Lecture 6.
Prior distributions

Summary

1. Introduction
2. Bivariate conjugate: normal
3. ‘Non-informative’ / reference priors
   - Discrete uniform distributions
   - Jeffreys priors
   - Location parameters
   - Proportions
   - Counts and rates
   - Scale parameters
4. Representation of informative priors
   - Elicitation
   - Data plus judgement
5. Mixture

Introduction

- The need for prior distributions should not be an embarrassment!
- Here we focus on breaking multivariate blocks into independent univariate priors
- It is quite reasonable that the prior should influence the analysis, as long as the influence is recognised and justified
- Importance of transparency and sensitivity analysis

Need to think about and understand **all** *prior assessments*

Normal data with unknown mean and unknown variance (for reference)

Suppose we have an independent sample of data

$$y_i \sim \text{Normal}(\mu, \sigma^2), \quad i = 1 \ldots n$$

where σ^2 and μ are unknown. The bivariate conjugate prior $p(\mu, \sigma^2)$ is expressed as $p(\mu, \sigma^2)p(\sigma^2)$, where

$$\mu|\sigma^2 \sim \text{Normal}(\gamma, \sigma^2/n_0)$$

where γ is assumed specified, and

$$1/\sigma^2 \sim \text{Gamma}(\alpha, \beta) :$$

we say that σ^2 has an *Inverse-Gamma* distribution, or alternatively that $\omega = 1/\sigma^2$ has a Gamma distribution.

Then after observing $\boldsymbol{y} = y_1, ..., y_n$, the joint posterior $p(\mu, \sigma^2|\boldsymbol{y})$ is expressed as $p(\mu|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y})$, where

$$p(\mu \mid \sigma^2, \boldsymbol{y}) = \text{Normal}(\gamma_n, \tau_n^2)$$

$$\text{where} \quad \gamma_n = \frac{n_0\gamma + n\overline{y}}{n_0 + n} \quad \text{and} \quad \tau_n^2 = \frac{\sigma^2}{n_0 + n}$$

and

$$1/\sigma^2|\boldsymbol{y} \sim \text{Gamma}(\alpha_n, \beta_n)$$

where $\alpha_n = \alpha + \frac{n}{2}$, $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^n (y_i - \overline{y})^2 + \frac{n_0 n(\overline{y}-\gamma)}{n_0+n}$

- Nice conjugate result, but prior dependence of μ and σ^2 may be unrealistic
- We shall see that assuming independence gives attractive results

Revision of a useful result in distribution theory

If $Y \sim \text{Normal}(\mu, \sigma^2/\lambda)$, $\lambda \sim \chi_v^2/v$, then

$$(Y - \mu)/\sigma \sim t_v.$$

Prove it! This is for a sampling distribution, but can use the same result for parameter distributions (taking care with changes in notation).

Can therefore show that the marginal prior distribution for μ in the bivariate conjugate prior is such that $(\mu - \gamma)\sqrt{n_0\alpha/\beta} \sim t_{2\alpha}$ with a corresponding result for the posterior, so that

$$(\mu - \gamma_n)\sqrt{(n_0 + n)\alpha_n/\beta_n} \sim t_{2\alpha_n}$$

Suppose we try to be 'uninformative' by letting $n_0, \alpha, \beta \rightarrow 0$. Then we get that

$$(\mu - \overline{y})/\sqrt{s^2/n} \sim t_n$$

where $s^2 = \sum_{i=1}^n (y_i - \overline{y})^2/n$. So looks like the 'classical' result, but *with no loss of a degree of freedom*.

Will see how to deal with this later.

'Non-informative' / reference priors

- In some circumstances would like to minimise judgemental input
- There has been a long search for an 'off-the-shelf' *objective* prior to use in all circumstances
- Does not exist, although useful guidance exists (Berger, 2006)
- Sometimes use improper priors (that do not integrate to 1)
- OK if lead to well-behaved posterior distributions
- Care needed in certain circumstances - just because 'proper' does not mean not influential
- If the form of 'non-informative' prior matters, then you should not be trying to be non-informative!

Discrete uniform distributions

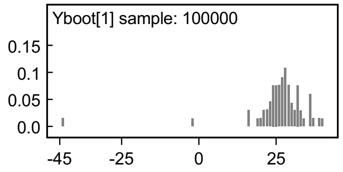
Bootstrapping in BUGS *: the Newcomb data*

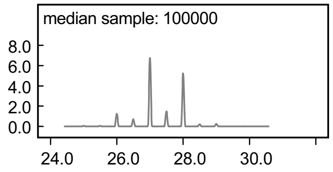
Gelman et al (2004) reanalyse 66 measurements made of the time taken for light to travel 7442 metres (recorded as deviations from 24800 nanoseconds), made by Simon Newcomb in 1882.

64 of these form a fairly symmetric distribution between around 17 and 40, while there are two gross outliers of -2 and -44.

We can adopt the basic bootstrap procedure of taking a series of repeat samples without replacement, and calculating the sample mean and median for each of these repeats

```
for(i in 1:N){p[i] <- 1/N} # set up uniform prior on 1 to N
for(j in 1:N){
  pick[j]   ~ dcat(p[])    # pick random number between 1 and N
  Yboot[j] <- Y[pick[j]]   # set jth bootstrap observation
}
mean  <- mean(Yboot[])
# find median of bootstrap sample, as N is even, this is halfway
# between the N/2 and N/2 +1 observation
n1 <- N/2 ; n2 <- n1 + 1
median <- (ranked(Yboot[],n1) + ranked(Yboot[],n2) )/2
```





(True value for the speed of light corresponds to 33, well outside the 95% bootstrap interval

The problem with uniform priors for continuous parameters

- Tempting to adopt a uniform prior for all θ
- But this does not generally imply a uniform distribution for a function of θ
- *eg* $\theta =$ chance a (biased) coin comes down heads, assume $\theta \sim$ Uniform$(0, 1)$
- Let $\phi = \theta^2 =$ chance of it coming down heads in both of the next 2 throws
- $p(\phi) = 1/(2\sqrt{\phi})$: a beta(0.5, 1) distribution and is certainly not uniform.

Jeffreys priors

- Harold Jeffreys (1939) suggested *invariant* prior distributions
- *ie* a 'Jeffreys' prior for θ would be formally compatible with a Jeffreys prior for any 1-1 transformation $\phi = f(\theta)$
- $p_J(\theta) \propto I(\theta)^{1/2}$ where $I(\theta)$ is Fisher information for θ

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

- Jeffreys' prior is invariant to reparameterisation because

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

and so

$$p_J(\phi) = p_J(\theta)\left|\frac{d\theta}{d\phi}\right|$$

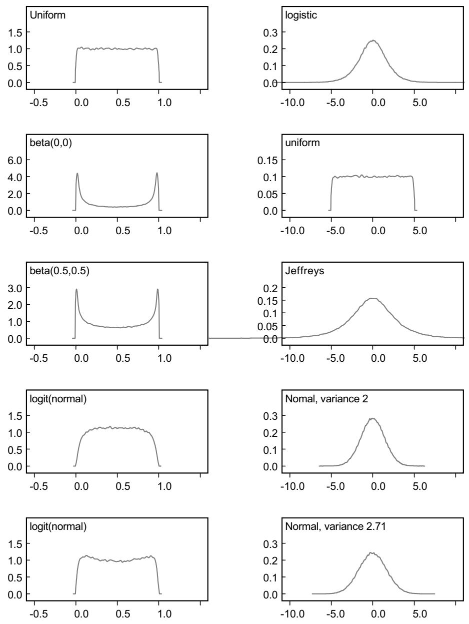
Location parameters

- Location parameter θ: $p(y|\theta)$ is a function of $y - \theta$, and so the distribution of $y - \theta$ is independent of θ
- $p_J(\theta) \propto$ constant
- In BUGS could use `dflat()` to represent this distribution
- Tend to use proper distributions such as `dunif(-100,100)` or `dnorm(0,0.0000001)`
- We recommend the former with appropriately chosen limits

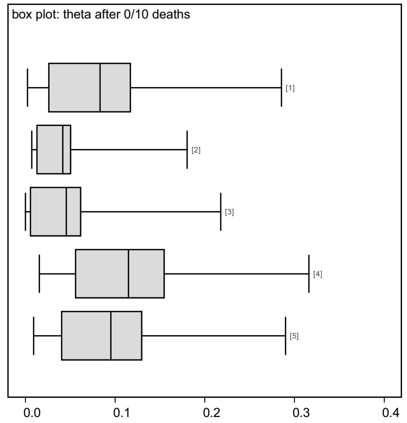
Proportions

The appropriate prior distribution for the parameter θ of a Bernoulli or Binomial distribution is one of the oldest problems in statistics

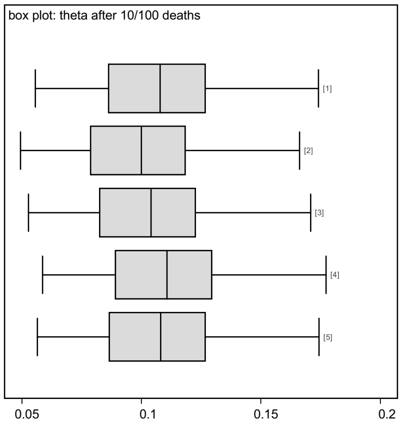
1. Bayes and Laplace suggesting a uniform prior, which is also a Beta(1, 1) (logistic on $\phi = \text{logit}\theta$): *Principle of Insufficient Reason*, is that it leads to a discrete uniform distribution for the predicted number y of successes in n future trials, so that $p(y) = 1/n, y = 0, 1, ..., n$.
2. An (improper) uniform prior on $\phi = \text{logit}\theta$ is formally equivalent to the (improper) Beta$(0, 0)$ distribution, where $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$
3. Jeffreys principle leads to a Beta$(0.5, 0.5)$ distribution, so that $p_J(\theta) = \pi^{-1}\theta^{\frac{1}{2}}(1-\theta)^{\frac{1}{2}}$
4. $\phi \sim$ Normal$(0, 2)$ gives a density for θ that is ‘flat’ at $\theta = 0.5$
5. $\phi \sim$ Normal$(0, 2.71)$ is close to a standard logistic distribution.



Suppose we observe 0/10 deaths. What is sensitivity to prior distribution on mortality rate?



And if observe 10/100 deaths?



Counts and rates

- Fisher information for Poisson data is $I(\theta) = 1/\theta$
- Jeffreys prior : $p_J(\theta) \propto \theta^{-\frac{1}{2}}$ (improper)
- Can be approximated in BUGS by a `dgamma(0.5, 0.00001)` distribution
- The same prior is appropriate if θ is a rate parameter per unit time, so that $Y \sim$ Poisson(θt).

Scale parameters

- σ is a scale parameter if $p(y|\sigma) = \sigma^{-1} f(y/\sigma)$ for some function f, so that the distribution of Y/σ does not depend on σ
- Jeffreys prior is $p_J(\sigma) \propto \sigma^{-1}$
- Implies that $p_J(\sigma^k) \propto \sigma^{-k}$, for any choice of power k
- Thus for the normal distribution, parameterised in BUGS in terms of the precision $\omega = 1/\sigma^2$, would have $p_J(\omega) \propto \omega^{-1}$
- can be approximated in BUGS by, say, a `dgamma(0.001,0.001)`, which also can be considered an inverse-gamma distribution on the variance σ^2
- Alternatively, we note that the Jeffreys prior is equivalent to $p_J(\log \sigma^k) \propto$ const, i.e. an improper uniform prior
- Hence it may be preferable to give $\log \sigma^k$ a uniform prior on a suitable range, for example `omega ~ dunif(-10, 10)` for a Normal precision

Multiparameter situations

- Multivariate Jeffreys can lead to unfortunate results
- Normal with unknown mean and variance: Jeffreys rule applied directly gives $p_J(\mu, \sigma^2) \propto 1/\sigma^3$, and leads to result shown earlier in which limiting dependent conjugate analysis does not lose degree of freedom in t posterior.
- Jeffreys suggested imposing location/scale independence and assessing univariate priors, so that

$$p_J(\mu, \sigma^2) = p_J(\mu) p_J(\sigma^2) \propto 1/\sigma^2.$$

- Then we can show that we match 'classical' analysis using t_{n-1} degrees of freedom.

Other situations

- Sampling distribution Uniform$(0, \theta)$, $p(y|\theta) = 1/\theta,\ 0 < y < \theta$, non-standard since range depends on parameter, think of θ as scale parameter, Jeffreys prior $p_J(\theta) \propto 1/\theta$.
- Jeffreys suggests (more informally) $p(\theta) \propto 1/\theta$ for other parameters restricted to $(0, \infty)$
- Care needed in handling variances for random-effects (see later)
- Berger and Bernardo have a theory of multivariate *reference* priors which may require an ordering of importance of the parameters
- Yang and Berger (1997) provide a ‘catalog of Noninformative’ priors
- Can be very complex and often no clear ‘standard’

Representation of informative priors

Pure elicitation

- Elicitation of subjective probability distributions is not a straightforward task
- Many well-known biases have been identified
- O’Hagan et al (2006) provide some ’*Guidance for best practice*’
- Emphasise that probability assessments are constructed by the questioning technique, rather than being ’*pre-formed quantifications of pre-analysed belief*’ [p 217]

- best to interview subjects face-to-face, with feedback and continual checking for biases
- conduct sensitivity analysis to the consequence of the analysis,
- avoid verbal descriptions of uncertainty.
- elicit intervals with moderate rather than high probability content, say by focussing on 33% and 67% quantiles
- use multiple experts with reporting a simple average,
- acknowledge imperfections in the process, and that even genuine ‘expertise’ cannot guarantee a suitable subject

Or simply ask for an interval and afterwards elicit a ‘confidence’ in that assessment *eg* in *Elicitor* (Kynn, 2006)

- Advantage to use conjugate forms where the prior distribution can then be interpreted as representing ‘implicit data’
- *ie* prior estimate of the parameter and an ‘effective prior sample size’
- even possible to include the prior information as ‘data’ and use standard software for statistical analysis
- For regression coefficients generally appropriate to assume a Normal distribution:
- For log-odds-ratios in logistic regression, log-rate-ratios in Poisson regression, log-hazard-ratios in Cox regression: prior variance is approximately $4/n_0$, where n_0 is the *effective number of events*

A simple moment-based method:

- ask either directly for the mean and standard deviation,
- or elicit an approximate 67% interval (*i.e.* the parameter is assessed to be twice as likely to be inside the interval as outside it)
- treat the interval as representing the mean ± 1 standard deviation
- solve for the parameters of the prior distribution

Good practice to iterate between alternative representations of the prior distribution, say as a drawn distribution, percentiles, moments, and interpretation as 'implicit data', in order to check the subject is happy with the implications of their assessments

Power calculations

- randomised trial is planned with n patients in each of two arms
- response within each treatment arm is assumed to have between-patient standard deviation σ
- treatment estimate Y assumed to have a Normal$(\theta, 2\sigma^2/n)$ distribution
- trial designed to have two-sided Type I error α and Type II error β in detecting a true difference of θ in mean response between the groups will require a sample size per group of

$$n = \frac{2\sigma^2}{\theta^2}(z_{1-\beta} + z_{1-\alpha/2})^2,$$

- Alternatively, for fixed n, the power of the study is

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

If we assume $\theta = 5, \sigma = 10$, $\alpha = 0.05, \beta = 0.10$, so that the power of the trial is 90%, then we obtain $z_{1-\beta} = 1.28$, $z_{1-\alpha/2} = 1.96$, $n = 84$.

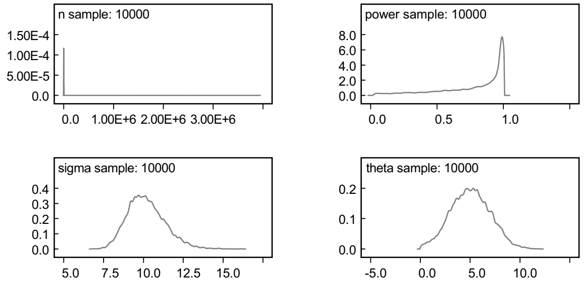
Wish to acknowledge uncertainty about θ and σ.

1. assume past evidence suggests θ is likely to lie anywhere between 3 and 7, which we interpret as a 67% interval and so assume $\theta \sim$ Normal$(5, 2^2)$
2. Remember that

$$\omega \sim \text{Gamma}(n_0/2, n_0\hat{\sigma}_0^2/2)$$

3. assess our estimate of $\sigma = 10$ as being based on around 40 observations, from which we assume a Gamma(a, b) prior distribution for $\omega = 1/\sigma^2$ with mean $a/b = 1/10^2$ and effective sample size $2a = 40$, from which we derive $\omega \sim$ Gamma$(20, 2000)$

```
omega   ~ dgamma(20,2000)   #  prior variance estimate 100, based on 40 observations
sigma  <- 1/sqrt(omega)
theta   ~ dnorm(5,0.25) # prior mean estimate 5, with sd 2
n      <- 2 * pow( (1.28 +1.96) * sigma / theta , 2)# sample size for 90% power
power  <- phi( sqrt(84/2)* theta /sigma  -1.96)     # power  for n = 84
prob70 <- step(power-.7)                            # probability that power is greater
```



```
node     mean    sd      MC error  2.5%    median  97.5%   start  sample
n        1841.0  58530.0 601.5     24.56   86.26   1463.0  1   10000
power    0.7788  0.2574  0.00218   0.1186  0.8921  1.0     1   10000
```

Median power for $= 84$ is 90%, with 30% probability that power is less than 70%

Median n for 90% power is 86, but with huge uncertainty

Priors based on Data + judgement

- have historical data and we could obtain a prior distribution for θ based only on an empirical estimate $\widehat{\theta}_H$ and its estimated standard error.
- Direct use would essentially be pooling the data
- May prefer to *downweight* historical data, maybe due to limits in *relevance* and *rigour*

Two basic methods:

1. *Power prior:*
2. *Bias modelling:*

Power prior: (Ibrahim and Shen, 2000)

this discounts the ‘effective prior sample size’ by a factor κ, *eg*

- a fitted Beta(a, b) would become a Beta$(\kappa a, \kappa b)$,
- a Gamma(a, b) would become a Gamma$(\kappa a, \kappa b)$,
- a Normal(γ, τ^2) would become a Normal$(\gamma, (\tau/\kappa)^2)$
- only increases variability

Bias modelling:

- assume that $\theta = \theta_H + \delta$, where δ is an (unknown) bias
- assume $\delta \sim [\mu_\delta, \sigma_\delta^2]$, where $[,]$ indicates a mean and variance but otherwise unspecified distribution
- assume the historical data gives a prior $\theta_H \sim [\widehat{\theta}_H, \tau_H^2]$
- then we obtain a prior distribution for θ of

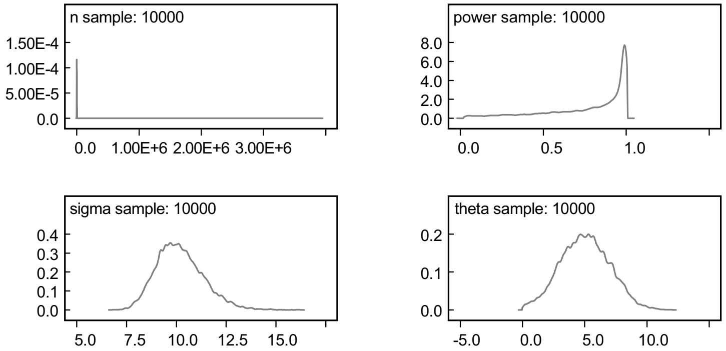
$$\theta \sim [\gamma_H + \mu_\delta, \tau_H^2 + \sigma_\delta^2]$$

- thus prior mean is shifted and prior variance is increased

Power calculations (continued)

Power with some discounted prior distributions.

```
#omega ~ dgamma(20,2000)
 omega ~ dgamma(10,1000)  #discounted by 2 (k = 0.5)
#theta ~ dnorm(5,0.25)
 theta ~ dnorm(4,0.125)I(0,)  # discounted by 4 (k=0.25) and shifted down b
                              # constrained to be > 0
```



```
node      mean      sd         MC error      2.5%      median     97.5%    start  sample
n      4.542E+6  4.263E+8    4.26E+6      20.96      125.6     14270.0   1    10000
power    0.6536  0.3315     0.003406     0.04353    0.7549      1.0      1    10000
```

Huge uncertainty

46% probability that power is less than 70%

Mixture of prior distributions

- Suppose we doubt which of two or more prior distributions is appropriate to the data in hand
- *eg* might suspect that *either* a drug will produce similar effect to other related compounds, *or* if it doesn't behave like these compounds we are unsure about its likely effect
- For two possible prior distributions $p_1(\theta)$ and $p_2(\theta)$ the overall prior distribution is then a *mixture*

  $$p(\theta) = qp_1(\theta) + (1 - q)p_2(\theta),$$

  where q is the assessed probability that p_1 is 'correct'.

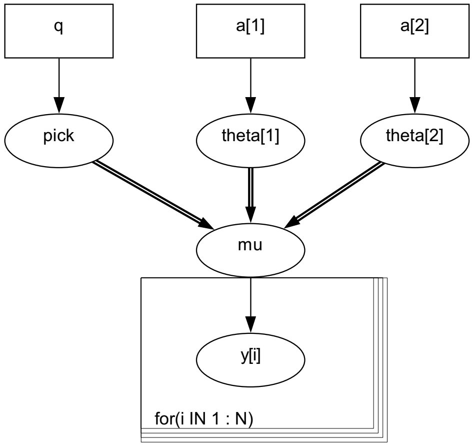
- If we now observe data y, it turns out that the posterior for θ is

  $$p(\theta|y) = q'p_1(\theta|y) + (1 - q')p_2(\theta|y)$$

  where

  $$\begin{aligned} p_i(\theta|y) &\propto p(y|\theta)p_i(\theta) \\ q' &= \frac{qp_1(y)}{qp_1(y) + (1-q)p_2(y)}, \end{aligned}$$

  where $p_i(y) = \int p(y|\theta)p_i(\theta)d\theta$ is the predictive probability of the data y assuming $p_i(\theta)$
- The posterior is a mixture of the respective posterior distributions under each prior assumption, with the mixture weights adapted to support the prior that provides the best prediction for the observed data.



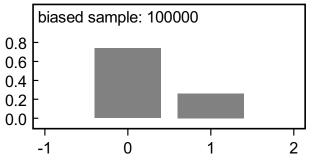
A biased coin?

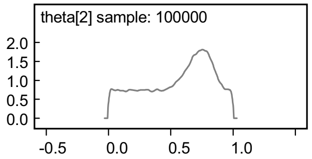
Suppose a coin is either unbiased or biased, in which case the chance of a 'head' is unknown. We assess a probability of 0.1 that it is biased, and then observe 15 heads out of 20 tosses — what is chance that coin is biased?

```
q[1]     <- 0.9; q[2] <- 0.1 # prior assumptions
y        <- 15; n <- 20      # data
y         ~ dbin(p, n)       # likelihood
p        <- theta[pick]      # could have included theta[pick] directly in dbin
pick      ~ dcat(q[])        # pick = 1 or 2

theta[1] <- 0.5              # if unbiased (assumption 1)
theta[2]  ~ dunif(0, 1)      # if biased, then uniform prior on prob of head

biased   <- pick-1           # 1 if biased, 0 otherwise
```

```
node          mean    sd     MC error  2.5%    median  97.5%   start  sample
biased       0.2619  0.4397  0.002027   0.0     0.0     1.0     1    100000
theta[2]     0.5594  0.272   9.727E-4   0.03284 0.6247  0.9664  1    100000
```

So the probability that the coin is biased has increased from 0.1 to 0.26 on the basis of the evidence provided.





The rather strange shape of the posterior distribution of `theta[2]` is explained below.

- 'Pick' is a variable taking on value 1 when first component is true, 2 if second
- But when `pick=1`, `theta[2]` is sampled from its prior distribution (Carlin and Chib, 1995)
- So posterior distribution of `theta[2]` is mixture of true posterior and its prior
- Could do separate run assuming each component true
- Or only use those values simulated when `pick=2` (need to sort outside WinBUGS)

- Essentially dealing with alternative model formulations
- q''s correspond to posterior probabilities of models
- Well-known difficulties with these quantities both in theory when calculating within MCMC
- In principle we can use the structure above to handle a list of arbitrary alternative models, but in practice considerable care is needed if the sampler is not to be go ‘off course’ when sampling from the prior distribution at each iteration when that model is not being ‘picked’
- It is possible to use ‘pseudo-priors’ to be used in these circumstances, where `pick` also dictates the prior to be assumed for `theta[j]` when `pick` $\neq j$ (Carlin and Chib, 1995)