $n = 20$ volunteers with the compound and observe $r = 15$ 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

Suppose we want to predict the consequences of further experimentation

The predictive distribution for $\tilde{r}_m$ (the number of treatment successes in $m$ new patients) is

$$\begin{aligned} p(\tilde{r}_m|r,n) &= \int p(\tilde{r}_m,\theta|r,n)d\theta \\ &= \int p(\tilde{r}_m|\theta,r,n)p(\theta|r,n)d\theta \end{aligned}$$

which generally simplifies to

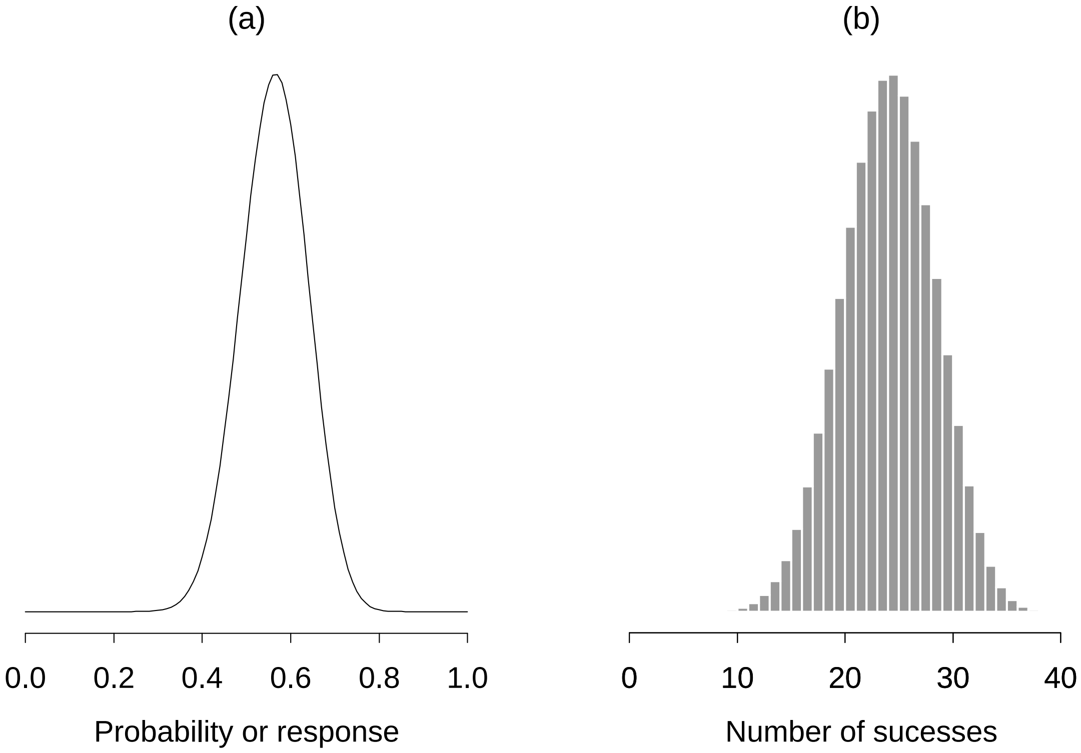
$$p(\tilde{r}_m|r,n) = \int p(\tilde{r}_m|\theta)p(\theta|r,n)d\theta$$

If $p(\theta|r,n) =$ Beta$(c,d)$, predictive distribution is known as the **Beta-Binomial**

$$p(\tilde{r}_m) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} \left( \begin{array}{c} m \\ \tilde{r}_m \end{array} \right) \frac{\Gamma(c+\tilde{r}_m)\Gamma(d+m-\tilde{r}_m)}{\Gamma(c+d+m)}.$$

(obtained by integrating Binomial likelihood wrt Beta posterior distribution for the success rate)

$$\mathsf{E}(\tilde{r}_m) = m\frac{c}{c+d}$$



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

(b) predictive Beta-Binomial distribution of the number of successes $\tilde{r}_{40}$ 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 $P(\tilde{r}_{40} \geq 25) = 0.329$

## Sequential learning

Suppose we obtain data $x$ and form the posterior $p(\theta|x)$ and then we obtain further data $z$. The posterior based on $x, z$ is given by:

$$p(\theta|x, z) \propto p(z|\theta) \times p(\theta|x)$$

'Today's posterior is tomorrow's prior !'

The resultant posterior is the same as if we have obtained the data $x, z$ together:

$$p(\theta|x, z) \propto p(x, z|\theta) \times p(\theta)$$

## Example: Drug (continued)

Suppose we observed a further $r' = 5$ positive responses in $n' = 10$ new volunteers

Using the Beta(24.2, 18.8) posterior from the previous trial as our prior, and conditioning on the new data gives a revised posterior for the success rate

$$\begin{aligned} p(\theta \mid r', n') &\propto p(r' \mid \theta, n') p(\theta) \\ &\propto \theta^{5} \, (1-\theta)^{10-5} \, \theta^{24.2-1} \, (1-\theta)^{18.8-1} \\ &= \theta^{5+24.2-1} \, (1-\theta)^{10-5+18.8-1} \\ &\propto \text{Beta}(29.2,\ 23.8) \end{aligned}$$

Alternatively, we could start with our original Beta(9.2, 13.8) prior and pool the data from the two trials to give a posterior for $\theta$

$$\begin{aligned} p(\theta \mid r + r', n + n') &\propto p(r + r' \mid \theta, n + n') p(\theta) \\ &\propto \theta^{15+5} \, (1-\theta)^{20+10-15-5} \, \theta^{9.2-1} \, (1-\theta)^{13.8-1} \\ &\propto \text{Beta}(29.2,\ 23.8) \end{aligned}$$

# Prior sensitivity

Choice of prior is very important.

1. If there is a priori information, use it to derive an appropriate prior
2. If there is no a priori information, use a “noninformative” prior

In either case, check for unintended sensitivity with respect to the choice of prior.

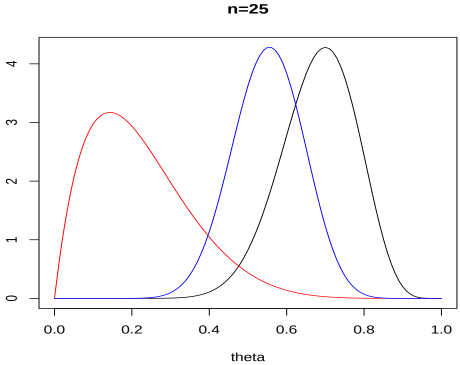
Ways to check for prior sensitivity:

1. Check for a conflict between prior and likelihood (prior/likelihood/posterior plot)
2. Predictive checks, i.e. if the model prediction is appropriate
3. In regression models: check residuals (Pearson, deviance)

**Example** $X \mid \theta \sim Bin(n, \theta)$, $\theta \in (0, 1)$.
Observe value $x$: $\bar{x} = \frac{x}{n} = 0.7$.
Prior: $\theta \sim Beta(a, b)$ (conjugate).
Posterior: $\theta \mid x \sim Beta(a + x, b + n - x)$.
*Choice of parameters for the prior: $a = 2$, $b = 7$ (also $b = 500, 5$)*
*Also compare with ‘non-informative’ Jeffreys prior.*

# Prior sensitivity

## Prior/likelihood/posterior plots

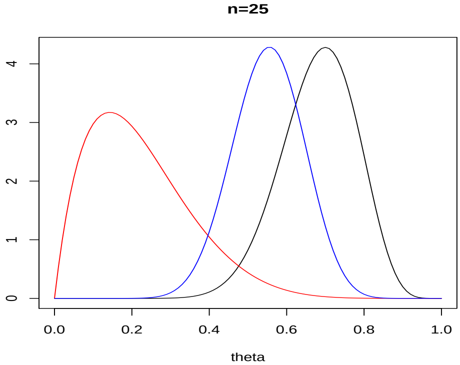


Interpretation?

*Prior has some effect on the posterior but most information comes from the likelihood.*

# Prior sensitivity

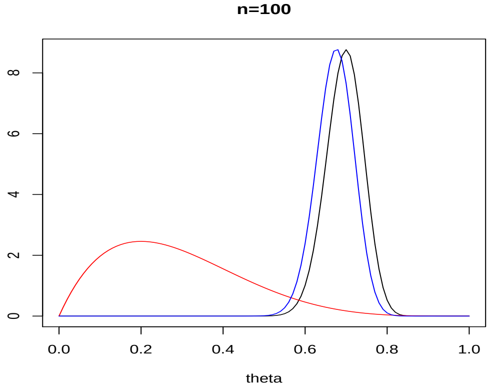
Prior/likelihood/posterior plots



Interpretation?
*Prior has some effect on the posterior but most information comes from the likelihood.*

# Prior sensitivity

Prior/likelihood/posterior plots

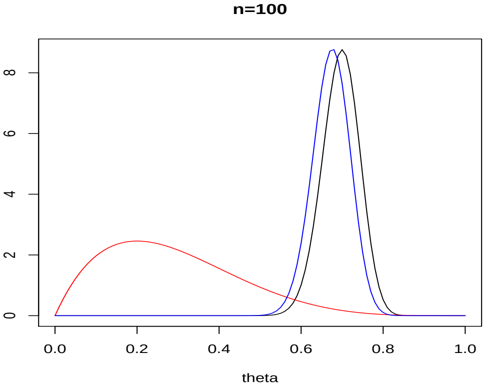


Interpretation?

*Prior has very weak effect on the posterior, most information comes from the likelihood.*

# Prior sensitivity

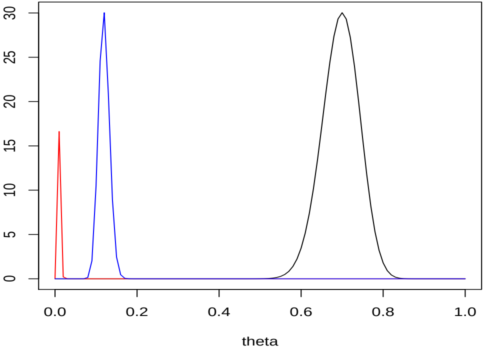
Prior/likelihood/posterior plots



Interpretation?
*Prior has very weak effect on the posterior, most information comes from the likelihood.*

# Prior sensitivity

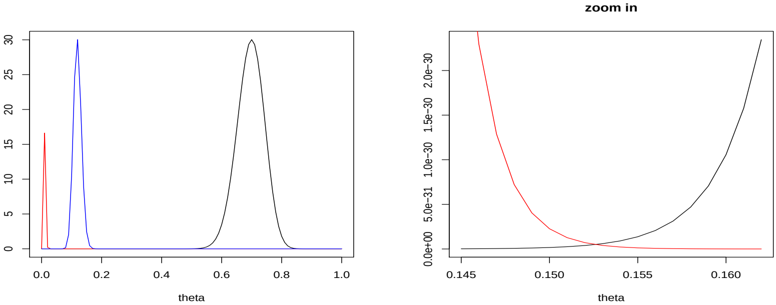
Prior/likelihood/posterior plots



Interpretation?

# Prior sensitivity

## Prior/likelihood/posterior plots

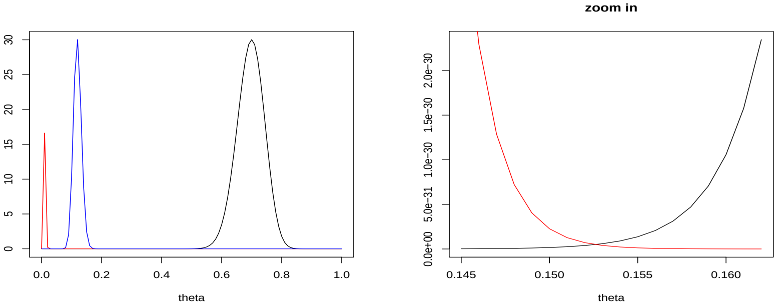


Interpretation.

*Support of the prior and likelihood have very little overlap, so the posterior is noninformative, and prior should be reconsidered.*

# Prior sensitivity

Prior/likelihood/posterior plots

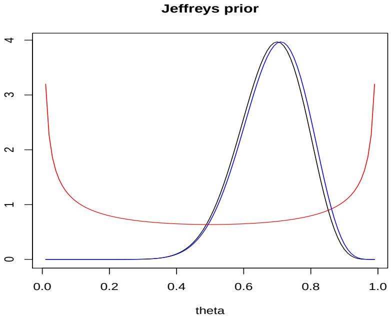


Interpretation.

*Support of the prior and likelihood have very little overlap, so the posterior is noninformative, and prior should be reconsidered.*

# Prior sensitivity

## Prior/likelihood/posterior plots

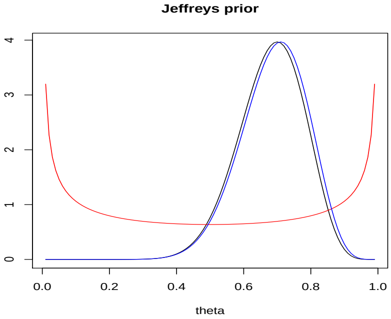


Interpretation. *Prior has effectively no effect on the posterior, the information essentially comes from the likelihood.*

**We will look at other methods for model checking and model comparison**: *residuals, prediction-based checks, DIC (Deviance Information Criterion).*

# Prior sensitivity

Prior/likelihood/posterior plots

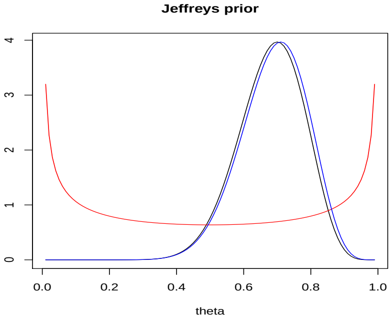


Interpretation. *Prior has effectively no effect on the posterior, the information essentially comes from the likelihood.*

**We will look at other methods for model checking and model comparison**: *residuals, prediction-based checks, DIC (Deviance Information Criterion).*

# Prior sensitivity

Prior/likelihood/posterior plots



Interpretation. *Prior has effectively no effect on the posterior, the information essentially comes from the likelihood.*

**We will look at other methods for model checking and model comparison**: *residuals, prediction-based checks, DIC (Deviance Information Criterion).*

# Possible prior distributions for $\theta \in [0, 1]$

**Informative**

1. $\theta \sim Beta(a, b)$ - conjugate
2. $\theta \sim pBeta(a, b) + (1 - p)Beta(c, d)$ - conjugate; if there is information that $\theta$ can come from two populations
3. $\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \sim N(\mu_0, \sigma_0^2)$ - non-conjugate; excludes values $\theta = 0, 1$
4. $\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \sim t_k(\mu_0, \sigma_0^2)$ - non-conjugate, 'robustified' ; excludes $\theta = 0, 1$

**'Non-informative'**

1. $\theta \sim Beta(1/2, 1/2)$ - Jeffreys prior
2. $\theta \sim Beta(1, 1) = U[0, 1]$ - uniform prior
3. $\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$: $\quad p(\text{logit}(\theta)) = 1$ - improper prior, corresponds to $\theta \sim Beta(1, 1)$
4. $\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \sim N(0, A^2)$ with large $A$, e.g. $A = 10^3$ - proper approximation to the above.

The lists are not exhaustive.

# Bayesian inference using count data

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

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

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

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

The posterior is another (different) Gamma distribution.

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

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

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

# Bayesian inference using the Normal distribution

## Known variance, unknown mean

Suppose we have a sample of Normal data $x_i \sim \mathrm{N}(\theta, \sigma^2) \quad (i = 1, ..., n)$. For now assume $\sigma^2$ is known and $\theta$ has a Normal prior $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$

Then the posterior distribution is

$$\begin{aligned} p(\theta|\boldsymbol{x}) &\propto \prod_i p(x_i \,|\, \theta)\ p(\theta) \\ &\propto \exp\left[-\frac{\sum_i (x_i - \theta)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\theta - \mu)^2 n_0}{2\sigma^2}\right] \end{aligned}$$

By matching terms in $\theta$ and writing $\sum x_i = n\overline{x}$ it can be shown that

$$\sum_i (x_i - \theta)^2 + (\theta - \mu)^2 n_0 = \left(\theta - \frac{n_0\mu + n\overline{x}}{n_0 + n}\right)^2 (n_0 + n) + \text{constant}$$

The term involving $\theta$ is exactly that arising from a Normal distribution, so

$$p(\theta|\boldsymbol{x}) = \mathrm{N}\left(\frac{n_0\mu + n\overline{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right)$$

- Same standard deviation $\sigma$ is used in the likelihood and the prior

- Prior variance is based on an implicit ‘prior sample size’ $n_0$

- As $n_0$ tends to 0, the variance becomes larger and the distribution becomes ‘flatter’, and in the limit the distribution becomes essentially uniform over $-\infty, \infty$

- Posterior mean $(n_0\mu + n\overline{x})/(n_0 + n)$ is a weighted average of the prior mean $\mu$ and parameter estimate $\overline{x}$, 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 $n$

*Large sample properties*

As $n \to \infty$,

$$\begin{aligned} \text{posterior mean,}(n_0\mu + n\bar{x})/(n_0+n) &\to \bar{x} \\ \text{posterior variance,}\sigma^2/(n_0+n) &\to \sigma^2/n \\ \text{and so posterior distribution,}p(\theta|\boldsymbol{x}) &\to \mathsf{N}(\bar{x}, \sigma^2/n) \end{aligned}$$

which do not depend on the prior

In the frequentist setting, the MLE is $\widehat{\theta} = \bar{x}$ with SE$(\widehat{\theta}) = \sigma/\sqrt{n}$, and sampling distribution

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

whereas in the Bayesian framework, the "dual statement" is made:

$$p(\theta \mid \bar{x}) \to \mathsf{N}(\bar{x}, \sigma^2/n)$$

## Prediction

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

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

which generally simplifies to

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

which can be shown to give

$$p(\tilde{x}|\boldsymbol{x}) \sim \mathsf{N}\left(\mu_n, \sigma_n^2 + \sigma^2\right)$$

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

## Bayesian inference for Normal data with unknown variance, known mean

Suppose now – and this is not realistic, but just to make a point – that $\theta$ is known, but the variance $\sigma^2$ unknown. It is often convenient in Bayesian statistics to work with the *precision* $\tau = \sigma^{-2}$ instead of the variance (mainly because, loosely speaking, the posterior precision of parameters is additive over independent data).

Let us take a flexible form of prior for $\tau$, say a Gamma distribution:

$$\tau \sim \text{Gamma}(\alpha, \beta)$$

Then the posterior is

$$\begin{aligned} p(\tau|x_1, x_2, \ldots, x_n) &\propto p(\tau)p(x_1, x_2, \ldots, x_n|\tau) \\ &\propto \tau^{\alpha-1}e^{-\beta\tau} \times \tau^{n/2} \exp\left[-(\tau/2)\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned}$$

that is,

$$\tau|x_1, x_2, \ldots, x_n \sim \text{Gamma}(\alpha + n/2,\ \ \beta + (1/2)\sum_{i=1}^{n}(x_i - \mu)^2)$$

That is, if the prior for $\sigma^2$ is Inverse-Gamma, so is the posterior.

## How do we interpret the ‘hyperparameters’ $\alpha$ and $\beta$?

$2\alpha$ and $2\beta$ can be interpreted as a ‘prior sample size’ and a ‘prior sum of squares’; the ratio $\beta/\alpha$ is a ‘prior guess‘ at the variance $\sigma^2$, and the size of $\alpha$ measures how sure we are about that guess. A common choice is to take both $\alpha$ and $\beta$ very small, then the posterior is approximately

$$\tau | x_1, x_2, \ldots, x_n \approx \text{Gamma}(n/2, (1/2)\sum_{i=1}^{n}(x_i - \mu)^2)$$

and so

$$E[\tau | x_1, x_2, \ldots, x_n] \doteq \left((1/n)\sum_{i=1}^{n}(x_i - \mu)^2)\right)^{-1}$$

– the posterior expectation of the precision is (approximately) the sample precision (but with a divisor of $n$ not $n-1$).

Note that the conjugate prior for variance of a Normal distribution can be parameterised in various ways:

- Gamma($\alpha$, $\beta$) prior on the inverse variance
- Inverse-gamma($\alpha$, $\beta$) prior on the variance
- Scaled-inverse-chi-squared($d$, $\phi^2$) on the variance, where $d = 2\alpha$ and $\phi^2 = \beta/\alpha$

Only the Gamma parameterisation is directly implemented in WinBUGS

## Multivariate Normal response - unknown variance

With multivariate data, we need to think about a variance matrix $\Sigma$ instead of a variance $\sigma^2$, or equivalently a precision matrix $\Omega$ instead of a precision $\tau$.

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 $\mathsf{W}_p(k, R)$ for a symmetric positive definite $p \times p$ matrix $\Omega$ has joint density function proportional to

$$|R|^{k/2}|\Omega|^{(k-p-1)/2}\,\mathrm{exp}\left(-(1/2)\mathrm{tr}(R\Omega)\right)$$

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, that is reverting from the multivariate to the univariate case, it is easy to see that the Wishart distribution becomes more familiar:

$$\mathsf{W}_1(k, R) \equiv \mathrm{Gamma}(k/2, R/2) \equiv (\chi^2_k)/R$$

If we use the Wishart distrubution 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 $k = p$.

# 2. Bayesian computation and MCMC

# Why is computation important?

- Bayesian inference centres around the posterior distribution

  $$p(\theta, \phi|x) \propto p(x|\theta, \phi) \times p(\theta, \phi)$$

  where $\theta$ is of interest, $\phi$ is nuisance

- $p(x|\theta, \phi)$ and $p(\theta, \phi)$ will often be available in closed form, but $p(\theta, \phi|x)$ is usually not analytically tractable, and we want to
  - obtain marginal posterior $p(\theta|x) = \int p(\theta, \phi|x)\ d\phi$
  - calculate properties of $p(\theta|x)$, such as mean, tail areas etc.

→ numerical integration becomes vital

## General Monte Carlo Integration

Standard software packages such as R have in-built algorithms for sampling from binomial and other standard 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 $(\theta, \phi)$, *i.e.*

$$(\theta^{(1)}, \phi^{(1)}), (\theta^{(2)}, \phi^{(2)}), ..., (\theta^{(N)}, \phi^{(N)}) \quad \sim \quad p(\theta, \phi|x)$$

- Then
  - $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are a sample from the marginal posterior $p(\theta|x)$
  - $E(g(\theta)) \ = \ \int g(\theta)p(\theta|x)d\theta \ \approx \ \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)})$

    → this is Monte Carlo integration

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

## How do we sample from the posterior?

- In general, we want samples from the joint posterior distribution $p(\boldsymbol{\theta}|x)$ (where now we use $\boldsymbol{\theta}$ to denote vector of all model parameters, including nuisance parameters)

- *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 $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, ...$ forms a Markov chain if

  $$\theta^{(i+1)} \quad \sim \quad p(\theta|\theta^{(i)})$$

  *i.e.* conditional on the value of $\theta^{(i)}$, $\theta^{(i+1)}$ is independent of $\theta^{(i-1)}, ..., \theta^{(0)}$

- Theorems exist which show that

  $\frac{1}{n}\sum_{i=1}^{n} g(\theta^{(i)}) \rightarrow E(g(\theta))$ as $n \rightarrow \infty$

  when $\theta^{(1)}, ..., \theta^{(n)}$ are sampled from a suitable Markov chain

## The Gibbs sampler

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

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

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

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

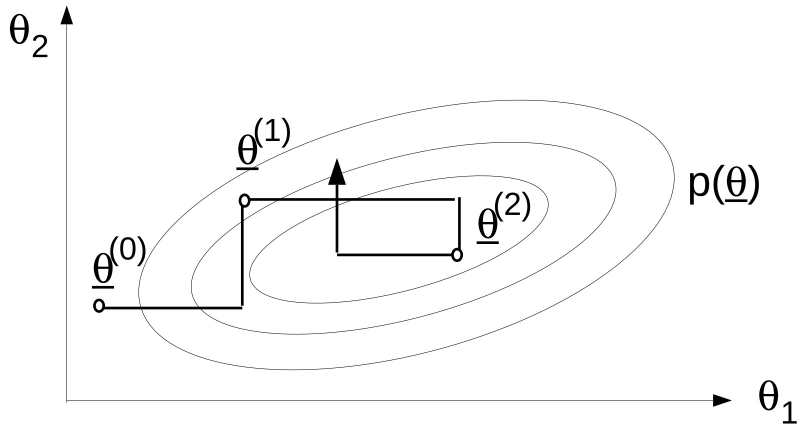
   .....

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

3) Repeat step 2 many 1000s of times
   - eventually obtain sample from $p(\boldsymbol{\theta}|x)$

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

## Gibbs sampling ctd.



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

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

# Performance of MCMC methods

There are three main issues to consider

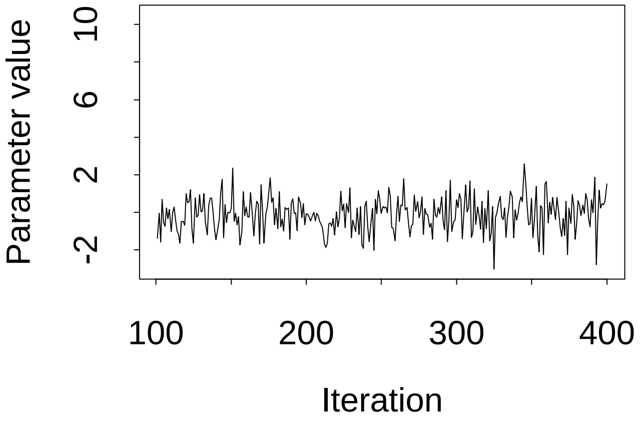
- Convergence (how quickly does the distribution of $\boldsymbol{\theta}^{(t)}$ approach $p(\boldsymbol{\theta}|x)$?)
- Efficiency (how well are functionals of $p(\boldsymbol{\theta}|x)$ estimated from $\{\boldsymbol{\theta}^{(t)}\}$?)
- Simplicity (how convenient is the method to use?)

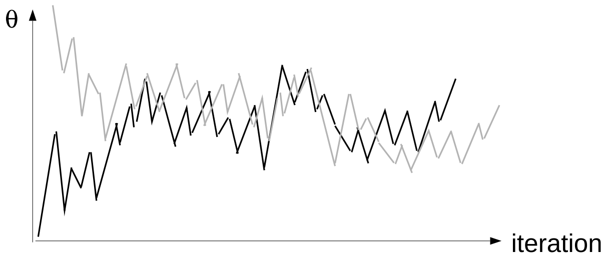
Note that computer effort should be measured in seconds, not iterations!

## 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.





- One approach is to run many long chains with widely differing starting values.

Another: use Brooks-Gelman-Rubin statistic

## How many iterations after convergence?

- After convergence, further iterations are needed to obtain samples for posterior inference.
- More iterations = more accurate posterior estimates.
- Accuracy of the posterior estimates can be assessed by the Monte Carlo standard error for each parameter (i.e. sd of difference between the mean of the sampled values of the parameter and the true posterior mean):
  - Posterior mean estimated by sample mean $\mathbb{E}(\theta) \approx \frac{1}{n}\sum \theta^{(i)}$
  - If samples were generated independently, could estimate SE of the mean as $S/\sqrt{n}$ where $S^2 = \frac{1}{n-1}\sum(\theta^{(i)} - \overline{\theta})^2$ is the sample variance
  - But, this will underestimate the true MC standard error due to autocorrelation in the samples generated using MCMC
  - Various remedies to obtain better estimate of MC error
    - ∗ WinBUGS uses a 'batch means' method — replaces sample variance $S^2$ by variance of batched means, which are assumed independent
    - ∗ Alternatively, replace posterior sample size $n$ by 'effective sample size' $n/\delta$ where $\delta = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ is the autocorrelation time and $\rho(k)$ is the lag $k$ autocorrelation in the sample of $\theta^{(i)}$'s

# 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

$\Rightarrow$ can address wide range of substantive questions by appropriate summaries of the posterior

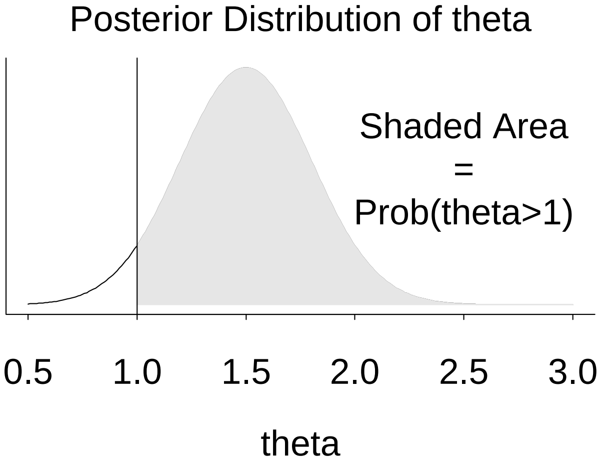
- Typically report either mean or median of the posterior samples for each parameter of interest as a point estimate
- 2.5% and 97.5% percentiles of the posterior samples for each parameter give a 95% posterior credible interval (interval within which the parameter lies with probability 0.95)

## Probability statements about parameters

- Already noted that classical inference cannot provide probability statements about parameters
- In contrast, 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 the 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 $g(\theta)$
  - not always possible, e.g. attributable risk = function of RR and prob. of exposure

- Easy using MCMC: just calculate required function $g(\theta)$ as a logical node at each iteration and summarise posterior samples of $g(\theta)$

⇒ in `WinBUGS`, include variables representing required functions as extra terms in the model code, and set sample monitors on these functions

## *Example: Surgical — Intervals on Ranks*

- Recent trend in UK towards ranking 'institutional' performance e.g. schools, hospitals

- 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, e.g.
  - Surgical mortality rates in 12 hospitals carrying out cardiac surgery
  - Fit conjugate beta prior × binomial likelihood independently to each hospital to estimate posterior distribution of mortality rate, $\pi_i$
  - Using MCMC, rank sampled values of $(\pi_1, ...., \pi_{12})$ at each iteration → sample from posterior distribution of ranks for each hospital

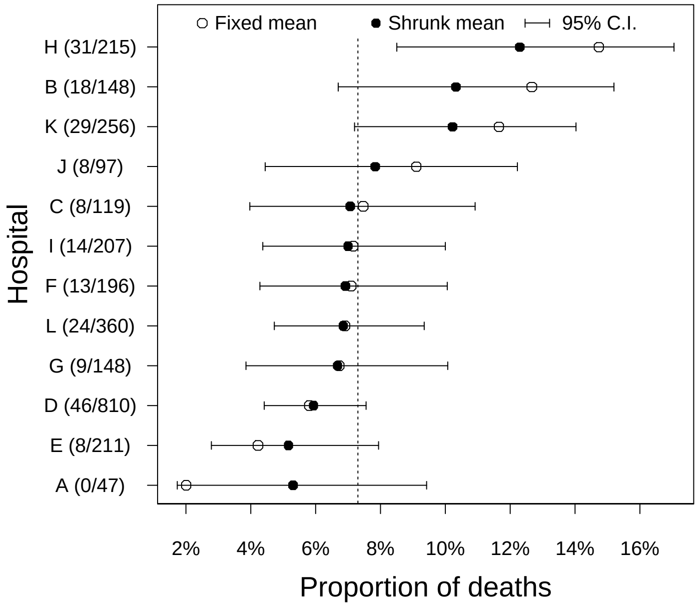
`WinBUGS` contains 'built-in' options for ranks:

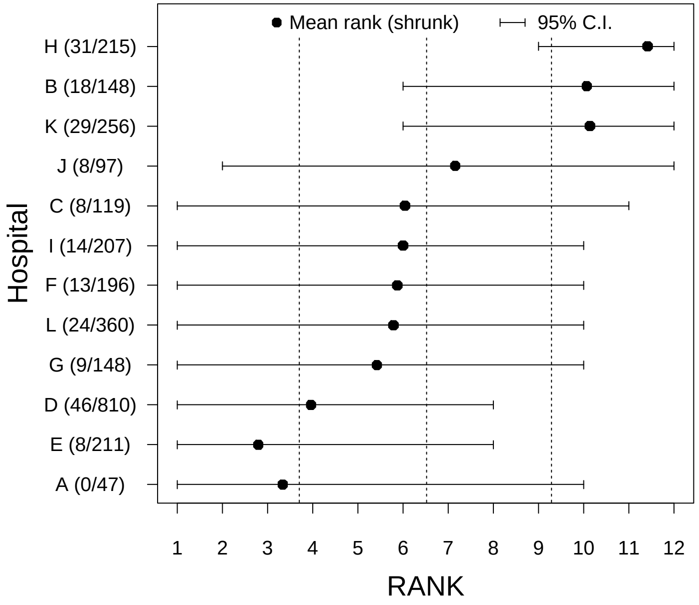
- `Rank` option of `Inference` menu monitors the rank of the elements of a specified vector

- `rank(x[], i)` returns the rank of the $i^{th}$ element of `x`

- `ranked(x[], i)` returns the value of the $i^{th}$-ranked element of `x`

# Example 1: Neonatal cardiac surgery mortality in 12 European hospitals





# The BUGS program

**Bayesian inference Using Gibbs Sampling**

- Language for specifying complex Bayesian models
- Constructs object-oriented internal representation of the model graph by identifying parents and children
- Builds up an arbitrarily complex model through specification of local structure
- Simulation from full conditionals using Gibbs sampling
- Current version (WinBUGS 1.4.3) runs in Windows, and incorporates the DoodleBUGS graphical model editor and a script language for running in batch mode
- Mac: JAGS are available to run under a Windows emulator, called from R

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

- An open source version of BUGS (called OpenBUGS) is under development, and includes versions of BUGS that run under LINUX (LinBUGS) and that can be run directly from R (BRugs). See `http://www.rni.helsinki.fi/openbugs`

# Running WinBUGS

1. Open *Specification tool* and *Update* from *Model* menu, and *Samples* from *Inference* 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. Click on *Update* to burn in.

9. Type nodes to be monitored into *Sample Monitor*, and click *set* after each.

10. Perform more updates.

11. Type * into *Sample Monitor*, and click *stats* etc to see results on all monitored nodes.

## Some useful facilities in WinBUGS

- `Info/Node info` gives current value of parameters
- `Inference/Correlations` gives posterior correlations
- `Inference/Compare` provides useful comparative plots, e.g.
  - Caterpillar or Box plot of an array: right-click, `Properties/Special` allows ranking etc
  - Model fit of a series: put fitted values (means) into `node`, raw data in `other`, independent predictor in `axis`

## 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)
  }
```

- `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 - 0.7)` $=$ 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, 0.7)` $=$ 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 $p_{ik} = \sum_j \pi_j \Lambda_{ijk}$.

- See 'Model Specification/Logical nodes' in manual for full syntax.

## Some common Distributions

| Expression | Distribution | Usage |
|---|---|---|
| `dbin` | binomial | `r ~ dbin(p,n)` |
| `dnorm` | normal | `x ~ dnorm(mu,tau)` |
| `dpois` | Poisson | `r ~ dpois(lambda)` |
| `dunif` | uniform | `x ~ dunif(a,b)` |
| `dgamma` | gamma | `x ~ dgamma(a,b)` |

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

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

## The WinBUGS data formats

WinBUGS accepts data files in:

1. Rectangular format

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

2. R 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.

## 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`
- Andrew Gelman's `bugs` function for R is most developed - reads in data, writes script, monitors output etc.
- OpenBUGS site http://mathstat.helsinki.fi/openbugs/ provides an open source version, including BRugs which works from within R

## 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)

# 3. Bayesian non-hierarchical 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 $x_{1i}, ..., x_{pi}$, $i = 1, ..., n$

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

An equivalent Bayesian formulation would typically specify

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \\ (\beta_0, \beta_1, ..., \beta_p, \sigma^2) &\sim \text{Prior distribution} \end{aligned}$$

## 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

Invariance arguments can suggest suitable scale on which to be ‘uniform’

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)

Recall our linear regression example:

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \end{aligned}$$

If we wished to assume ‘non-informative’ priors for the regression coefficients and residual variance, we might specify

$$\begin{aligned} \beta_k &\sim \text{Uniform}(-\infty, \infty), \qquad k = 0, ..., p \\ \log \sigma &\sim \text{Uniform}(-\infty, \infty) \end{aligned}$$

These priors are improper. Alternative ’just proper’ priors specification could be

$$\begin{aligned} \beta_k &\sim \text{Normal}(0, 100000), \qquad k = 0, ..., p \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

## Comparison of classical and Bayesian estimates

Let $\boldsymbol{Y}$ denote the $n \times 1$ vector of responses and $\boldsymbol{X}$ denote the $n \times (1 + p)$ matrix of covariates.

*Classical estimate*

MLE of $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ is

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ \text{SE}(\widehat{\boldsymbol{\beta}}) &= \sqrt{s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}} \end{aligned}$$

where $s^2 = \widehat{\sigma}^2 = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/(n-p)$

*Bayesian approach*

Assuming uniform prior on $(\beta_0, \beta_1, ..., \beta_p, \log\sigma^2)$, *conditional* posterior for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}, \boldsymbol{X}) = \text{MVN}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},\ \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

Integrating out $\sigma^2$ gives the *marginal* posterior

$$p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) = \text{MV-t}_{n-p}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},\ s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

Comments

- Uniform priors give closed-form posterior for linear model
- In general, we may not want to be restricted to uniform or conditionally conjugate priors for sampling variance and regression coefficients (although still convenient in many cases)
- May want to ’robustify’ model by replacing Normal errors with heavy-tailed distribution (e.g. t, double exponential)
- May wish to fit non-linear and generalised linear regressions

## Example: Linear regression of Stack Loss data

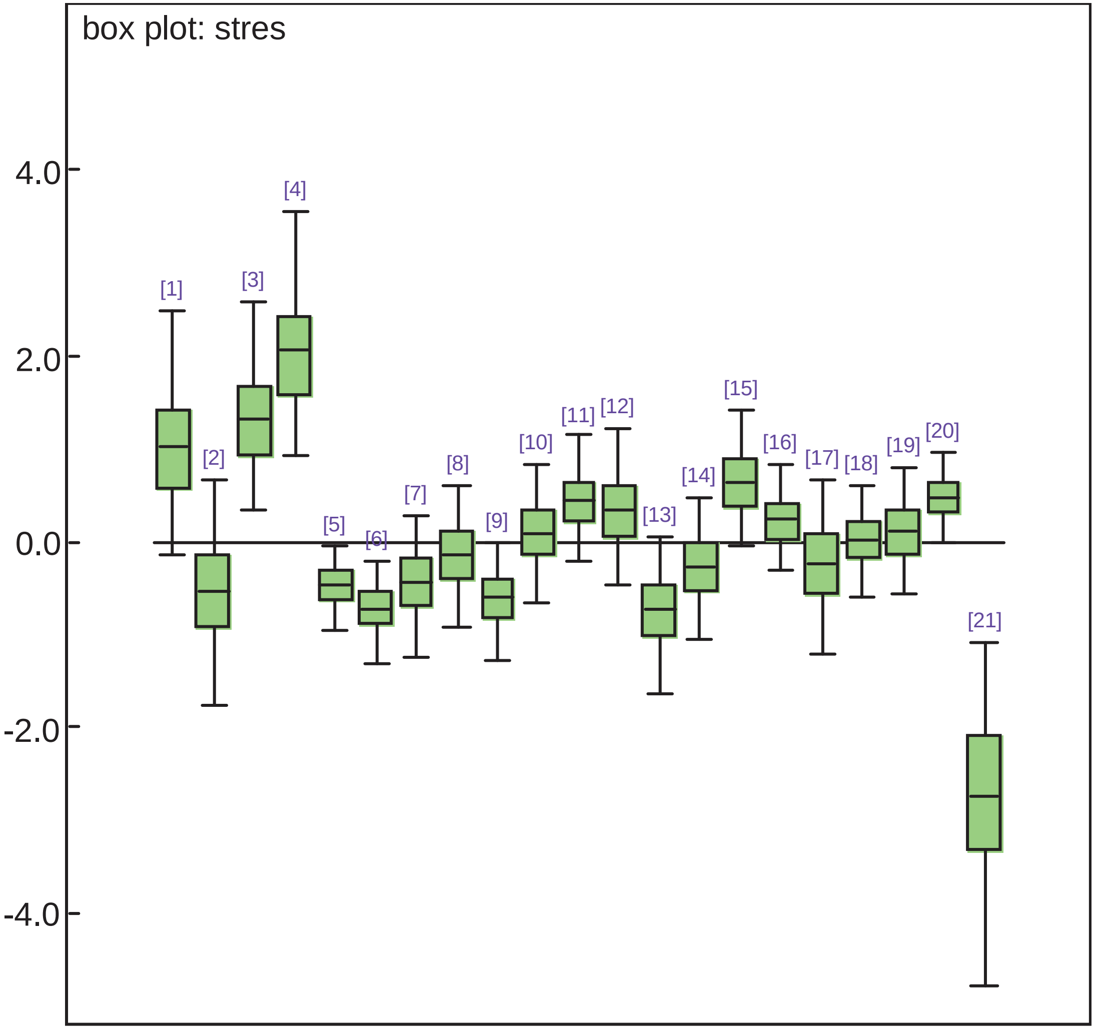
- 21 daily responses of stack loss, $y_i$, the amount of ammonia escaping from industrial chimneys
- Covariates: air flow $x_1$, temperature $x_2$ and acid concentration $x_3$
- Transformed covariates: $z_{ki} = (x_{ki} - \overline{x}_k)/\text{sd}(x_k), \quad k = 1, 2, 3$
- Model specification:

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \quad i = 1, ..., 21 \\ \mu_i &= \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \\ \beta_k &\sim \text{Normal}(0, 10000) \quad k = 0, ..., 3 \end{aligned}$$

- Results:

| | posterior mean | 95% interval |
|---|---:|---|
| $\beta_1$ | 6.55 | (3.91, 9.08) |
| $\beta_2$ | 4.09 | (1.71, 6.60) |
| $\beta_3$ | –0.81 | (–2.57, 0.94) |

But, examination of standardised residuals indicates some evidence of outliers



Robustify model by replacing Normal likelihood by t-4 likelihood

- Model specification:

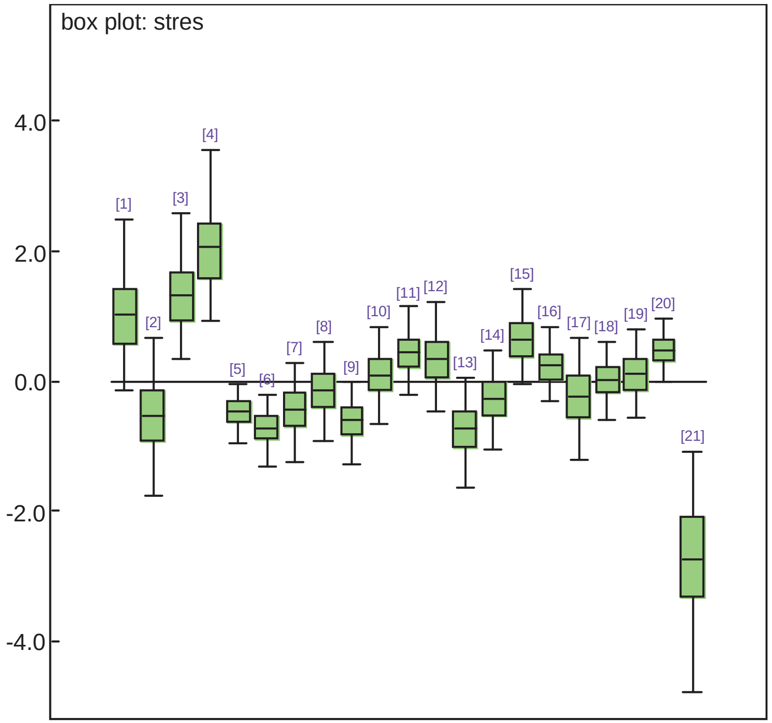
$$\begin{aligned} y_i &\sim \text{t}_4(\mu_i, \sigma^2) \quad i = 1, ..., 21 \\ \mu_i &= \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 z_{3i} \\ 1/\sigma^2 &\sim \text{Gamma}(0.001, 0.001) \\ \beta_k &\sim \text{Normal}(0, 10000) \quad k = 0, ..., 3 \end{aligned}$$

- Results:

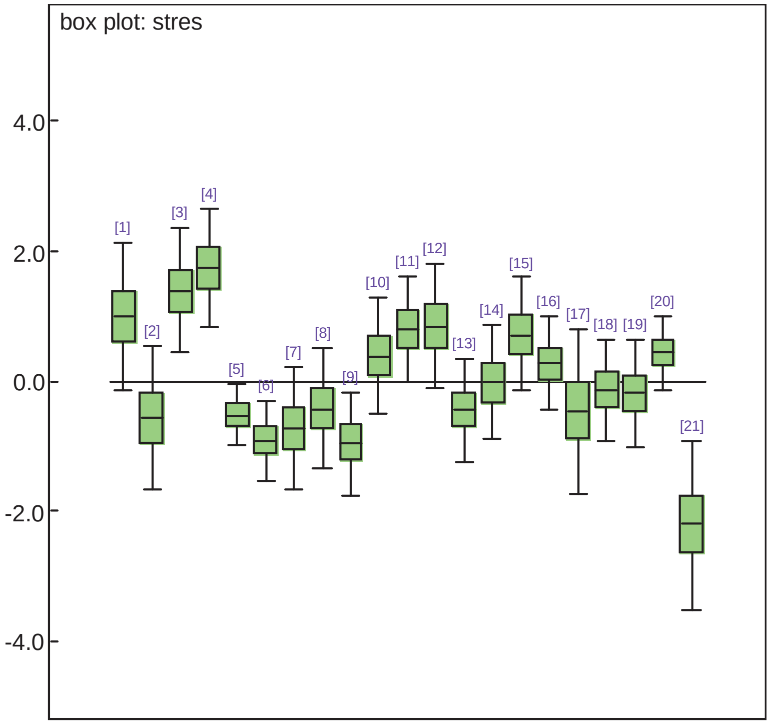
| | posterior mean | 95% interval |
|---|---:|:---:|
| $\beta_1$ | 7.65 | (4.92, 10.16) |
| $\beta_2$ | 2.74 | (0.59, 5.39) |
| $\beta_3$ | −0.68 | (−2,12, 0.69) |

# Standardised residuals

## Normal errors

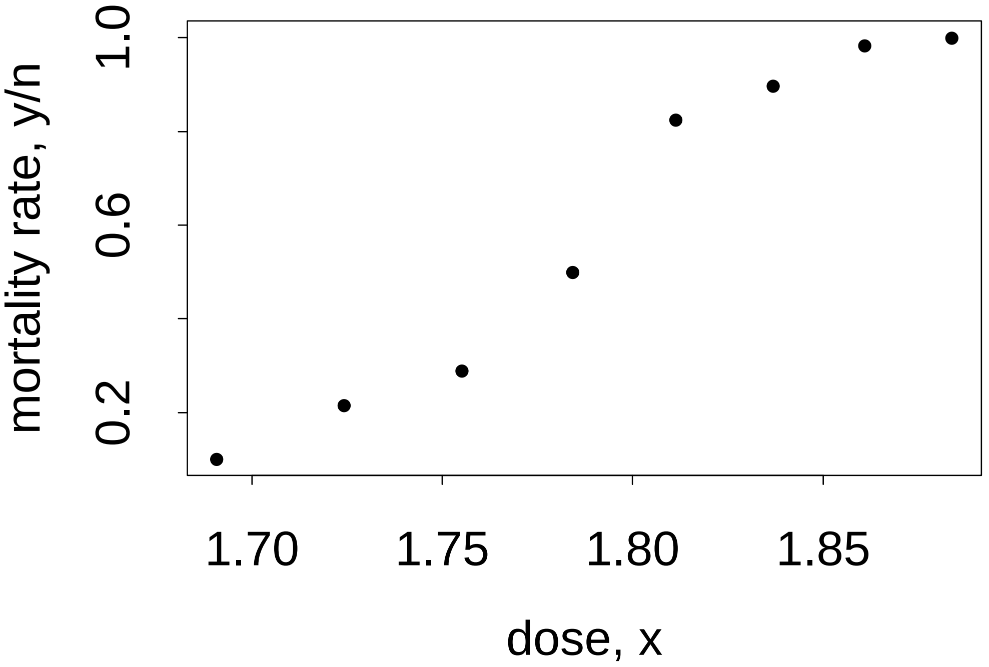


## $t_4$ errors



## Example: Generalised Linear regression of Beetles data

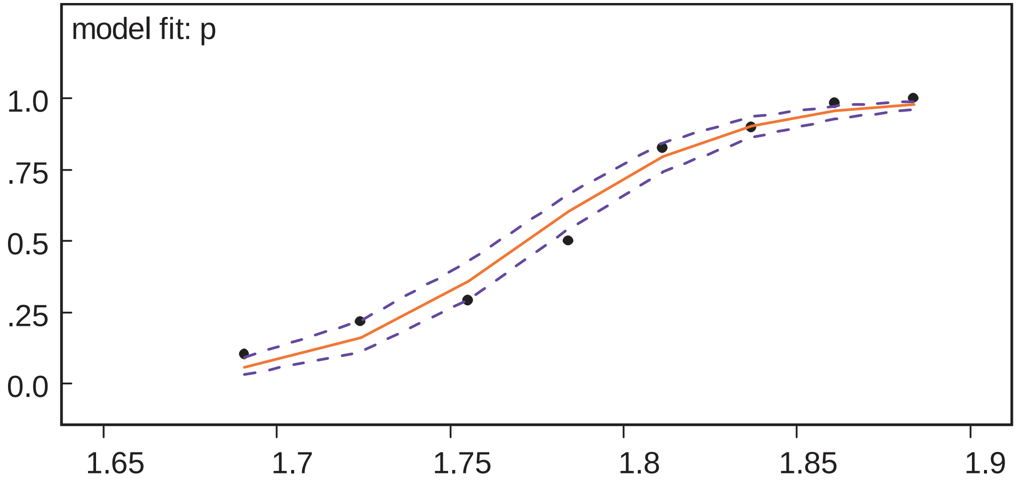
Dobson (1983) analyses binary dose-response data from a bioassay experiment in which the numbers of beetles killed after 5 hour exposure to carbon disulphide at N=8 different concentrations are recorded.



We start by fitting a logistic regression model

$$\begin{aligned} y_i &\sim \text{Binomial}(p_i, n_i) \\ \text{logit} p_i &= \alpha + \beta(x_i - \overline{x}) \\ \alpha &\sim \text{Normal}(0, 10000) \\ \beta &\sim \text{Normal}(0, 10000) \end{aligned}$$

**Beetles: logistic regression model fit (red = posterior mean of $p_i$; blue = 95% interval; black dots = observed rate $y_i/n_i$)**

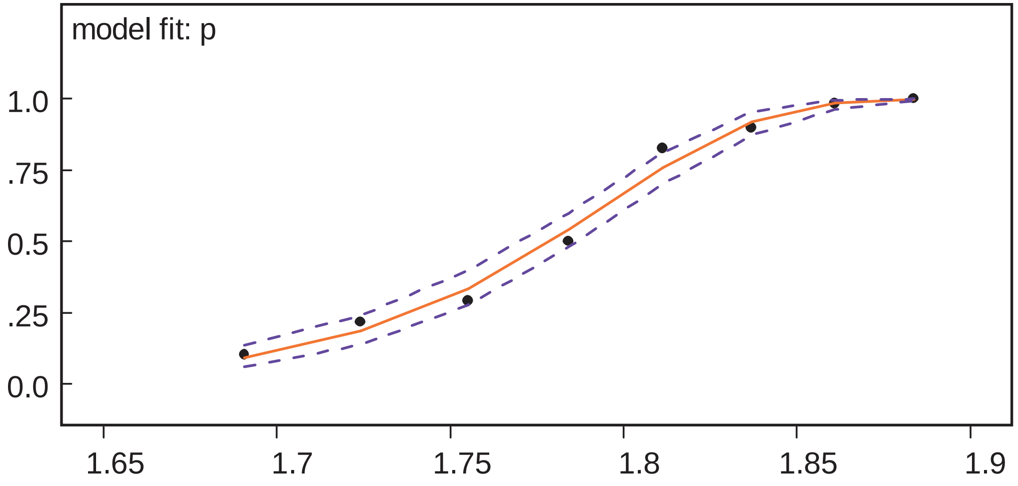


| dose level $i$ | obs. rate $y_i/n_i$ | posterior mean of $p_i$ | 95% interval |
|---|---|---|---|
| 1 | 0.10 | 0.06 | (0.03, 0.09) |
| 2 | 0.22 | 0.16 | (0.11, 0.22) |
| 3 | 0.29 | 0.36 | (0.29, 0.43) |
| 4 | 0.50 | 0.61 | (0.54, 0.67) |
| 5 | 0.83 | 0.80 | (0.74, 0.85) |
| 6 | 0.90 | 0.90 | (0.86, 0.94) |
| 7 | 0.98 | 0.96 | (0.93, 0.97) |
| 8 | 1.00 | 0.98 | (0.96, 0.99) |

Some evidence of lack of fit at extremes, so try alternative complementary log-log link function

$$\begin{aligned} y_i &\sim \text{Binomial}(p_i, n_i) \\ \text{cloglog} p_i &= \alpha + \beta(x_i - \overline{x}) \\ \alpha &\sim \text{Normal}(0, 10000) \\ \beta &\sim \text{Normal}(0, 10000) \end{aligned}$$

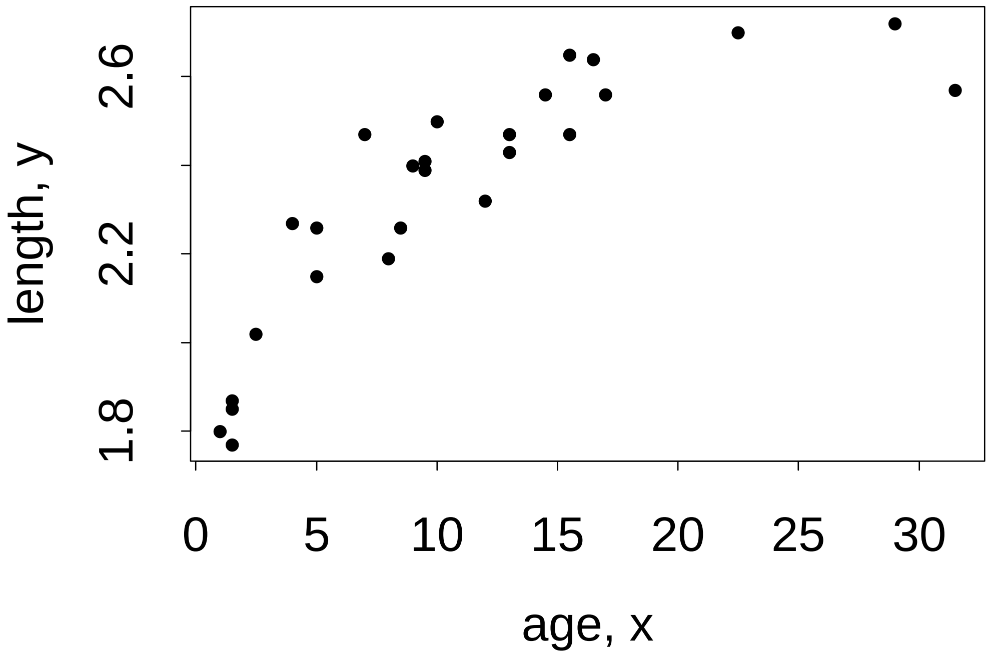
**Beetles: cloglog regression model fit (red = posterior mean of $p_i$; blue = 95% interval; black dots = observed rate $y_i/n_i$)**



| dose<br>level $i$ | obs. rate<br>$y_i/n_i$ | posterior<br>mean of $p_i$ | 95%<br>interval |
|---|---|---|---|
| 1 | 0.10 | 0.09 | (0.06, 0.14) |
| 2 | 0.22 | 0.19 | (0.14, 0.24) |
| 3 | 0.29 | 0.34 | (0.28, 0.40) |
| 4 | 0.50 | 0.54 | (0.48, 0.60) |
| 5 | 0.83 | 0.76 | (0.70, 0.81) |
| 6 | 0.90 | 0.92 | (0.87, 0.95) |
| 7 | 0.98 | 0.98 | (0.96, 0.99) |
| 8 | 1.00 | 1.00 | (0.99, 1.00) |

## Example: Non linear regression of Dugongs data

Carlin and Gelfand (1991) consider data on length ($y_i$) and age ($x_i$) measurements for 27dugongs (sea cows) captured off the coast of Queensland



A frequently used nonlinear growth curve with no inflection point and an asymptote as $x_i$ tends to infinity is

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \alpha - \beta\gamma^{x_i} \end{aligned}$$

where $\alpha, \beta > 0$ and $\gamma \in (0, 1)$

Vague prior distributions with suitable constraints may be specified as e.g.

$$\begin{aligned} \alpha &\sim \text{Uniform}(0, 100) \\ \beta &\sim \text{Uniform}(0, 100) \\ \gamma &\sim \text{Uniform}(0, 1) \end{aligned}$$

Alternatively, vague Normal priors with appropriate bounds could be specified for $\alpha$ and $\beta$, e.g.

$$\begin{aligned} \alpha &\sim \text{Normal}(0, 10000)I(0, ) \\ \beta &\sim \text{Uniform}(0, 10000)I(0, ) \end{aligned}$$

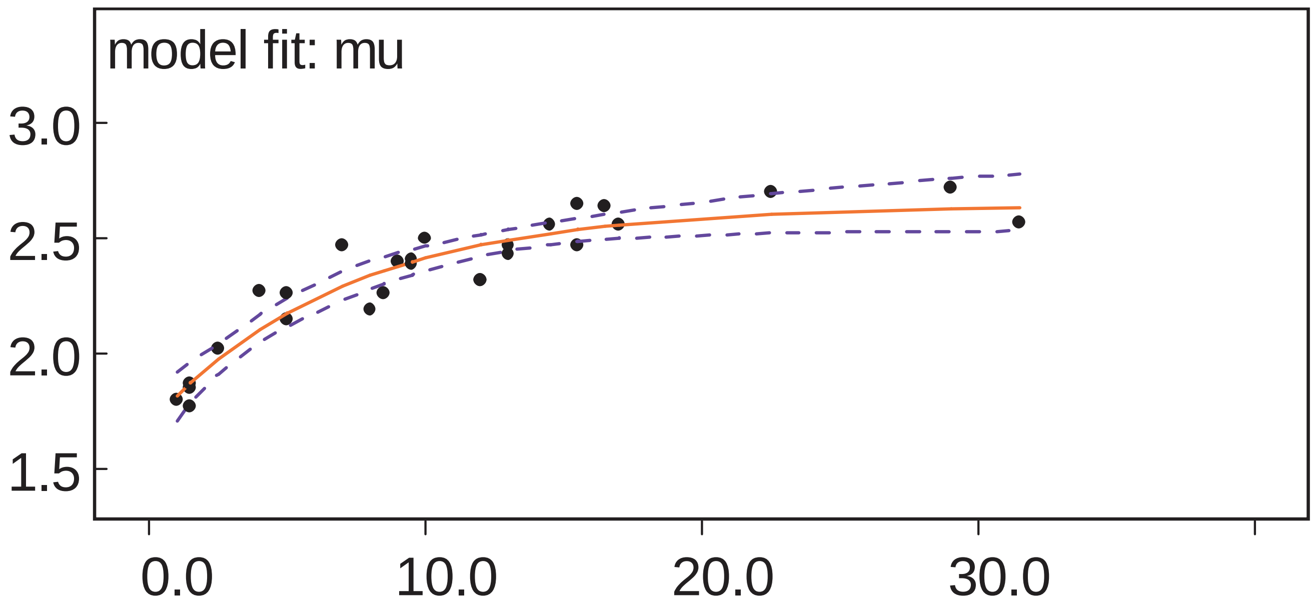
For the sampling variance, could specify uniform prior log variance or log sd scale

$$\log \sigma \sim \text{Uniform}(-10, 10)$$

or gamma prior on precision scale

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

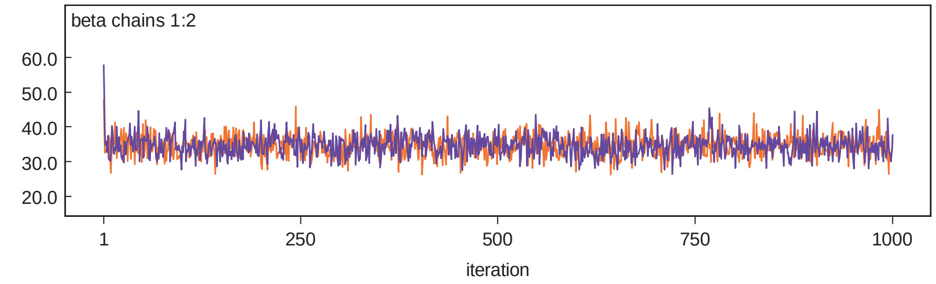
## Dugongs: model fit (red = posterior mean of $\mu_i$; blue = 95% interval)



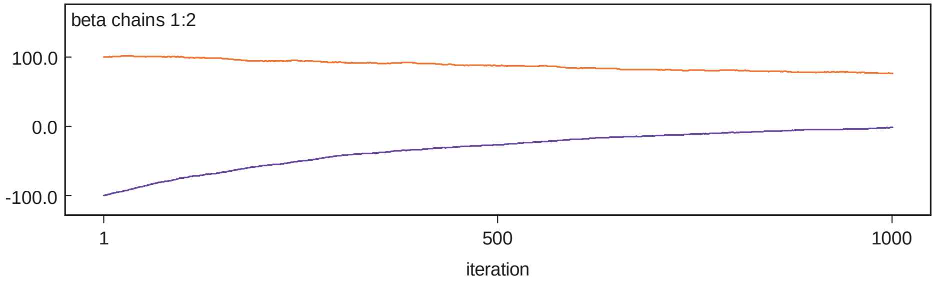
Re-parameterisation can often improve convergence

Recall Beetles example from previous lecture
→ Note the importance of centering the covariate (dose) in this example to reduce correlations between the parameters

History plot for slope, $\beta$: Centred covariate



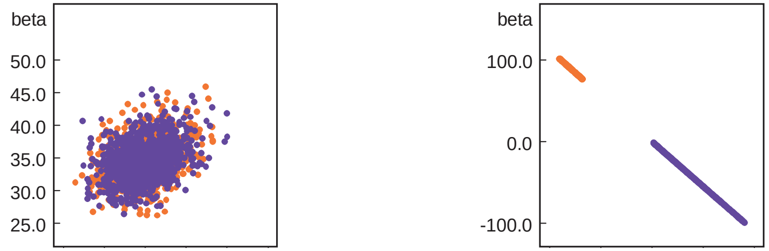
History plot for slope, $\beta$: Uncentred covariate



Bivariate scatter plot showing correlation between sampled values of $\alpha$ and $\beta$

Centered covariate

Uncentred covariate



# Model checking for non-hierarchical models

'Standard' checks based on fitted model, such as

- *residuals:* plot versus covariates, checks for auto-correlations and so on
- *prediction:* check accuracy on external validation set, or cross validation
- etc...

All this applies in Bayesian modelling, but in addition:

- parameters have distributions and so residuals are variables
- should check for conflict between prior and data
- should check for unintended sensitivity to the prior
- using MCMC, have ability to generate replicate parameters and data, so predictive checks and approximations to cross validation are easy

# Residuals in non-hierarchical models

- Standardised Pearson residuals $(y-\mu)/\sigma$ where $\mu = \mathrm{E}[Y|\theta]$, $\sigma^2 = \mathrm{V}[Y|\theta]$, $\theta$ - parameters

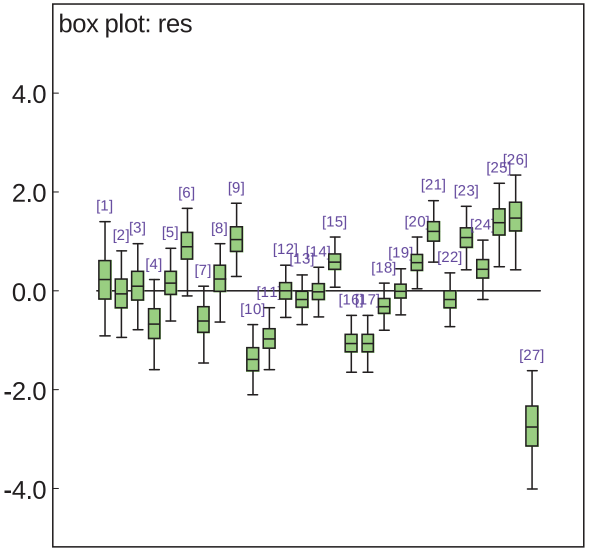
- In Bayesian analysis these are random quantities, with distributions

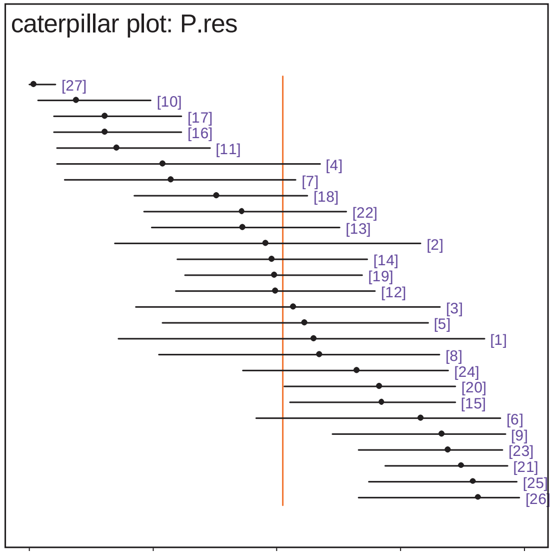
- If assuming Normality, then

$$P(Y) = \Phi[(Y-\mu)/\sigma]$$

has a Uniform[0,1] distribution under true $\mu$ and $\sigma$

## Dugongs: residuals





## Deviance residuals

Need to first define deviance as

$$D(\theta) = -2\log p(y|\theta)$$

Define saturated deviance as

$$D_S(\theta) = -2\log p(y|\theta) + 2\log p(y|\widehat{\theta}(y))$$

where $\widehat{\theta}(y)$ are appropriate 'saturated' estimates: typically, when $\boldsymbol{E}[Y] = \theta$, we set $\widehat{\theta}(y) = y$ (McCullagh and Nelder).

Standardised deviances for common distributions are:

$$\begin{array}{lll} y_i \sim \text{Binomial}(\theta_i, n_i): & D_S(\theta) = & 2\left\{\sum_i y_i \log\left[\frac{y_i/n_i}{\theta_i}\right] + (n_i - y_i)\log\left[\frac{(1-y_i/n_i)}{1-\theta_i}\right]\right\} \\ y_i \sim \text{Poisson}(\theta_i): & D_S(\theta) = & 2\left\{\sum_i y_i \log\left[\frac{y_i}{\theta_i}\right] - (y_i - \theta_i)\right\} \\ y_i \sim \text{Normal}(\theta_i, \sigma_i^2): & D_S(\theta) = & \sum_i \left[\frac{y_i-\theta_i}{\sigma_i}\right]^2 \end{array}$$

Deviance residuals

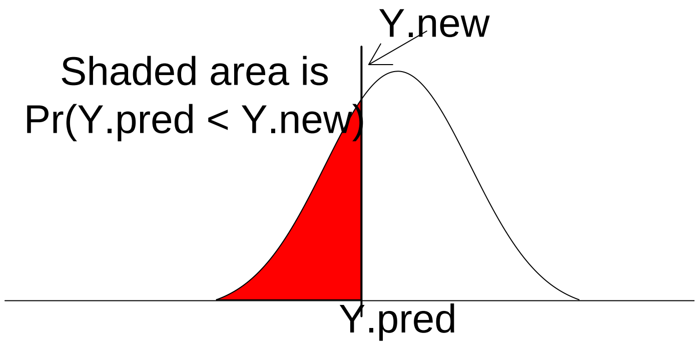
$$dr_i \quad = \quad sign_i\sqrt{D_{Si}}$$

where $sign_i$ is the sign of $y_i - \mathbb{E}(y_i|\text{parents}(y_i))$.

# Prediction

Ideally would like to test fit on new data $Y^{\text{new}}$ (prediction)

- For Normal model, could then calculate expectation of $P$-value $\Phi[(Y^{\text{new}} - \theta)/\sigma]$ under current posterior distribution, and see whether extreme
- Equivalently, could predict new observation $Y^{\text{pred}}$, and estimate $P$-value by posterior probability that $Y^{\text{pred}} < Y^{\text{new}}$



- Second approach more general
  - Each can be approximated by just using current data instead of new data (conservative)
  - See Gelman et al (2004) for further details

# Bayesian model comparison using DIC

- Natural way to compare models is to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model
- Spiegelhalter et al (2002) proposed a Bayesian model comparison criterion based on this principle:

  Deviance Information Criterion, DIC = 'goodness of fit' + 'complexity'
- They measure fit via the deviance

$$D(\theta) = -2 \log L(\text{data}|\theta)$$

- Complexity measured by estimate of the 'effective number of parameters':

$$\begin{aligned} p_D &= E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \\ &= \overline{D} - D(\bar{\theta}); \end{aligned}$$

  i.e. posterior mean deviance minus deviance evaluated at the posterior mean of the parameters
- The DIC is then defined analagously to AIC as

$$\text{DIC} = D(\bar{\theta}) + 2p_D = \overline{D} + p_D$$

  Models with smaller DIC are better supported by the data

## Running from ‘scripts’

Once a program is working it is more convenient to use ’scripts’ to carry out a simulation in the background.

```
# Script for running analysis display(‘log’)
check(‘c:/winbugs/stacks-mod’)    # check syntax of model
data(‘c:/winbugs/stacks-dat’)     # load data file
compile(2)                        # generate code for 2 simulations
inits(1,‘c:/winbugs/stacks-in1’)  # load initial values 1
inits(2,‘c:/winbugs/stacks-in2’)  # load initial values 2
set(beta)                         # monitor regression coeffs.
set(mu)                           # monitor fitted values
set(stres)                        # monitor residuals
update(11000)                     # perform 11000 simulations
history(beta)                     # trace plot of samples for beta
gr(beta)                          # Gelman-Rubin diagnostic for convergence
beg(1001)                         # Discard first 1000 iterations as burn-in
stats(*)         # Calculate summary statistics for all monitored quantities
density(beta)    # Plot marginal posterior distribution of each beta
```

# 4. **Bayesian hierarchcial models**

# Graphical Models

## Model building

Statistical modelling of complex systems involve usually many interconnected random variables.

How to build the connections ?

**Key idea: conditional independence**

It is helpful to represent the modelling process by a graph

- nodes: all random quantities
- links (directed or undirected): association between the nodes

Directed edges: natural ordering of association, “causal” influence

Undirected edges: symmetric association, correlation

The graph is used to represent a set of *conditional independence* statements

## Independence and Conditional independence

Two variables, $X$ and $Y$, are *statistically independent* if

$$p(X,Y) \;=\; p(X)\,p(Y).$$

Equivalently, variables $X$ and $Y$ are statistically independent if

$$p(Y \mid X) \;=\; p(Y).$$

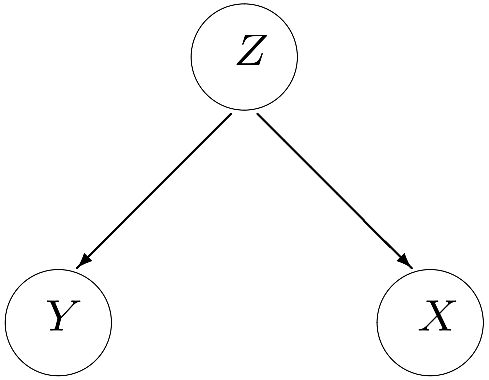
*Conditional independence:*

Given 3 variables $X, Y$ and $Z$, we say that $X$ and $Y$ are conditionally independent given $Z$, denoted by $X \perp\!\!\!\perp Y \mid Z$,

if

$$p(X,Y \mid Z) \;=\; p(X \mid Z)p(Y \mid Z)$$

We can draw this relationship in a *graph*:



**Genetic Example:**

Consider a family with 2 parents and 2 children.

Let $X$ and $Y$ denote the genotype of the 2 children and $Z$ the genotype of the parents.

If we know the genotypes of the parents, the genotypes of the children are conditionally independent: $X \perp\!\!\!\perp Y \mid Z$

$$p(X, Y \mid Z) \;=\; p(X \mid Z)p(Y \mid Z)$$

But, if we have no information on the parents, the genotypes of the children are marginally (unconditionally) dependent .

## Directed Acyclic Graphs, (DAG)

- Only contain directed edges
- No directed cycles allowed
  $\Rightarrow$ each node has a well defined set of parents: parents[$v$] and descendants
- Used to build models directionally, e.g. disease → symptoms, parameters → data, cause → effect

DAG: set of nodes $V = \{v\} +$ a set of directed edges

The joint distribution associated with the DAG is then specified by:

$$p(V) \;=\; \prod_{v \epsilon V} p(v \,|\, \text{parents}[v])$$

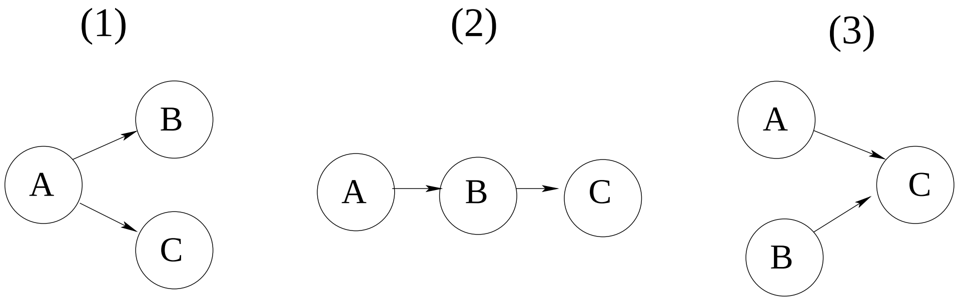
This is a recursive factorisation that will be extensively used in computations for hierarchical models

Some basic local structures:

(1) $A$ is parent of $B$ and $C$: Joint distribution is $p(A)p(B|A)P(C|A)$

(2) Markov chain: path from $A$ to $C$: Joint distribution is $p(A)P(C|B)p(B|A)$

(3) $A$ and $B$ are marginally independent but dependent after observing $C$:
Joint distribution is $p(A)p(B)P(C|A,B)$



## Summary

- Using a DAG is an easy and interpretable way of specifying joint distributions through simple local terms

  We shall see that most hierarchical models can be built using DAGs

- Further CI properties can be deduced by moralising the graph

- The graph is used for simplifying computations

## Further reading

Spiegelhalter (1998) (Tutorial on Bayesian graphical models)

Spiegelhalter et al (1995) (Discussion of link between graphical models and Bayesian computation)

Richardson and Best (2003) (Use of Bayesian graphical models to build complex models in environmental epidemiology)

# Hierarchical Models

Many statistical problems involve multiple parameters

**Why ?**

It is necessary to reflect the complexity of observables and different patterns of heterogeneity, dependence, mismeasurements ...

Some examples:

- "treatment effect" for different categories of patient (age, stage of disease, prior treatment,...)
- "study effect" in meta-analyses
- relative risks for a disease outcome in different areas/age/time periods
- "subject effect" in growth curves models
- "frailty effect" in correlated or familial survival studies
- .....

How to make inference on multiple parameters $\{\theta_1, \ldots \theta_I\}$ measured on $I$ units (persons, centres, areas, ... ) *which are related or connected by the structure of the problem ?*

We can identify three different assumptions:

1. **Identical parameters:** All the $\theta$'s are identical, in which case all the data can be pooled and the individual units ignored.

2. **Independent parameters:** All the $\theta$'s are entirely unrelated, in which case the results from each unit can be analysed independently (for example using a fully specified prior distribution within each unit)

   → individual estimates of $\theta_i$ are likely to be highly variable (unless very large sample sizes)

3. **Exchangeable parameters:** The $\theta$'s are assumed to be 'similar' in the sense that the 'labels' convey no information

Under broad conditions an assumption of exchangeable units is mathematically equivalent to assuming the $\theta$'s are drawn at random from some population distribution, just as in a traditional random effects model

## Exchangeability

'Exchangeability' is a formal expression of the idea that we find no systematic reason to distinguish the individual random variables $X_1, ..., X_n$

→ A *judgement* that they are 'similar' but not identical

We judge that $X_1, ..., X_n$ are exchangeable if the probability that we assign to any set of potential outcomes, $p(x_1, ..., x_n)$, is unaffected by permutations of the labels attached to the variables

e.g. suppose $X_1, X_2, X_3$ are the first three tosses of a (possibly biased) coin, where $X_1 = 1$ indicates a head, and $X_1 = 0$ indicates a tail

We might judge $p(X_1 = 1, X_2 = 0, X_3 = 1) = p(X_2 = 1, X_1 = 0, X_3 = 1) = p(X_1 = 1, X_3 = 0, X_2 = 1)$: i.e. the probability of getting 2 heads and a tail is unaffected by the particular toss on which the tail comes

This is a natural judgement to make if we have no reason to think that one toss is systematically any different from another

Note that it does *not* mean we believe that $X_1, ..., X_n$ are independent: this would not allow us to learn about the chance of a head

## Representation theorem

de Finetti (1930) showed that if a set of binary variables $X_1, ..., X_n$ were judged exchangeable, then it implied that

$$p(x_1, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta)\ p(\theta)\ d\theta$$

Easy if argue from ‘right to left’

From ‘left to right’ is remarkable: exchangeable random quantities can be thought of as being *independently and identically distributed* drawn from some common *parametric distribution* depending on an unknown parameter $\theta$, which itself has a *prior distribution* $p(\theta)$

Thus, from a subjective judgment about observable quantities, one derives the whole apparatus of parametric models and Bayesian statistics!

## Link between exchangeability, representation thm and hierarchical models

Suppose $x_{ij}$ is outcome for individual $j$, unit $i$, with unit-specific parameter $\theta_i$

- Assumption of partial exchangeability of individuals within units can be represented by the following model:

$$\begin{aligned} x_{ij} &\sim p(x_{ij}|\theta_i) \\ \theta_i &\sim p(\theta_i) \end{aligned}$$

- Assumption of exchangeability of the units can be represented by the model:

$$\begin{aligned} \theta_i &\sim p(\theta_i|\phi) \\ \phi &\sim p(\phi) \end{aligned}$$

  – can be considered as a common prior for all units, but one with unknown parameters

Note that there does not need to be any actual sampling — perhaps these $I$ units are the only ones that exist — since the probability structure is a consequence of the belief in exchangeability rather than a physical randomisation mechanism

We emphasise that an assumption of exchangeability is a *judgement* based on our knowledge of the context.

Assuming $\theta_1, ...\theta_I$ are drawn from some common prior distribution whose parameters are unknown is known as a **hierarchical** or **multi-level** model

## General form of a Bayesian hierarchical model

Observables $x$,

Parameters $\theta = (\theta_1, \ldots, \theta_n)$

- likelihood $p(x|\theta)$ models the structure of observables (1st level)
- prior $p(\theta)$ is decomposed into conditional distributions expressing judgements about exchangeability:
  $p(\theta|\phi_2)$ (2nd level),
  $p(\phi_2|\phi_3)$ (3rd level),
  $\ldots$
  and a marginal distribution $p(\phi_m)$, such that

$$p(\theta) = \int p(\theta|\phi_2)p(\phi_2|\phi_3)\ldots p(\phi_{m-1}|\phi_m)p(\phi_m)d\phi_2 d\phi_3 \ldots d\phi_m$$

  $\phi_k$ are called the hyperparameters of level $k$

Can view hierarchical models as a way of simplifying specification of the joint prior on $\theta$

Provides a way of ‘estimating’ the prior distribution

# Hierarchical models and shrinkage

Suppose in each unit we observe a response $x_i$ assumed to have a Normal likelihood

$$x_i \sim \mathsf{N}(\theta_i, s_i^2)$$

Unit means $\theta_i$ are assumed to be exchangeable, and to have a Normal distribution

$$\theta_i \sim \mathsf{N}(\mu, \sigma^2)$$

where $\mu$ and $\sigma^2$ are 'hyper-parameters' for the moment assumed known.

After observing $x_i$, Bayes theorem gives (see lecture 1)

$$\theta_i | x_i \sim \mathsf{N}(w_i\mu + (1 - w_i)x_i, (1 - w_i)s_i^2)$$

where $w_i = s_i^2/(s_i^2 + \sigma^2)$ is the weight given to the prior mean

An exchangeable model therefore leads to the inferences for each unit having *narrower* intervals than if they are assumed independent, but *shrunk* towards the prior mean response

$w_i$ controls the 'shrinkage' of the estimate towards $\mu$, and the reduction in the width of the interval for $\theta_i$

Shrinkage ($w_i$) depends on precision of the individual unit $i$ relative to the variability between units

The unknown hyper-parameters $\mu$ and $\sigma$ may be

1. Estimated directly from the data, known as the ‘empirical Bayes’ approach (standard random-effects modelling — e.g. DerSimonian and Laird (1986) method of moments)

2. Given a prior distribution, known as the ‘full Bayes’ approach

Results often similar provided each unit is not too small and there are a reasonable number of units.

But, full Bayes approach correctly allows for uncertainty in hyper-parameters and provides much greater flexibility

## Comments

Hierarchical models allow “borrowing of strength” across units

- posterior distribution of $\theta_i$ for each unit borrows strength from the likelihood contributions for *all* the units, via their joint influence on the posterior estimates of the unknown hyper-parameters

→ improved efficiency

MCMC allows considerable flexibility over choice of random effects distribution (not restricted to normal random effects)

Judgements of exchangeability need careful assessment

- units suspected a priori to be systematically different might be modelled by including relevant covariates so that residual variability more plausibly reflects exchangeability
- subgroups of prior interest should be considered separately

## Example: Surgical — Hierarchical models for binary data

Suppose we observe $I$ sets of binomial data, e.g.

- $I$=12 Hospitals performing cardiac surgery

- Number of surgical failures (deaths) per centre

| | *Hospital* | | | | | | |
|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **....** | **J** | **K** | **L** |
| **No. of ops.** $n$ | 47 | 148 | 119 | .... | 97 | 256 | 360 |
| **No. of deaths** $r$ | 0 | 18 | 8 | .... | 8 | 29 | 24 |

How would you model these data?

Pooled approach: fit same beta-binomial model to all the hospitals:

*Likelihood:* $$\prod_{i=1}^{I} p(r_i|\pi, n_i) \quad = \quad \prod_{i=1}^{I} \text{Binomial}(n_i,\ \pi)$$

*Prior:* $$p(\pi) \quad = \quad \text{Beta}(a, b)$$
$(a,\ b$ known constants)

*Posterior:* $$\begin{aligned} p(\pi \mid \underline{r}, \underline{n}) \quad &\propto \quad \prod_{i=1}^{I} p(r_i \mid \pi, n_i) p(\pi) \\ &= \quad \text{Beta}(\sum_i r_i + a,\ \sum_i (n_i - r_i) + b) \end{aligned}$$

BUT is it reasonable to assume *common* probability $\pi$ of death for each hospital?

The beta-binomial model above assumes that each outcome (proportion of failures per hospital) is *independent and identically distributed* according to the binomial probability distribution with parameter $\pi$

- Does this model adequately describe the random variation in outcomes for each hospital?
- Are the hospital failure rates more variable than our model assumes?

## Some reasons for excess variation in response

- Individual heterogeneity *i.e.* systematic differences between units which are not attributable to random variation
  - this concept is often termed *frailty* in survival analysis
  - for binary/count data this is often termed *overdispersion*
- Repeated response measurements from the same unit tend to be *correlated*

  ⇒ 2 responses from the same unit will be more alike than 2 responses from different units

  ⇒ variation in responses is not completely random
- Failure to measure or include a relevant explanatory variable
- Inaccurate measurement of relevant explanatory variables

## Modelling the excess variation

We could modify our beta-binomial model to allow for a *different* failure probability, $\pi_i$ for each hospital $i$:

$$\begin{aligned} p(r_i \mid n_i, \pi_i) &= \text{Binomial}(n_i, \pi_i) \\ p(\pi_i) &= \text{Beta}(a, b) \end{aligned}$$

*Interpretation*

- If $a$ and $b$ are elicited separately for each hospital, or chosen to represent vague prior information, then equivalent to estimating $\pi$ *independently* for each hospital

- If $a$ and $b$ are treated as unknown and assigned a prior, then equivalent to viewing hospitals as exchangeable

  → assumes mortality rates are *similar* but not identical

  → Beta($a$, $b$) prior describes distribution of true mortality rates across hospitals

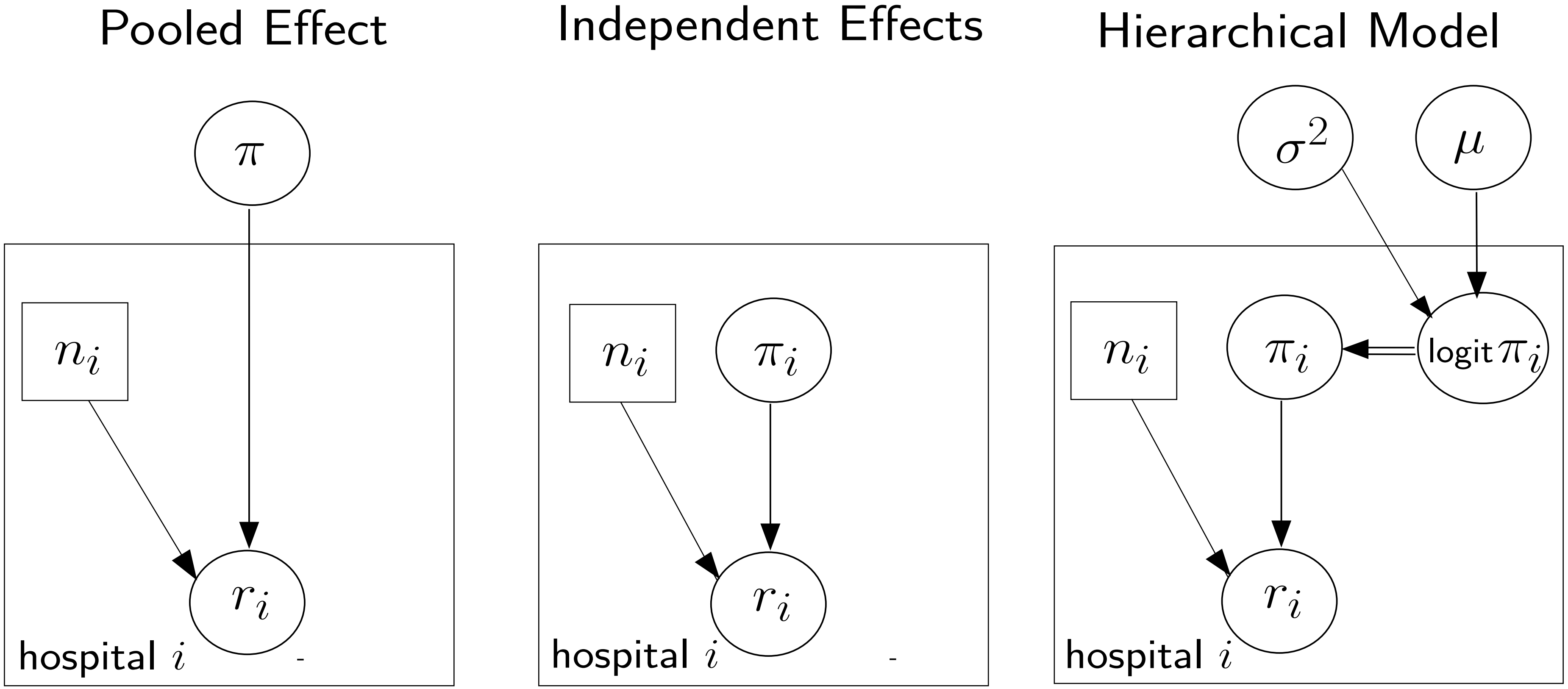
- What prior distributions would you specify for hyper-parameters $a$ and $b$?

- Beta prior is convenient choice in non-hierarchical models due to conjugacy, but no need to be restricted here

- More natural to work within generalised linear (mixed) model framework

| | | | |
|---|---|---|---|
| Likelihood (Level 1) | $r_i$ | $\sim$ | Binomial$(n_i, \pi_i)$ |
| Exchangeable rates (Level 2) | logit $\pi_i$ | $\sim$ | N$(\mu, \sigma^2)$ |
| Hyperpriors | $\mu$ | $\sim$ | Uniform or Normal |
| | $\sigma^2$ | $\sim$ | (See later) |

Graphical models (DAGs) for surgical example



Pooled Effect

Independent Effects

Hierarchical Model

Additional notation:

- Double edged arrow denotes logical (rather than stochastic) relationship

## Further reading

WinBUGS examples volumes I and II (lots of examples of Bayesian hierarchical models)

Congdon (2001) (lots of examples of Bayesian hierarchical models)

Gelman et al (2004) Chapters 5, 13, 14

# Priors for hyper-parameters in hierarchical models

Consider a hierarchical model with exchangeable random effects

$$\theta_i \quad \sim \quad \mathrm{N}(\mu, \sigma^2) \quad i = 1, ..., I$$

What priors should we assume for the hyper-parameters $\mu$ and $\sigma^2$?

Often want to be reasonably non-informative about the mean and variance of the random effects

- For location parameters (i.e. random effects mean, $\mu$), a uniform prior on a wide range, or a Normal prior with a large variance can be used
  - ! remember that WinBUGS parameterised 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 $\theta$
- 'Non-informative' priors for random effects variances are more tricky

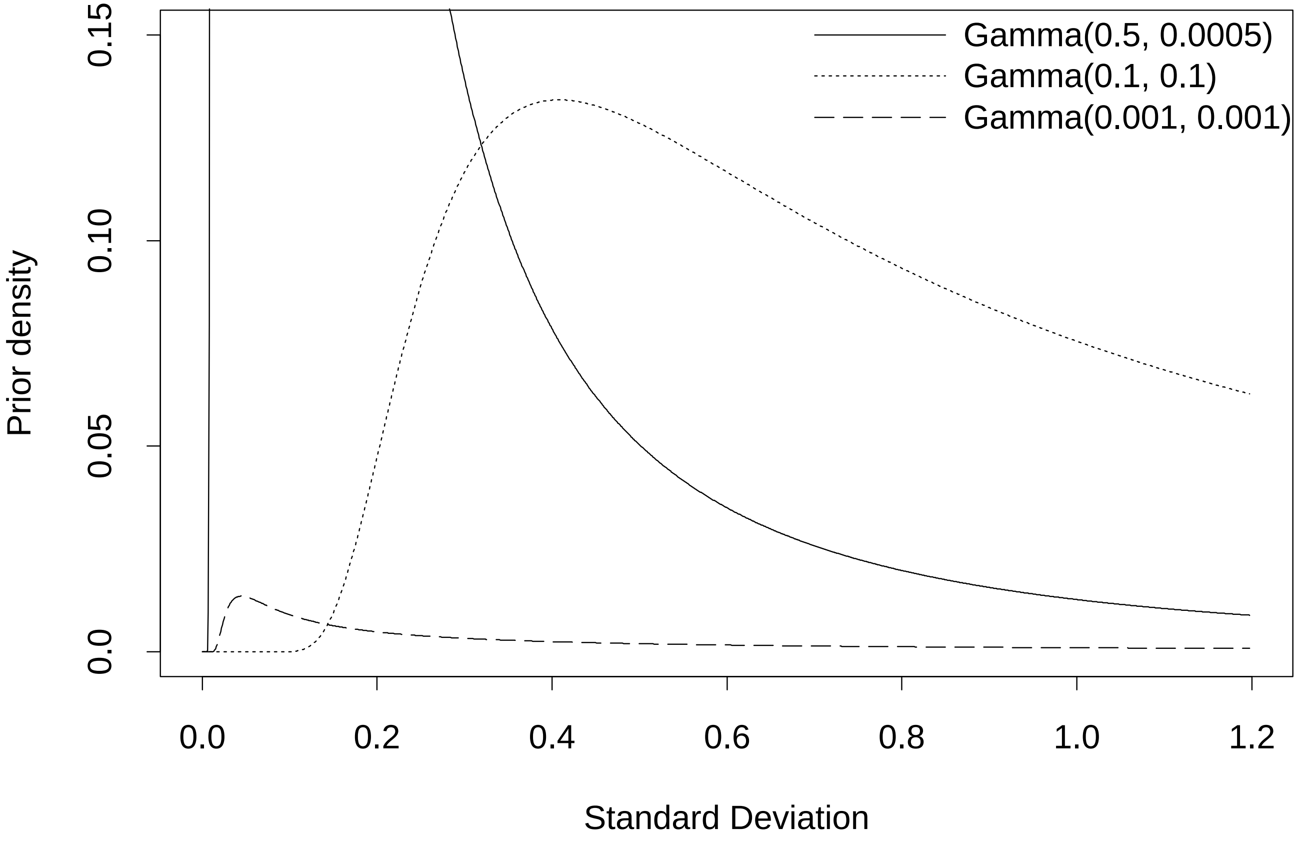
- Standard 'non-informative' (Jeffreys) prior for a Normal variance $\sigma^2$ is

$$\log \sigma^2 \sim \text{Uniform} \Rightarrow p(\sigma^2) \propto \sigma^{-2} \propto \text{Gamma}(0,0)$$

  - Note: Uniform prior on log $\sigma^2$ is equivalent to Uniform prior on $\log \sigma$ or $\log \sigma^{-2}$ (or indeed any power of $\sigma$)

- This prior is improper (doesn't integrate to 1)
  - OK as prior on sampling variance as still gives proper posterior
  - If used as prior for random effects variance → *improper* posterior
    - prior has infinite mass at zero, but $\sigma^2 = 0$ is supported by non-negligible likelihood

- Gamma($\epsilon$, $\epsilon$)— with small positive $\epsilon$ — is 'just proper' form of Jeffreys prior
  - A Gamma(0.001, 0.001) prior for the random effects *precision* is often used, as it also has nice conjugacy properties with the Normal distribution for the random effects
  - But inference may still be sensitive to choice of $\epsilon$
    - sensitivity particularly a problem if data (likelihood) supports small values of $\sigma^2$ (i.e. little evidence of heterogeneity between units)
    - See Gelman (2005) for further discussion



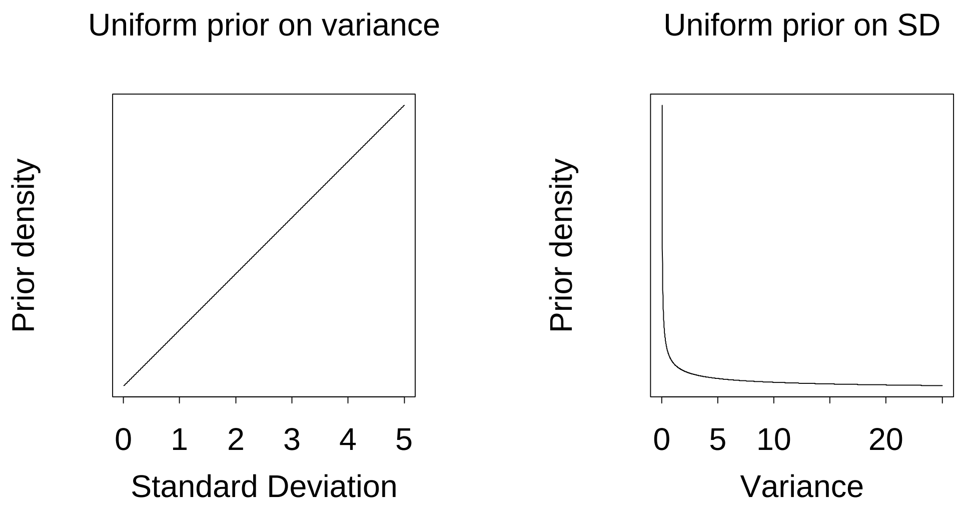
Some different gamma($a$, $b$) priors for the precision, shown on the scale of the standard deviation

Other options for ‘vague’ or ‘weakly informative’ priors on variance components:

- Uniform priors over a finite range on the variance or standard deviation, e.g.

$$\sigma^2 \;\sim\; \text{Uniform}(0,25); \quad \sigma \sim \text{Uniform}(0,5)$$

  Appropriate upper bound will depend on scale of measurement of random effects



- Half-normal or half-t on standard deviation, e.g.

$$\sigma \;\sim\; \text{Normal}(0,100)I(0,)$$

  Again, value chosen for variance of half-normal will depend on scale of measurements for continuous data

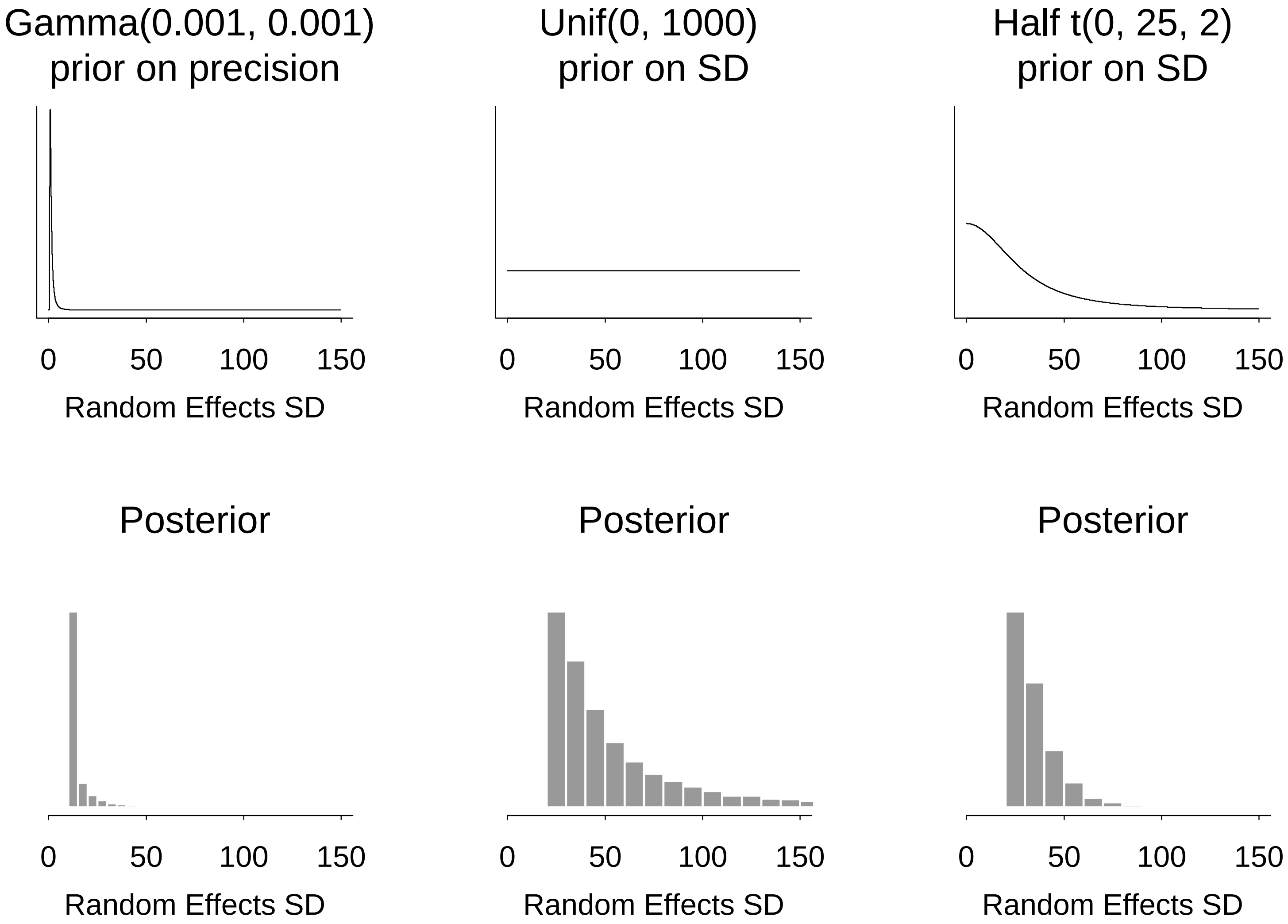
## Example: Comparing performance on Aptitude Test in schools

(Gelman 2005)

- Two-level normal hierarchical model with school-specific random effect
- Observed effects for 8 schools range from $-2.75$ (SE 16.3) to 28.4 (SE 14.9)
- Likelihood supports wide range of values for between-school SD

  → plausible that all schools have same underlying effect (i.e. between-school SD $= 0$)

  → data also support large between-school variation, although implausible that SD $> 50$
- Consider subset of first 3 schools (→ very sparse data) and compare Gamma(0.001, 0.001) prior on random effects precision with Uniform(0, 1000) prior and half-t(0, 25, 2) prior on random effects sd



Gamma(0.001, 0.001) prior on precision

Unif(0, 1000) prior on SD

Half t(0, 25, 2) prior on SD

Posterior

Posterior

Posterior

## Weakly informative prior based on plausible range of variation for log normal random effects

Suppose we assume a normal random effects distribution on the log odds ratio or log rate or log relative risk scale

- Assume a priori that log ORs or log rates or log RRs, $\theta_i$, are N$(\mu, \sigma^2)$

  $\Rightarrow$ 90% of values of $\theta$ lie in interval $\mu \pm 1.645\sigma$

  $\Rightarrow \quad \theta_{95\%} - \theta_{5\%} = 2 \times 1.645 \times \sigma$

  $\Rightarrow \quad \sigma = (\theta_{95\%} - \theta_{5\%})/3.29$

  $\Rightarrow \quad \sigma^{-2} = 3.29^2/\left(\theta_{95\%} - \theta_{5\%}\right)^2$

- Suppose we believe 3-fold variation between the ORs or rates or RRs for the upper and lower 5% of units is reasonable

  - 3-fold variation in ORs or rates $\Rightarrow$ on log scale, $\theta_{95\%} - \theta_{5\%} = \log 3$

  - So our prior guess at $\sigma^{-2}$ would be

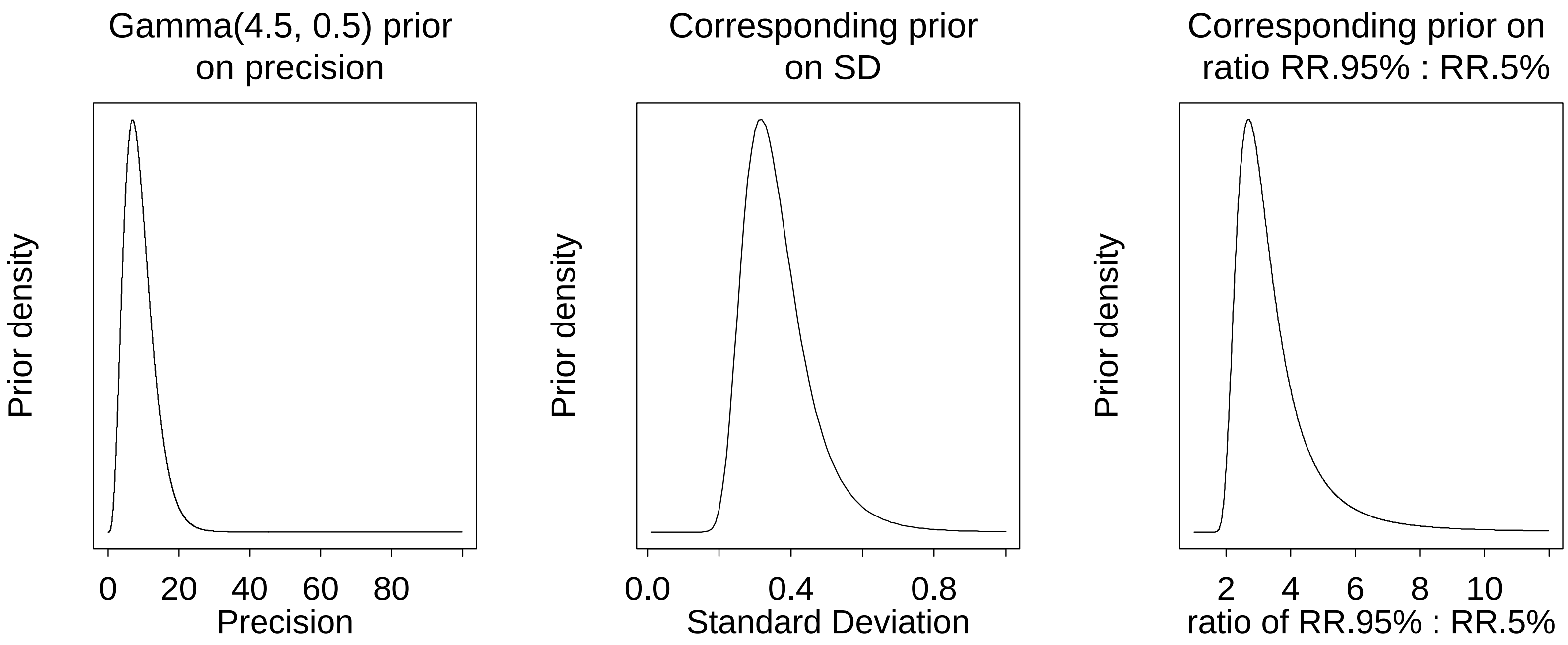
$$\sigma^{-2} \;\approx\; 3.29^2/(\log 3)^2 \;\approx\; 9$$

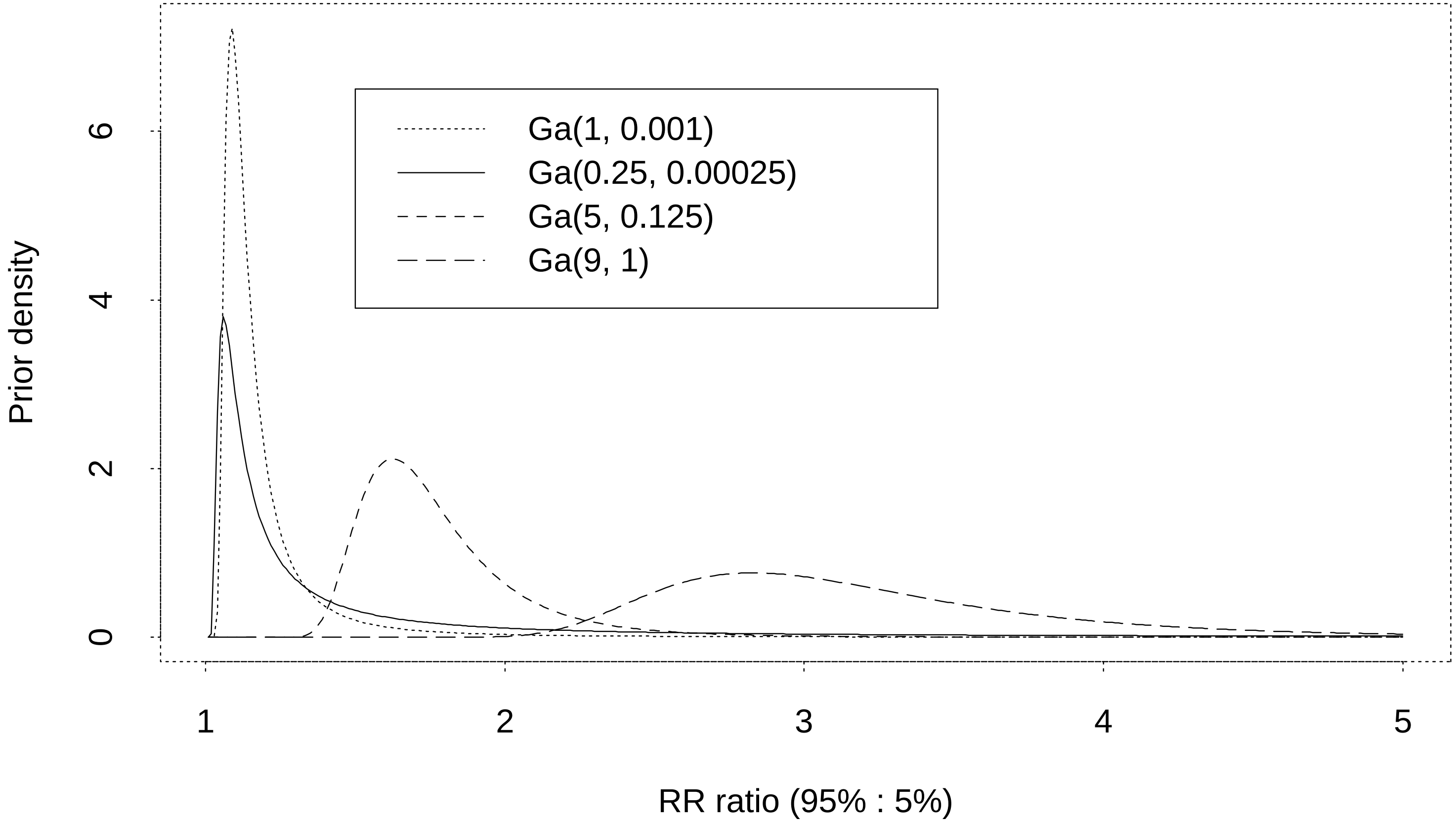
- To reflect our uncertainty in this prior guess, suppose we believe that 10-fold variation between the upper and lower 5% of units is very unlikely (say, less than a 1% chance) ⇒ lower 1% quantile of our prior distibution for the precision, $\sigma^{-2}$, would be

$$\sigma^{-2}_{1\%} \approx 3.29^2/(\log 10)^2 \;\approx\; 2$$

- So, choose gamma distribution with mean ≈ 9 and lower $99^{th}$ percentile ≈ 2:

$$\sigma^{-2} \sim \text{Gamma}(4.5, 0.5)$$





Other choices of gamma prior for the random effects precision, shown on the scale of the ratio between 95$^{th}$ and 5$^{th}$ percentiles of the prior distribution for $e^{\theta_i}$

Alternatively, if $\theta_{95\%} - \theta_{5\%} = 2 \times 1.645 \times \sigma$
$\Rightarrow$ exp($2 \times 1.645 \times \sigma$) is ratio of 95% to 5% in distribution of odds ratios or rates
$\Rightarrow$ exp($2 \times 1.96 \times \sigma$) is ratio of 97.5% to 2.5% in dist. of odds ratios or rates

| $\sigma$ | 95% range ($e^{3.92\sigma}$) | 90% range ($e^{3.29\sigma}$) |
|---:|---:|---:|
| 0.0 | 1.00 | 1.00 |
| 0.1 | 1.48 | 1.39 |
| 0.2 | 2.19 | 1.93 |
| 0.3 | 3.24 | 2.68 |
| 0.4 | 4.80 | 3.73 |
| 0.5 | 7.10 | 5.18 |
| 0.75 | 18.92 | 11.79 |
| 1.0 | 50.40 | 26.84 |
| 2.0 | 2540.20 | 720.54 |

- $\sigma$ around 0.1 to 0.5 may appear reasonable in many contexts
- $\sigma$ around 0.5 to 1.0 might be considered as fairly high
- $\sigma$ around 1.0 would represent fairly extreme heterogeneity

Similar approach can be used to 'calibrate' $\sigma$ in terms of, say, intra-class correlation $\sigma^2_{\text{btw}}/(\sigma^2_{\text{btw}} + \sigma^2_{\text{wth}})$ in normal-normal hierarchical models

## Example: Surgical (continued) — sensitivity to priors

Hierarchical model for surgical data:

$$
\begin{aligned}
r_i &\sim \text{Binomial}(n_i, \pi_i) \\
\text{logit}\pi_i &\sim \text{Normal}(\mu, \sigma^2) \\
\mu &\sim \text{Uniform}(-1000, 1000)
\end{aligned}
$$

Consider 5 alternative priors for random effects variance:

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

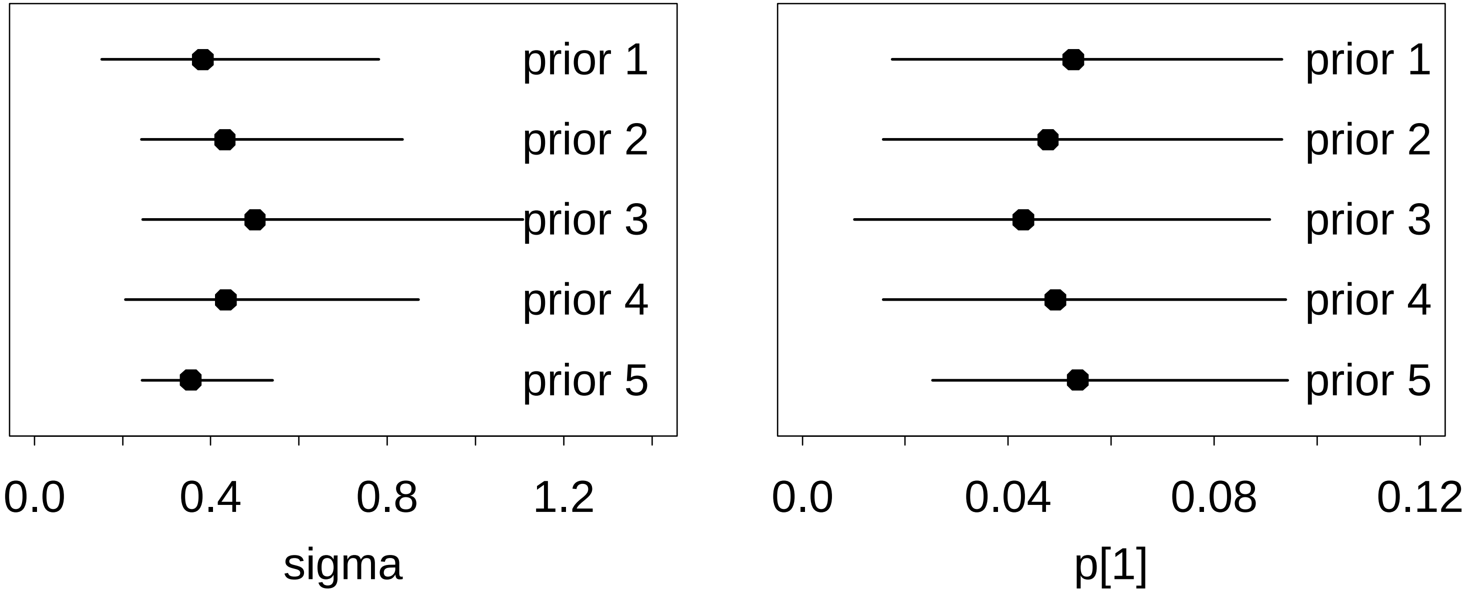
2. $\sigma^{-2} \sim \text{Gamma}(0.1, 0.1)$

3. $\sigma^2 \sim \text{Uniform}(0, 100)$

4. $\sigma \sim \text{Uniform}(0, 100)$

5. $\sigma^{-2} \sim \text{Gamma}(4.5, 0.5)$

Posterior median and 95% intervals for the random effects standard deviation (`sigma`) and the mortality rate in hospital 1 (`p[1]`), under each prior



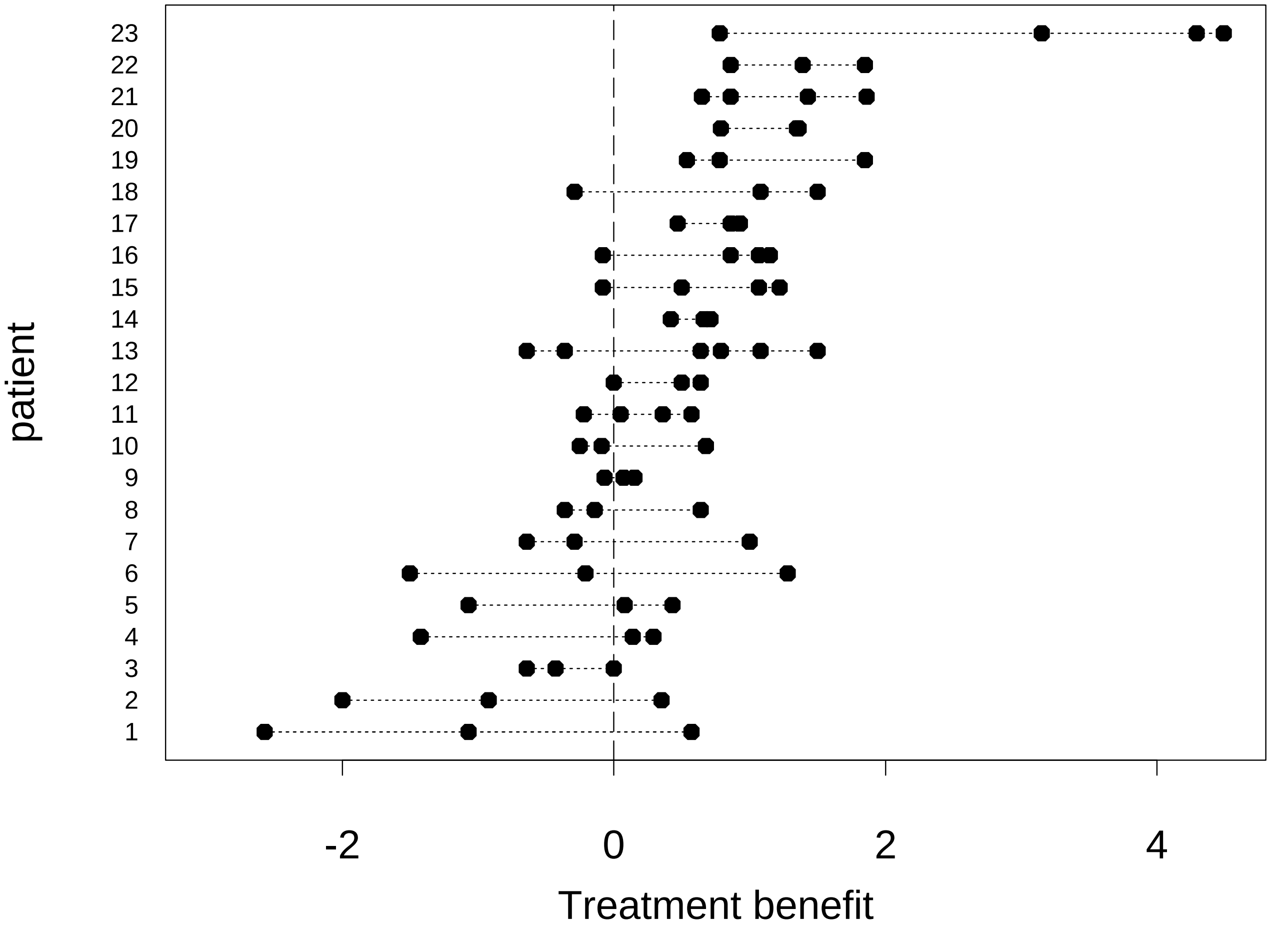
**Further reading**

Gelman (2005)

Spiegelhalter et al (2004) Section 5.7.3

# 5. More complex hierarchical models

# Raw data for each patient



## Statistical model

If $y_{kj}$ is the $j^{th}$ measurement on the $k^{th}$ individual, we assume

$$y_{kj} \sim \mathsf{N}(\theta_k, \sigma_k^2)$$

Assume both $\theta_k$'s and $\sigma_k^2$'s are *exchangeable*, in the sense there is no reason to expect systematic differences and we act as if they are drawn from some common prior distribution.

Note: alternative assumptions are either that $\theta_k$ and $\sigma_k^2$ are same for all patients (pooled model) or that they are independent (fixed effects) for each patient

We make the specific distributional assumption that

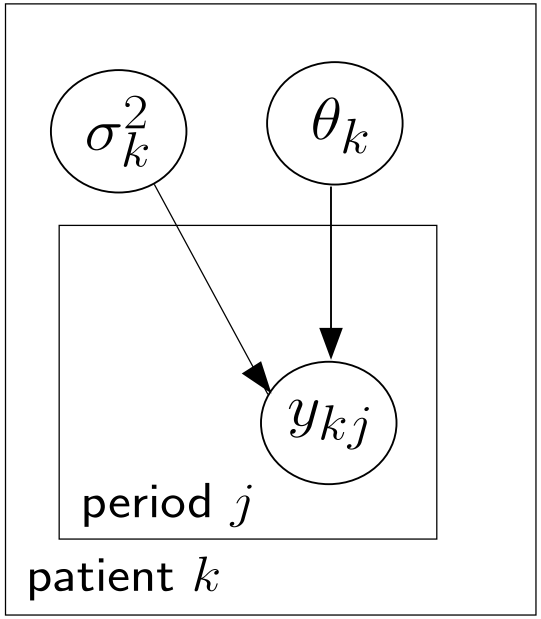
$$\begin{array}{rcl} \theta_k & \sim & \mathsf{N}(\mu_\theta, \phi_\theta^2) \\ \log(\sigma_k^2) & \sim & \mathsf{N}(\mu_\sigma, \phi_\sigma^2) \end{array}$$

A normal distribution for the log-variances is equivalent to a log-normal distribution for the variances

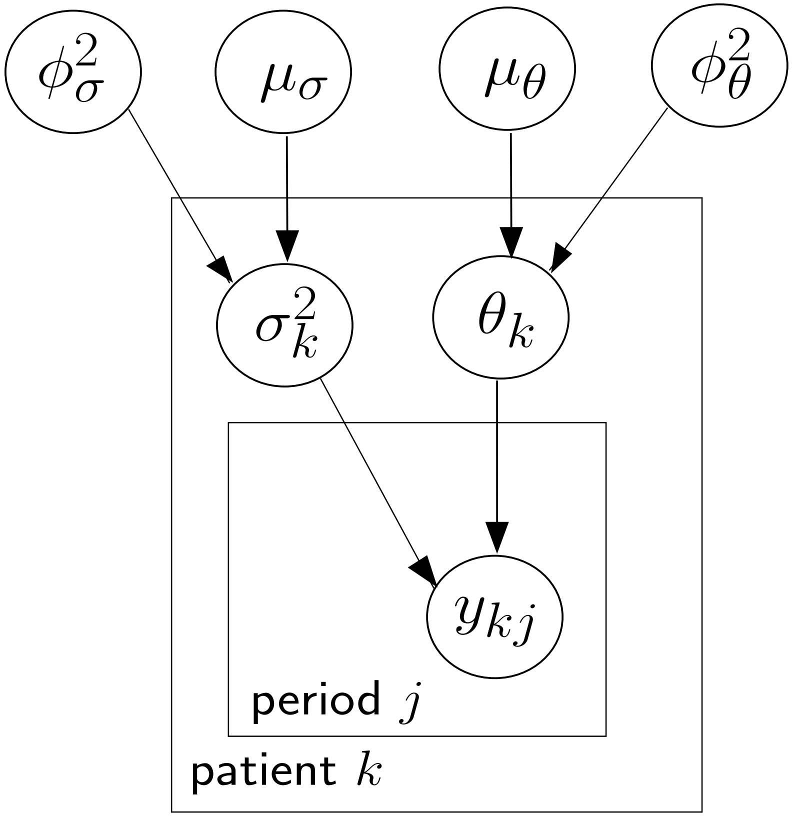
Uniform priors adopted for $\mu_\theta, \phi_\theta, \mu_\sigma$ and $\phi_\sigma$.

# Graphical model

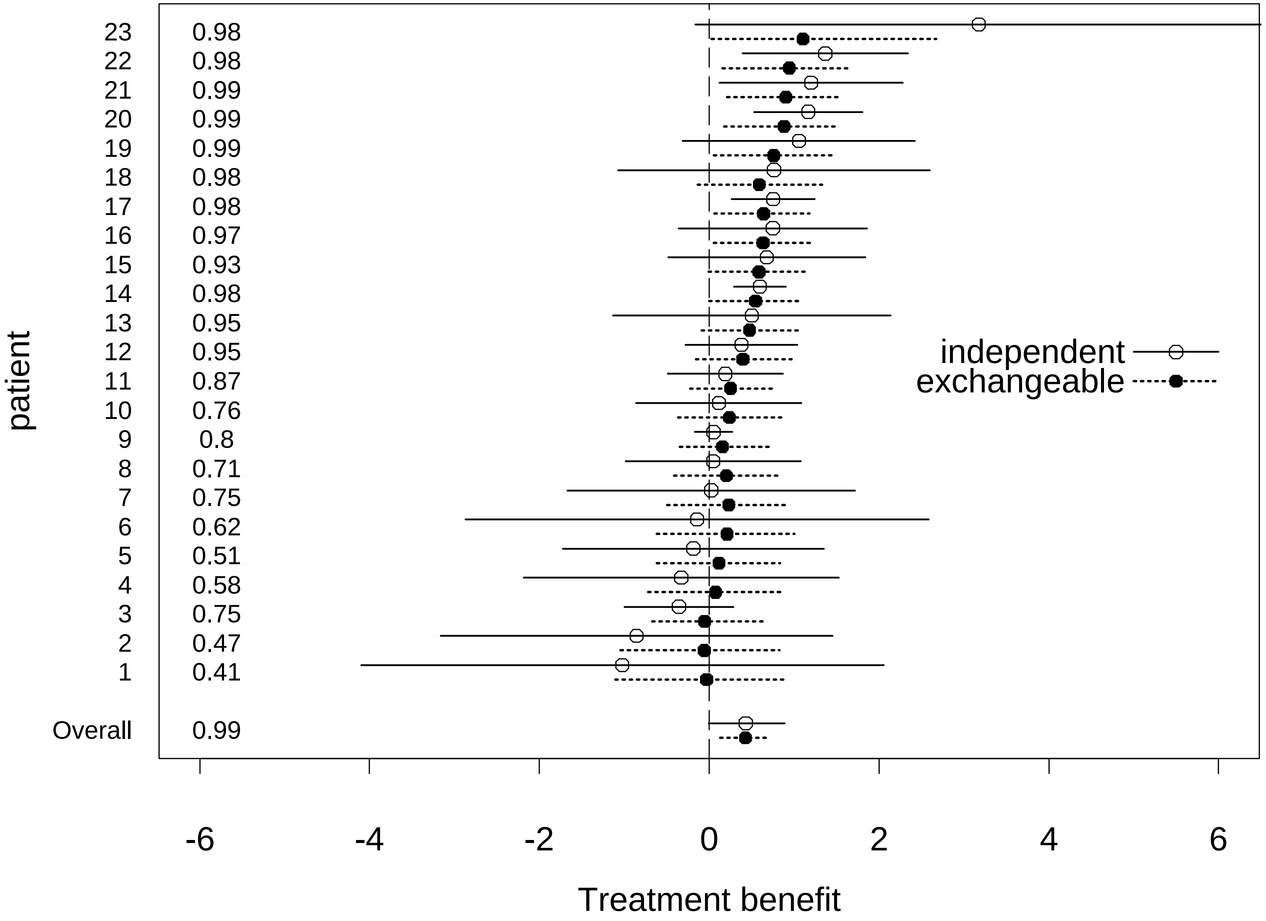
## Independent effect



## Exchangeable means and variances



# Estimates and 95% intervals for treatment effect, and posterior probability that effect $> 0$



## Interpretation

- Exchangeable model shrinks in the extreme patients, reflecting the limited information from each individual (see patient 23)

- It might be felt the model is exercising undue influence in this situation

- Despite shrinkage, narrower intervals mean that 9 patients have 95% intervals excluding 0 compared to 6 with the independent analysis

- One consequence of allowing exchangeable variances is that patient 9 has a *wider* interval under the exchangeable model
  - patient 9's observations were very close together → very narrow interval under independence model

- Straightforward to include patient-level covariates

- Sensitivity analysis to the shape of both the sampling and the random-effects distribution: say assuming $t$-distributions.

# 6. Missing data and Measurement error

# Missing data

*1. Classical approach:*

- Complete case analysis
  - Inefficient since throwing away data
  - Can be biased
- Imputation
  - 'Fill in' missing data with imputed values, then estimate parameters assuming imputed values were actually observed
  - Naive approach: replace missing data by mean of observed responses
    - underestimates true variation in response
    - may be biased
  - Multiple imputation (Rubin, 1978)
    - Generate $K > 1$ sets of imputations Re-estimate model using each 'completed' data set
    - Pool parameter estimates to obtain single estimate
    - Estimate variance by combining within and between-imputation variances

*2. Bayesian approach:*

- Inference based on joint posterior distribution of the parameters and missing data given the observed data and modelling assumptions

- Using MCMC ⇒ obtain samples of all the unknowns (*i.e.* parameters **and missing data**)

⇒
  - Missing values 'automatically' imputed
  - New values are sampled for the missing observations at each iteration (*cf* multiple imputation)
  - Posterior estimates of model parameters will be fully adjusted for uncertainty in the imputed observations (conditional on the assumed model)

- Missing value code in `BUGS` is `NA`

## Example: Rats — repeated responses with missing values

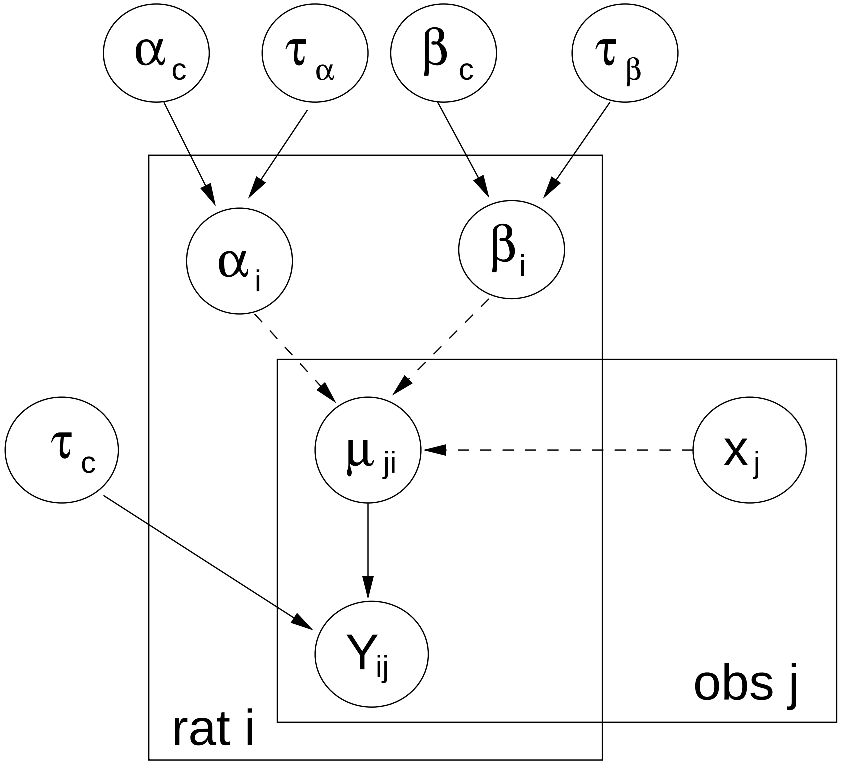
- 30 rats
- Weight measured weekly for five weeks
- Suppose a random 33% of measurements are missing

| | Weights $Y_{ij}$ of rat $i$ on day $x_j$ | | | | |
|---|---|---|---|---|---|
| | $x_j = 8$ | 15 | 22 | 29 | 36 |
| Rat 1 | 151 | 199 | 246 | 283 | 320 |
| Rat 2 | NA | 199 | 249 | 293 | NA |
| ....... | | | | | |
| Rat 6 | NA | 210 | 252 | NA | NA |
| ....... | | | | | |
| Rat 30 | 153 | NA | 244 | 286 | 324 |

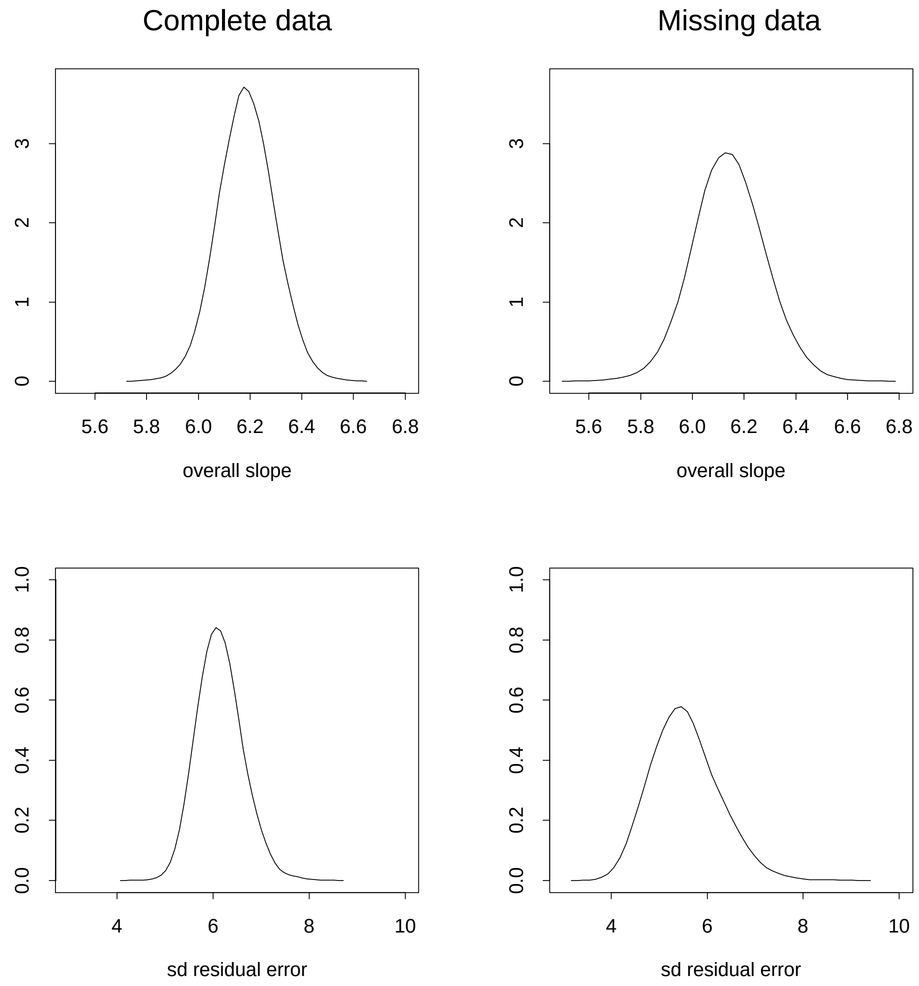
.

Model as random effects linear growth curve

$$
\begin{aligned}
Y_{ij} &\sim \text{Normal}(\mu_{ij}, \tau_c) \\
\mu_{ij} &= \alpha_i + \beta_i x_j \\
\alpha_i &\sim \text{Normal}(\alpha_0, \tau_\alpha) \\
\beta_i &\sim \text{Normal}(\beta_0, \tau_\beta)
\end{aligned}
$$



.



Note: Results with missing data same as complete-case analysis here, since missing values are only in response, and mechanism is assumed ignorable.

## When can we ignore the mechanism causing the missing data?

Some notation:

$$
\begin{aligned}
Y &= \text{complete response vector} \\
Y_{obs} &= \text{observed part of } Y \\
Y_{miss} &= \text{missing part of } Y \\
M &= \text{missing data indicator} \\
&= \text{0 when } Y \text{ observed} \\
& \quad\; \text{1 when } Y \text{ missing} \\
X &= \text{covariates} \\
\theta &= \text{ parameters of interest} \\
\phi &= \text{nuisance parameters}
\end{aligned}
$$

For the Rats example:

$Y$ matrix:

```
151 199 246 283 320
NA 199 249 293 NA
.......
153 NA 244 286 324
```

$M$ matrix:

```
0 0 0 0 0
1 0 0 0 1
......
0 1 0 0 0
```

$$\theta = \alpha_i, \beta_i, \alpha_c, \beta_c, \tau_c, \tau_\alpha, \tau_\beta$$

- Likelihood for observed data $(Y_{obs}, M)$ is given by integrating joint density over $Y_{miss}$:

$$p(Y_{obs}, M|\theta, \phi) = \int p(Y_{obs}, Y_{miss}|\theta)p(M|Y_{obs}, Y_{miss}, \phi)dY_{miss}$$

- Bayesian inference under full model for $Y$ and $M$ is then based on the posterior

$$p(\theta, \phi|Y_{obs}, M) \propto p(\theta, \phi)p(Y_{obs}, M|\theta, \phi)$$

- Ignoring the missing data mechanism is equivalent to basing inference about $\theta$ instead on the posterior $p(\theta|Y_{obs})$

  → When is this valid?

  → Valid if $p(\theta, \phi|Y_{obs}, M) \propto p(\theta|Y_{obs})$

- If missing data mechanism does not depend on $Y_{miss}$ (i.e. data are ‘missing at random’, MAR) then

$$\begin{aligned} p(M|Y_{obs}, Y_{miss}, \phi) &= p(M|Y_{obs}, \phi) \\ \Rightarrow \;\; p(Y_{obs}, M|\theta, \phi) &= \int p(Y_{obs}, Y_{miss}|\theta) p(M|Y_{obs}, \phi) dY_{miss} \\ &= p(M|Y_{obs}, \phi) p(Y_{obs}|\theta) \end{aligned}$$

- Further, if $\theta$ and $\phi$ are assumed a priori independent then

$$p(\theta, \phi) = p(\theta) p(\phi)$$

- Then

$$\begin{aligned} p(\theta, \phi|Y_{obs}, M) &\propto p(\theta) p(Y_{obs}|\theta) \times p(\phi) p(M|Y_{obs}, \phi) \\ &\propto p(\theta|Y_{obs}) p(\phi|Y_{obs}, M) \\ &\propto p(\theta|Y_{obs}) \quad \text{as a function of } \theta \end{aligned}$$

So, missing data mechanism is ignorable for Bayesian inference about $\theta$ if data are MAR and $\theta$ and $\phi$ are a priori independent (Little and Rubin, 2002)

## Example of an ignorable missing data mechanism

- Example 1: Suppose some measurements are missing because the lab technician only had time to weigh (a randomly selected) $\frac{2}{3}$ of the rats on each occasion
  $\Rightarrow \pi(M|X, Y, \phi) = \pi(M|\phi) = \phi = 0.333$

## Example of a non-ignorable missing data mechanisms

- Example: On each occasion, suppose the fattest looking rats were removed for use in another experiment, and so their weights were not actually recorded for the current study
  $\Rightarrow p(M|X, Y, \phi) = p(M|Y_{miss}, Y_{obs}, \phi)$
  - Failure to explicitly model missing data mechanism will lead to biased estimates of $\theta$

## Handling missing data in WinBUGS

*Missing response data, assuming missing data mechanism is ignorable*

- denote missing observations by `NA` in the data file
- specify response distribution (likelihood) as you would for complete data
- missing data are treated as additional unknown parameters

⇒ WinBUGS will automatically simulate values for the missing observations according to the specified likelihood distribution, conditional on the current values of all relevant unknown parameters

Missing response data is essentially a prediction problem

*If missing data mechanism is non-ignorable*

- include missing data indicator $M$ in data file
- include extra term in model specifying the likelihood for $M$ — this distribution should depend on $Y_{miss}$ in some way, e.g.

```
          M[i] ~ dbern(delta[i])
  logit(delta[i]) <- phi[1] + phi[2]*Y[i]
```

- Usually need informative priors on parameters of missing data model (`phi`) as no information in the data. See Best *et al.* (1996) for an example.

*Missing covariate data*

- denote missing observations by `NA` in the data file
- specify prior distribution for the covariate
  - e.g. if $X$ is a continuous covariate containing some missing values, could specify $X_i \sim$ Normal$(\mu, \sigma^2)$ or build regression model relating $X_i$ to other observed covariates
  - can then assume vague priors for $\mu$ and $\sigma^2$; posterior distribution of $\mu$ and $\sigma^2$ will be informed by the observed part of the vector of $X$'s
- WinBUGS will automatically simulate values from the posterior distribution of the missing covariates (which will depend on the prior for the $X$'s and the likelihood contribution from the corresponding response variable)

## Example: Childhood malaria in the Gambia

Diggle et al (2002)

*Data:*

- 2035 children in 65 villages in the Gambia
- Response: Binary indicator of presence of malarial parasites in blood sample taken from each child
- Covariates include: child's age and use of bed nets, inclusion/exclusion of village from primary health care system and greenness of surrounding vegetation (from satellite information)

*Questions of interest include:*

- Does sleeping under a bed net reduce risk of malaria?

Basic model is a logistic regression of the probability of malaria, with bed net use (binary) as well as other variables as predictors

- Consider a slightly modified version of the malaria dataset:
  - BEDNET = binary indicator of whether child sleeps under a (treated) bed net
  - Suppose the value of BEDNET is missing for 30% of children
- Consider 2 alternative models for the missing covariate:
  1. Probability of BEDNET = 1 is same for all children *a priori*

$$\begin{aligned} \text{BEDNET}_i &\sim \text{Bernoulli}(q) \\ q &\sim \text{Beta}(1, 1) \end{aligned}$$

  2. Probability of BEDNET = 1 depends on whether or not village belongs to primary health care system (PHC)

$$\begin{aligned} \text{BEDNET}_i &\sim \text{Bernoulli}(q_i) \\ \text{logit} q_i &= \gamma_1 + \gamma_2 \text{PHC}_i; \quad (+ \text{ vague priors on } \gamma_1 \text{ and } \gamma_2) \end{aligned}$$

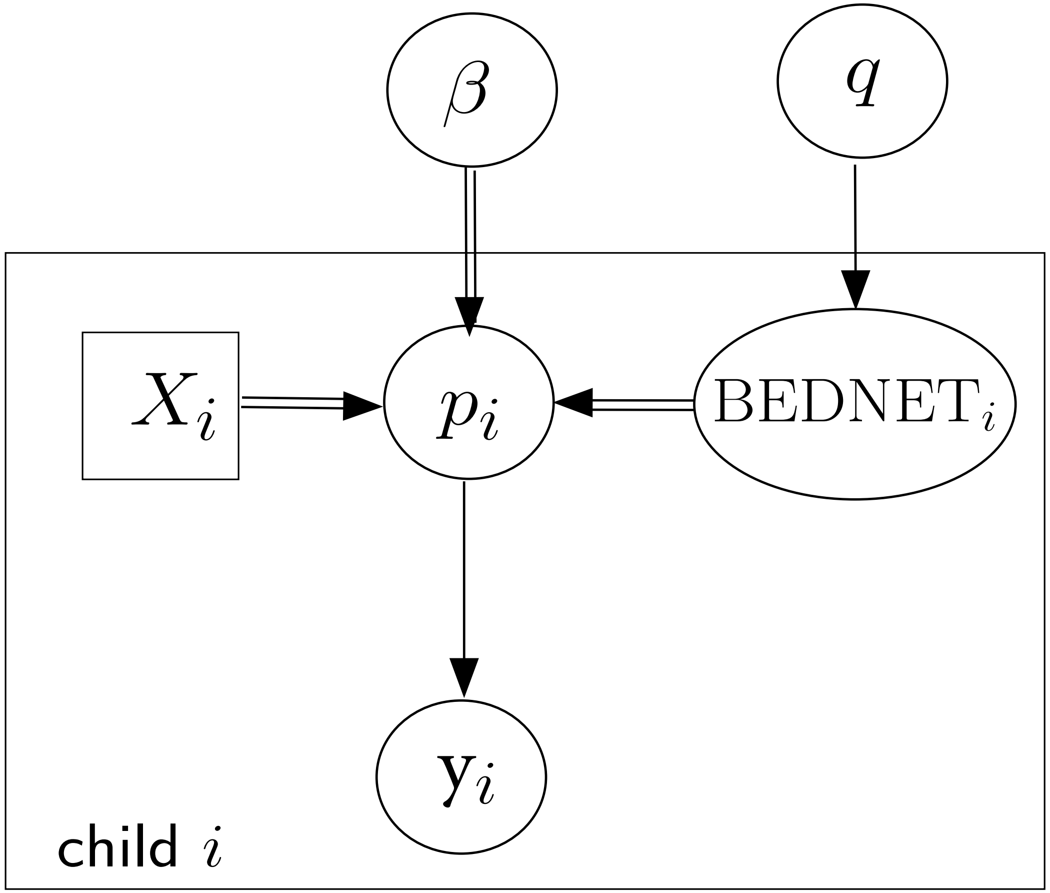
## WinBUGS code for model 1

```
model {
  for(i in 1:2035) {
    Y[i] ~ dbern(p[i])
    logit(p[i]) <- alpha + beta.age[AGE[i]] + beta.bednet*BEDNET[i] +
                   beta.green*(GREEN[i] - mean(GREEN[])) + beta.phc*PHC[i]
  }
  # model for missing exposure variable
  for(i in 1:2035) { BEDNET[i] ~ dbern(q)  } # prior model for whether or not child
                                             # i sleeps under treated bednet
  q ~ dbeta(1, 1) # vague prior (uniform) on prob of sleeping under treated bednet

  # vague priors on regression coefficients
  alpha ~ dflat()
  beta.bednet ~ dflat()
  ........etc.......

  # calculate odds ratios of interest
  OR.bednet <- exp(beta.bednet)      # odds ratio of malaria for children using
                                     # treated bednets
  PP.bednet <- step(0.8 - OR.bednet) # probability that using treated bed net
                                     # reduces risk of malaria by at least 20%
}
```

DAG for Model 1



WinBUGS code for model 2

- Replace model for missing exposure variable by

```
# model for missing exposure variable
for(i in 1:2035) {
   BEDNET[i] ~ dbern(q[i])  # prior model for whether or not child i
                            # sleeps under treated bednet
   logit(q[i]) <- gamma[1] + gamma[2]*PHC[i] # allow prob of using treated
                                             # bednet to depend on whether
                                             # or not village belongs to
                                             # primary health care system
}
for(k in 1:2) {   gamma[k] ~ dflat()     }
OR.treated.phc <- exp(gamma[2])              # odds ratio of sleeping under
                                             # treated bednet for children
                                             # living in villages in the PHC
```

Results

| | OR.bednet | | PP.bednet | OR.age2 | |
|---|---|---|---|---|---|
| | Mean | 95% interval | | Mean | 95% interval |
| No missing data | 0.57 | (0.45, 0.72) | 0.99 | 1.40 | (1.06, 1.81) |
| Model 1 | 0.66 | (0.49, 0.86) | 0.93 | 1.39 | (1.06, 1.80) |
| Model 2 | 0.64 | (0.47, 0.83) | 0.95 | 1.41 | (1.06, 1.83) |
| Single imputation* | 0.76 | (0.61, 0.95) | 0.68 | 1.40 | (1.05, 1.80) |
| Complete case | 0.63 | (0.47, 0.83) | 0.96 | 1.70 | (1.20, 2.35) |

*Imputed using observed proportion of bed net users

# 7. Practical introduction to Bayesian model criticism for hierarchical models

# General framework

Need to distinguish three stages

1. *Criticism:* exploratory checking of a single model, which may suggest -

2. *Extensions:* embed initial model in list of alternatives, which leads to -

3. *Comparison:* assess candidates in terms of their evidential support and influence on conclusions of interest.

There should be iteration between these stages.

# Model Criticism

- Model checking is not about asking
  - ‘Is model true or false?’
- Model checking is about asking
  - ‘Does the model provide an adequate description of important features of the data?’
  - ‘Do deficiencies in the model have a noticeable effect on substantive inference?’
- Important to think about which aspect of model you want to test.
  - Likelihood or prior
  - Individual data points or aggregated units

## Recall: residuals in non-hierarchical models

- Standardised Pearson residuals $(y - \theta)/\sigma$ where $\theta = \mathsf{E}[Y]$, $\sigma^2 = \mathsf{V}[Y]$
- In Bayesian analysis these are random quantities, with distributions
- If assuming Normality, then

  $$P(Y) = \Phi[(Y - \theta)/\sigma]$$

  has a Uniform[0,1] distribution under true $\theta$ and $\sigma$

## Model

Data $y$, parameters $\theta$

## Making predictions

- $y_f$ used to fit the model
- $y_c$ for criticism
- Predictive distribution

$$
\begin{aligned}
p(y_c^{pred}|y_f) &= \int p(y_c^{pred}|y_f,\theta)p(\theta|y_f)d\theta \\
&= \int p(y_c^{pred}|\theta)p(\theta|y_f)d\theta
\end{aligned}
$$

Different ways to choose $y_f$, $y_c$ ...

## Cross-Validation

$y_f$ and $y_c$ are non-overlapping

- Take out part of data $y_c$
- Run model on $y_f$ to obtain posterior distribution $p(\theta|y_f)$.
- Predict new data $p(y_c^{pred}|y_f)$.
- Compare observed data $y_c$ with predicted distribution $p(y_c^{pred}|y_f)$.

Model checking: repeat above for each data point (or unit). If too many points in disagreement conclude model is not adequate.

Detecting outliers: take one data point out at a time, find which ones don't agree with the predictive distribution from the rest of the data.

## Posterior Predictive Checks

For large data sets, unfeasible to run model with each data unit removed. (E.g. gene expression data has thousands of data points.)

$y_f = y$

- Run model on **all** the data to obtain posterior distribution $p(\theta|y)$.
- Predict new data $p(y_c^{pred}|y)$.
- Compare observed data $y_c$ with predicted distribution $p(y_c^{pred}|y)$.

This method is conservative, because the observed data $y_c$ is used to obtain the posterior distribution. If $y_c$ is an outlier, the posterior $p(\theta|y)$ will be different from $p(\theta|y_{\backslash c})$.

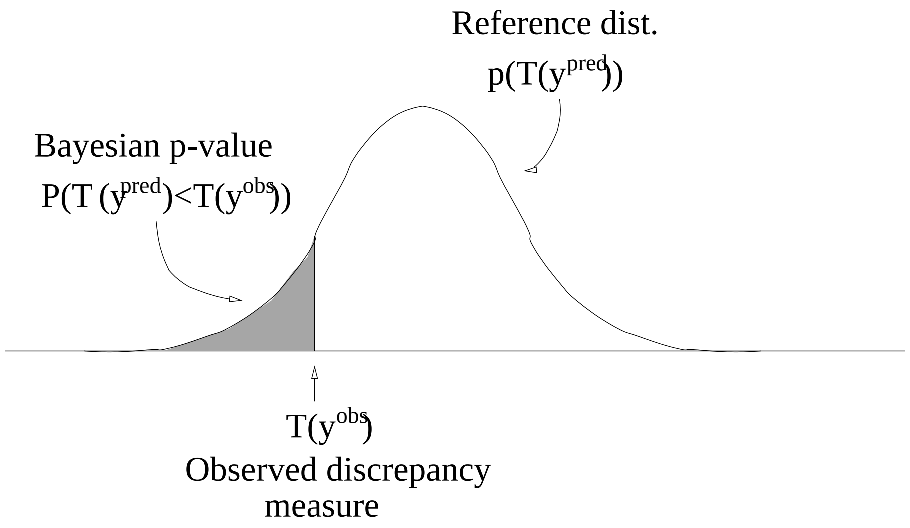
## How to compare observed data to predictive distribution?

1) Choose Discrepancy Function $T(y, \theta)$

- May be function of data and/or parameters
- Extreme $T()$ should indicate data conflict with model
- Often $T(y_c) = y_c$ to check for individual outliers

2) Checking Function

- For small data sets, simple plots of distributions
- Bayesian p-values



- (1-sided) probability that predicted data could be more extreme than observed, as measured by the discrepancy statistic $T(y_c)$:

$$p_{\mathsf{Bayes}} \;=\; \mathsf{Pr}\left(\,T(y_c^{pred}) \leq T(y_c) \mid y_f\,\right)$$

## Bayesian p-values

- Extreme $p$-values indicate conflict between data and aspect of the model under investigation
- Distribution of p-values is Uniform if model is "true"
- Usual warnings about p-values apply:
  - $p_{\text{Bayes}} \neq \Pr(\text{model is true} \mid \text{data})$
  - The $p$-value measures *statistical*, not *practical*, significance.
  - Suffers from usual problems of multiple testing

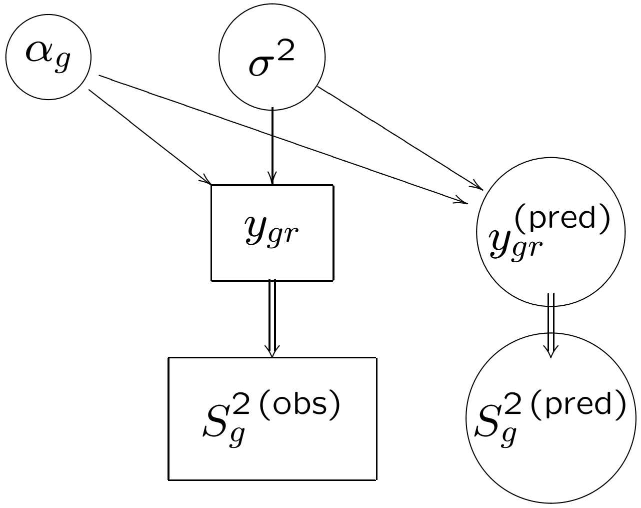
## Example of Posterior Predictive Model Checking

Gene expression data for genes $g = 1, ..., G$, repeat measurements $r = 1, ..., R$. Typical values, $G = 12000$, $R = 3$.

Log gene expression values $y_{gr}$, modelled as Normal, with the same variance $\sigma^2$ for all genes.

$$y_{gr} \sim N(\alpha_g, \sigma^2)$$

Want to test if assumption of equal variances reasonable.



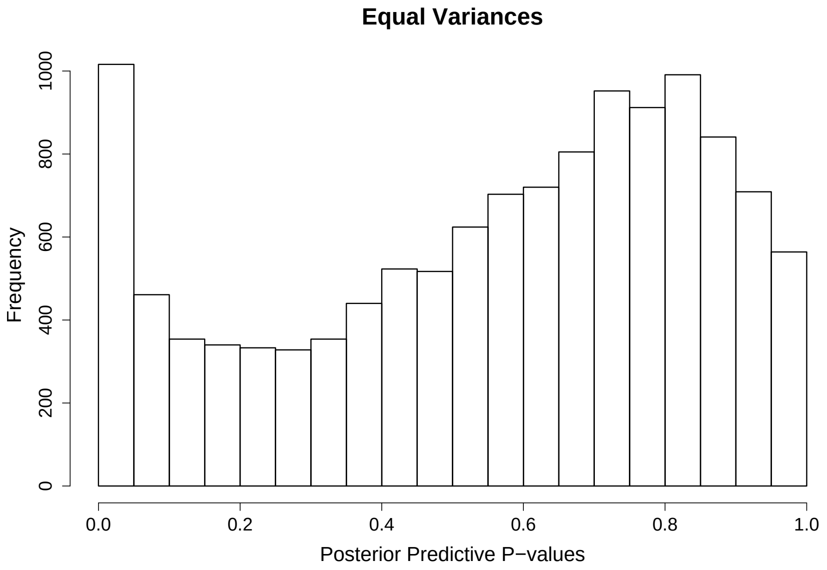
- For each $g$, predict new data $y_{gr}^{\text{(pred)}}$
- For each $g$ calculate sample variance $T(y_{gr}^{\text{(pred)}}) = S_g^{2\text{(pred)}}$ (sum of squares)
- Calculate $\mathbb{P}(S_g^{2\text{(pred)}} > S_g^{2\text{(obs)}})$

## Algorithm to calculate Bayesian p-values in MCMC

At each iteration $j = 1, ..., N$ of the MCMC sampler,

- Generate $y_{grj}^{\text{(pred)}}$ from the full conditional for $y_{gr}$
- Calculate $S_{gj}^{2\,\text{(pred)}} \equiv 1/R \sum_r (y_{grj}^{\text{(pred)}} - \bar{y}_{g.j}^{\text{(pred)}})^2$
- $M_{gj} \equiv I[S_{gj}^{2\,\text{(pred)}} \geq S_g^{2\,\text{(obs)}}]$

The Bayesian p-value for unit $g$ is $p_g = 1/N \sum_{j=1}^{N} M_{gj}$ (posterior expectation).



Posterior predictive p-values for gene expression Model 1.

Posterior predictive p-values are conservative; cross-validation p-values would be even more non-Uniform.

## Mixed Predictive Checks

Hierarchical models: data $y_i$, parameters $\theta_i, \psi$

Model: $\prod_i p(y_i|\theta_i)p(\theta_i|\psi)$

Posterior predictive checks test the likelihood. For hierarchical models, may want to check the assumptions on the intermediate level parameters.

$y_f = y$

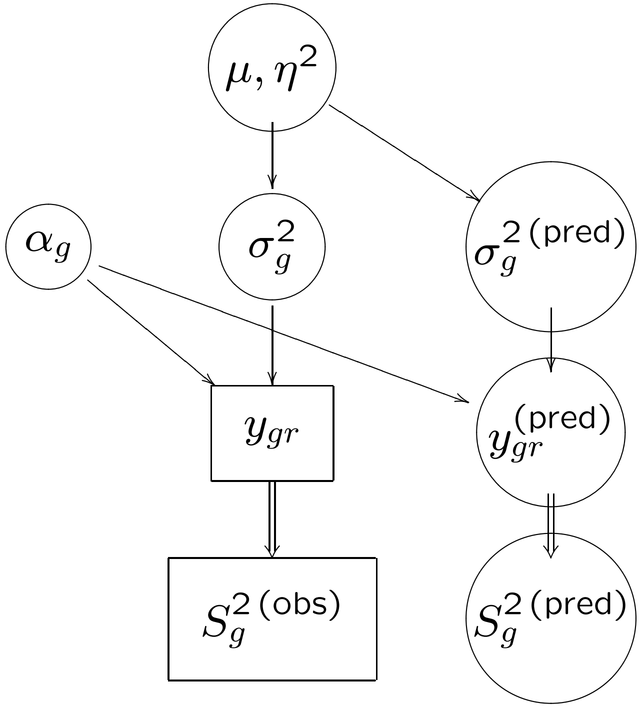
- Run model on **all** the data to obtain posterior distribution $p(\psi|y)$.
- Predict new parameters $p(\theta_c^{pred}|y)$
- Predict new data $p(y_c^{pred}|y)$.
- Compare observed data $y_c$ with predicted distribution $p(y_c^{pred}|y)$.

## Example of Mixed Predictive Model Checking

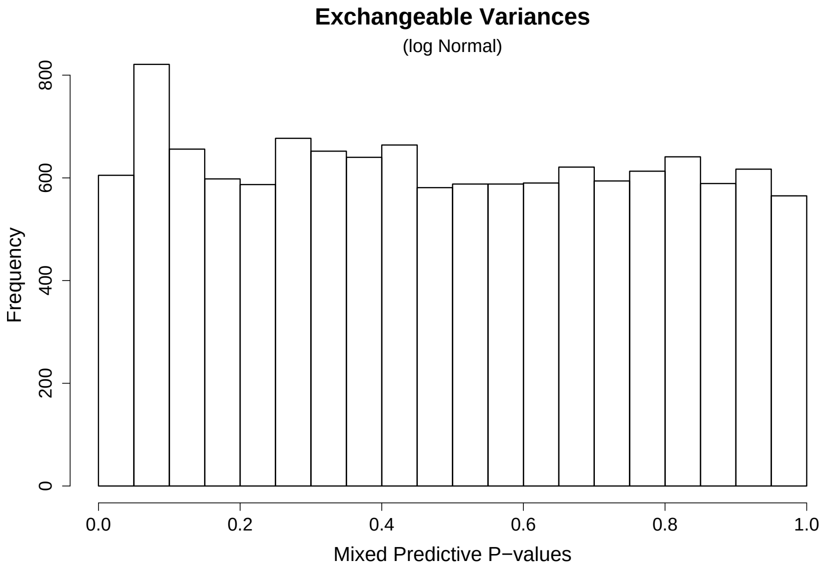
Gene expression data again, this time with exchangeable variances:

$$
\begin{aligned}
y_{gr} &\sim N(\alpha_g, \sigma_g^2) \\
\sigma_g^2 &\sim logNorm(\mu, \eta^2)
\end{aligned}
$$

Want to test if assumption of exchangeable variances reasonable.



- For each $g$, predict new data $\sigma_g^{\text{(pred)}}$
- For each $g$, predict new data $y_{gr}^{\text{(pred)}}$
- For each $g$ calculate sample variance $T(y_{gr}^{\text{(pred)}}) = S_g^{2\text{(pred)}}$ (sum of squares)
- Calculate $\mathbb{P}(S_g^{2\text{(pred)}} > S_g^{2\text{(obs)}})$

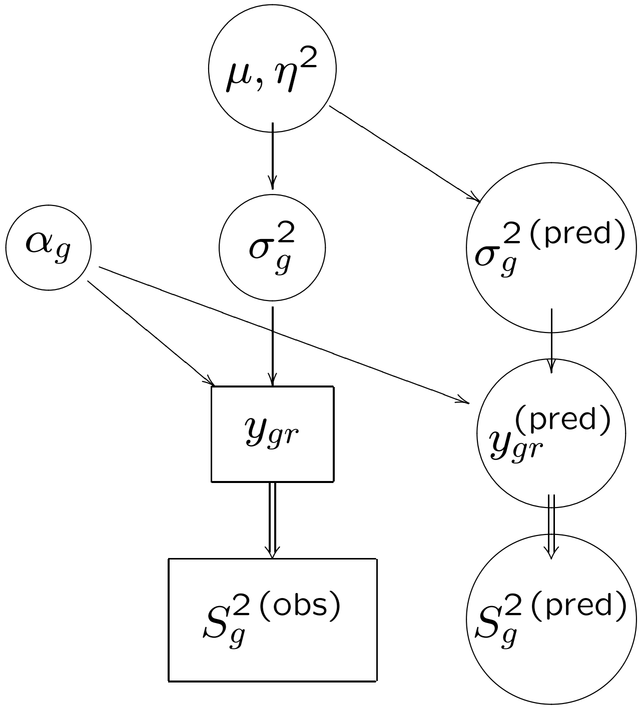


Mixed predictive p-values for gene expression Model 2.

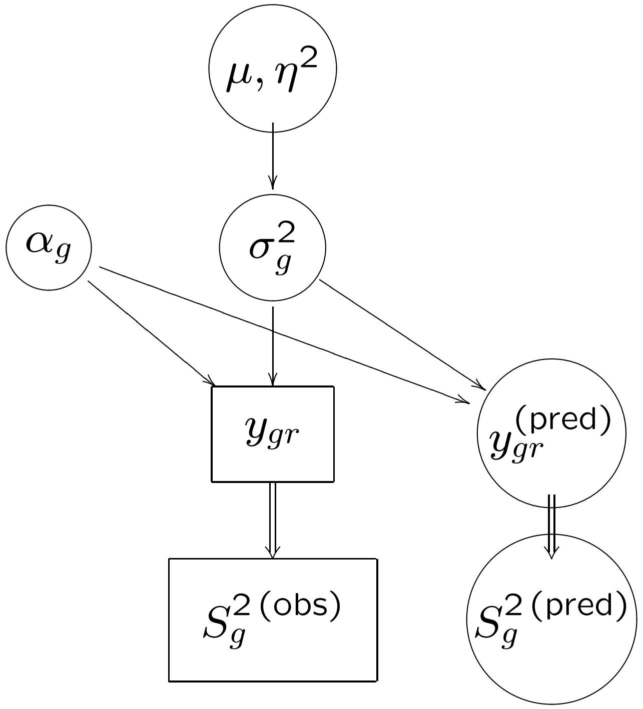
## Part of the WinBUGS code for the Mixed Predictive Checks

```
###### 1st level: likelihood and mixed predictive p-values
for( i in 1 : n ) {
    for( j in 1 : 3) {
       y[i, j] ~ dnorm(x[i], tau[i])
       ynew[i, j] ~ dnorm(x[i], taunew[i])
    }
    s2[i] <- pow(sd(y[i, ]), 2)
    s2new[i] <- pow(sd(ynew[i, ]), 2)
    pval[i] <- step(s2new[i] - s2[i])
}
###### 2nd level: exchangeable gene variances
for( i in 1 : n ) {
    tau[i] <- 1.0/sig2[i]
    taunew[i] <- 1.0/sig2new[i]
    sig2[i] <- dlnorm(mu,etaminus2)
    sig2new[i] <- dlnorm(mu,etaminus2)
}
```

# Different level of checking means a different predictive part of the model

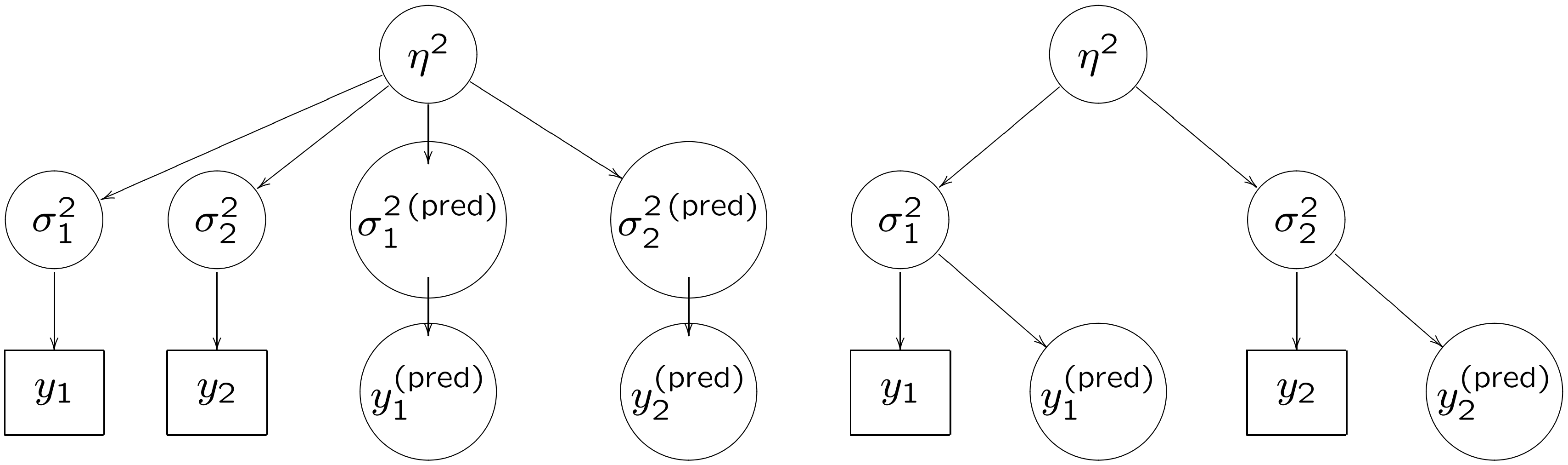


MIXED PREDICTIVE CHECKS



POSTERIOR PREDICTIVE CHECKS

# Mixed Checks are less conservative than Posterior Checks



MIXED PREDICTIVE CHECKS POSTERIOR PREDICTIVE CHECKS

Predicted values $y_i^{(\text{pred})}$ are much more sensitive to the data $y_i$ in the posterior checks than in the mixed checks.

(See from graph: for posterior checks $y_i^{(\text{pred})}$ and $y_i$ are connected through $\sigma_i^2$. For mixed checks the connection involves $\eta^2$, which is affected by all data points, not just $i$.)

## Example of Cross-Validation used to detect Outlying Data: Assessing prior-data conflict

Bristol Royal Infirmary Enquiry: investigation into outcomes of heart surgery on children between 1984 and 1995.

12 hospitals were compared. Several categories of operations.

Number of deaths $r_{hs}$ in hospital $h$, category $s$, number of operations $n_{hs}$.

$$r_{hs} \sim Bin(p_{hs}, n_{hs})$$

## Hierarchical Model:

Model the mortality rates $p_{hs}$ as **exchangeable** between hospitals:

$$\begin{aligned} \text{logit}(p_{hs}) &\sim N(\mu_s, \nu_s) \\ \log(\nu_s) &\sim N(\phi, \psi) \end{aligned}$$

Bayesian model fit by MCMC (WinBUGS), calculated predictive distribution for mortality rates $p_{hs}^{pred}$ conditional on data from all centres except Bristol.

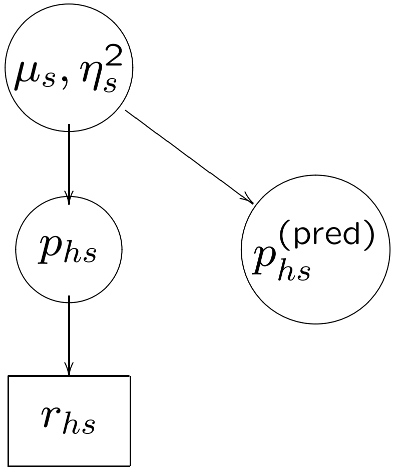
## Fixed Effects Model:

Model the mortality rates $p_{hs}$ as **independent** between hospitals:

$$\text{logit}(p_{hs}) = \mu_s + \beta_{hs}$$

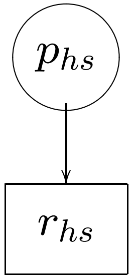
where $\mu_s$ and $\beta_{hs}$ are fixed effects.

Get posterior distribution for $p_{1s}$ for Bristol ($h = 1$).



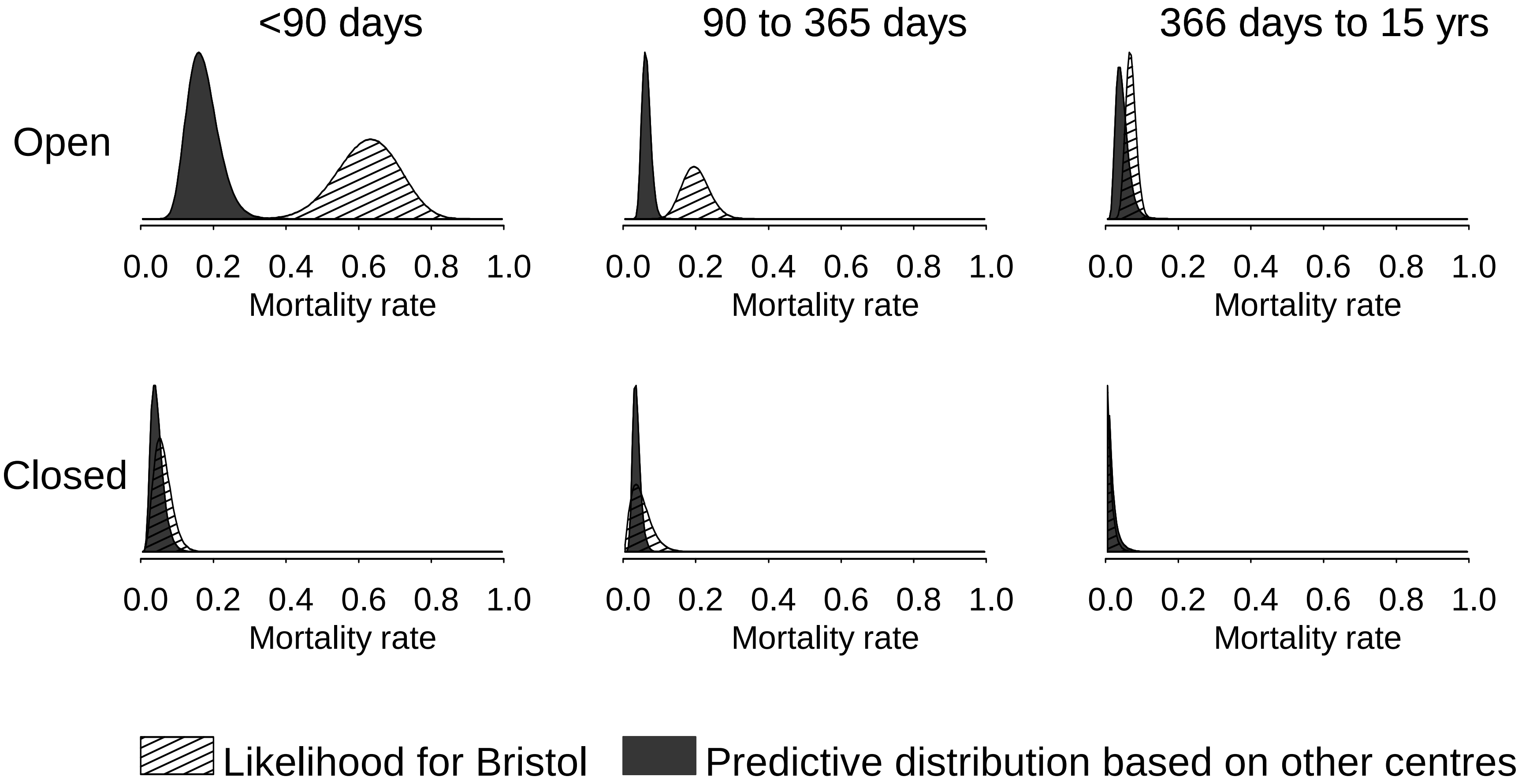
EXCHANGEABLE MODEL

"Prior from other hospitals"



INDEPENDENT MODEL

"Likelihood" for Bristol


<90 days
90 to 365 days
366 days to 15 yrs
Open
Closed
0.0
0.2
0.4
0.6
0.8
1.0
Mortality rate
Likelihood for Bristol
Predictive distribution based on other centres

## Summary: recommended strategy

1. If concern lies solely with the likelihood $p(y_i|\theta_i)$ for a vector $y_i$, then use posterior predictive replication of individual observations.

2. If concern lies solely with the prior $p(\theta_i|\psi)$ then

   (i) if closed form parameter estimates (e.g. sample variance $S_g^2$ in our example), use *mixed* replication

   (ii) if not, use *parameter* replication in which prior and likelihood replications are contrasted (e.g. the Bristol Inquiry example).

3. If we are concerned with both prior *and* likelihood, then

   - if $y_i$ is a vector, first use strategy 1, and then strategy 2 to check the prior.

   - if scalar, cannot separate prior from likelihood, so use 2(i).

# 8. Concluding Remarks

# Hierarchical models

Many interlinked arguments to favor the use of hierarchical models:

- by breaking down the problem in layers, able to separate structural judgments on observables and on parameters, and subjective information
- natural structure for expressing dependence, prior correlations, ... in a plausible way
  - ‘structural’ assumptions still need to be carefully justified
  - Bayesian non-parametric and semi-parametric models (e.g. mixtures) offer considerable flexibility for representing ‘similarity’ assumptions in hierarchical models
- reduces the arbitrariness of hyperparameter choice → robustify the inference
- through shrinkage and borrowing of strength, parameter estimates are stabilized
- computationally feasible via MCMC algorithms within Bayesian paradigm

## Comparison of with non-Bayesian approaches to random effects modelling

- Classical approach usually treats random effects as nuisance parameters to be integrated out

- Outside of normal linear hierarchical models, classical estimation methods provide exact solution to approximate problem, whereas Bayesian approach provides arbitrarily precise (modulo MC error) solution to exact problem

- For large samples (units and observations per unit), Bayesian and classical inference tend to give similar results

- For small samples
  - Little information available for estimating between-unit variability
  - Classical methods can be unreliable
  - Bayesian approach allows inclusion of external knowledge (prior) which can help stabilize the model

# MCMC

## Some strengths of MCMC

- Freedom in modelling
  - in principle, no limits
  - well-adapted for models defined on sparse graphs
- Freedom in inference
  - in principle, no limits
  - can estimate arbitrary functions of model parameters (e.g. ranks, probabilities of threshold exceedence)
  - opportunities for simultaneous inference
- Allows/encourages sensitivity analysis
- Coherently integrates uncertainty
- Only available method for many complex problems

## Some weaknesses and dangers

- Order $\sqrt{N}$ precision
- Possibility of slow convergence and difficulties is diagnosis
- Risk that fitting technology runs ahead of statistical science
- Risk of undisciplined, selective presentation
- Difficulty of validating code

**Thank you for your attention!**

## References

Bennett, JE and Wakefield, JC (2001). Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. *Biometrics*, **57**, 803-812.

Berger, JO (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

Berger JO and Berry DA (1988) !!!!

Berry, DA (1987). Interim analysis in clinical trials — the role of the likelihood principle. *American Statistician*, **41**, 117–22.

Berry, DA (1996). *Statistics: A Bayesian Perspective*, Duxbury, London.

Best, NG, Spiegelhalter, DJ, Thomas, A and Brayne, CEG (1996). Bayesian analysis of realistically complex models. *J R Statist Soc A*, **159**, 323–342.

Breslow, N (1990). Biostatistics and Bayes. *Statistical Science*, **5**, 269–298.

Brooks, SP (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69-100.

Brooks, SP and Gelman, A (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434-455.

Casella, G and George, EI (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

Congdon, P (2001) Bayesian statistical modelling. Wiley.

Cowles, MK and Carlin, BP (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Dempster, A (1998). Bayesian methods. In *Encyclopedia of Biostatistics*, (eds. P Armitage and T Colton). Wiley, Chichester, pp. 263–271.

Diggle, P (1988). An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.

Diggle, P, Moyeed, R, Rowlingson, B and Thomson, M (2002). Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Applied Statistics*, **51**, 493–506.

DerSimonian, R and Laird, N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-188.

Dunson, D (2001). Commentary: Practical advantages of Bayesian analysis in epidemiologic data. *American Journal of Epidemiology*, **153**, 1222–1226.

Fisher, LD (1996). Comments on Bayesian and frequentist analysis and interpretation of clinical trials — comment. *Controlled Clinical Trials*, **17**, 423–34.

Gelfand, AE and Smith, AFM (1990). Sampling-based approaches to calculating marginal densities. *J Amer Statistic Assoc*, **85**, 398–409.

Gelman, A (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, to appear.

Gelman, A, Carlin, JC, Stern, H and Rubin, DB (2004). *Bayesian Data Analysis*, 2nd edition, Chapman & Hall, New York.

Greenland, S (1997). Probability logic and probabilistic induction. *Epidemiology*, **9**, 322–332.

Gustafson, P (2003). *Measurement Error and Misclasification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Chapman & Hall/CRC Press.

Kass, RE and Wasserman, L (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–70.

Lee, PM (2004). *Bayesian Statistics: An Introduction*, 3rd edition, Arnold, London.

Lilford, RJ and Braunholtz, D (1996). The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal*, **313**, 603–607.

Little RJA and Rubin DB (2002). *Statistical Analysis with Missing Data*, 2nd edition, Wiley, New Jersey.

Marshall, EC and Spiegelhalter, DJ (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, **22**, 1649–60.

O'Hagan, A (1988). *Probability: Methods and Measurement*, Chapman and Hall, London.

O'Hagan, A (2003). HSSS Model Criticism. In *Highly Structured Stochastic Systems*, (eds. PJ Green, NL Hjort, and ST Richardson). Oxford University Press, Oxford.

Richardson, S (1996). Measurement error. In *Markov chain Monte Carlo in Practice*, (eds. DJ Spiegelhalter, WR Gilks, and S Richardson). Chapman & Hall, London, pp. 401-417.

Richardson, S and Best, NG (2003). Bayesian hierarchical models in ecological studies of health-environment effects, *Environmetrics*, **14**, 129-147.

Senn, S (1997). Statistical basis of public policy — present remembrance of priors past is not the same as a true prior. *British Medical Journal*, **314**, 73.

Spiegelhalter, DJ (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society, Series C*, **47**, 115–133.

Spiegelhalter, DJ, Thomas, A, and Best, NG (1995). Computation on Bayesian graphical models. In *Bayesian Statistics 5* (eds. JM Bernardo, JO Berger, AP Dawid and AFM Smith). Oxford University Press, Oxford), pp. 407-425.

Spiegelhalter, DJ, Gilks, WR and Richardson, S (1996). *Markov chain Monte Carlo in Practice*, Chapman & Hall, London.

Spiegelhalter, DJ, Abrams, K and Myles, JP (2004). *Bayesian Approaches to Clinical Trials and Health Care Evaluation*, Wiley, Chichester.

Spiegelhalter, DJ, Best, NG, Carlin, BP, and van der Linde, A (2002). Bayesian measures of model complexity and fit (with discussion). *J Roy Statist Soc B*, **64**, 583–639.