# 3. Learning from data using MCMC and BUGS

**Gianluca Baio**

Department of Statistical Science | University College London

g.baio@ucl.ac.uk
https://gianluca.statistica.it/
https://egon.stats.ucl.ac.uk/research/statistics-health-economics/
https://github.com/giabaio
https://github.com/StatisticsHealthEconomics
@gianlubaio

STAT0019 - Bayesian Methods in Health Economics, UCL

UCL Statistical Science

# Summary

- Already seen how to make predictions based on parameters with *known* uncertainty distributions
- Here we **learn** about parameters from **observed data** using *Bayes theorem*
- Simple example – trial with binary outcome
  - *"conjugate"*: no simulation needed
- Simulating posterior distributions of unknowns given data, using **MCMC (Markov Chain Monte Carlo)** in BUGS
  - Practical issues: putting data in
  - Convergence (knowing which / how many simulations to save)
- Expressing full uncertainty on any function / transformation of unknown parameters

## References


*The BUGS Book*, chapters 3, 4 Library Book website

*Bayesian Methods in Health Economics*, chapters 2, 4 Library Book website (CRC) Book website Code

## Updating beliefs with new evidence

- External evidence about unknown quantities $\theta$ **that is not based on current data** is expressed as a **prior** probability distribution $p(\theta)$
- Evidence from available data $y$ expressed as **sampling distribution** $p(y \mid \theta)$

## Updating beliefs with new evidence

- External evidence about unknown quantities $\theta$ **that is not based on current data** is expressed as a **prior** probability distribution $p(\theta)$
- Evidence from available data $y$ expressed as **sampling distribution** $p(y \mid \theta)$

Two sources of evidence combined using **Bayes theorem**:

$$p(\theta \mid y) = p(\theta) \times \frac{p(y \mid \theta)}{p(y)}$$

which is essentially

$$p(\theta \mid y) \propto p(\theta) \times p(y \mid \theta)$$

**Posterior $\propto$ Prior $\times$ Likelihood**

... Posterior becomes your prior when next piece of evidence arrives...

## Practical benefits

- Ability to **synthesise** multiple datasets / sources of evidence in coherent manner
- ... Allows complexities about real-world data to be modelled (via MCMC methods)
- Naturally provides **predictions** of future events
- Full allowance for **uncertainty** in conclusions
- **Intuitive** communication

  – express uncertainty by probability statements about unknowns

## Practical benefits

- Ability to **synthesise** multiple datasets / sources of evidence in coherent manner
- ... Allows complexities about real-world data to be modelled (via MCMC methods)
- Naturally provides **predictions** of future events
- Full allowance for **uncertainty** in conclusions
- **Intuitive** communication

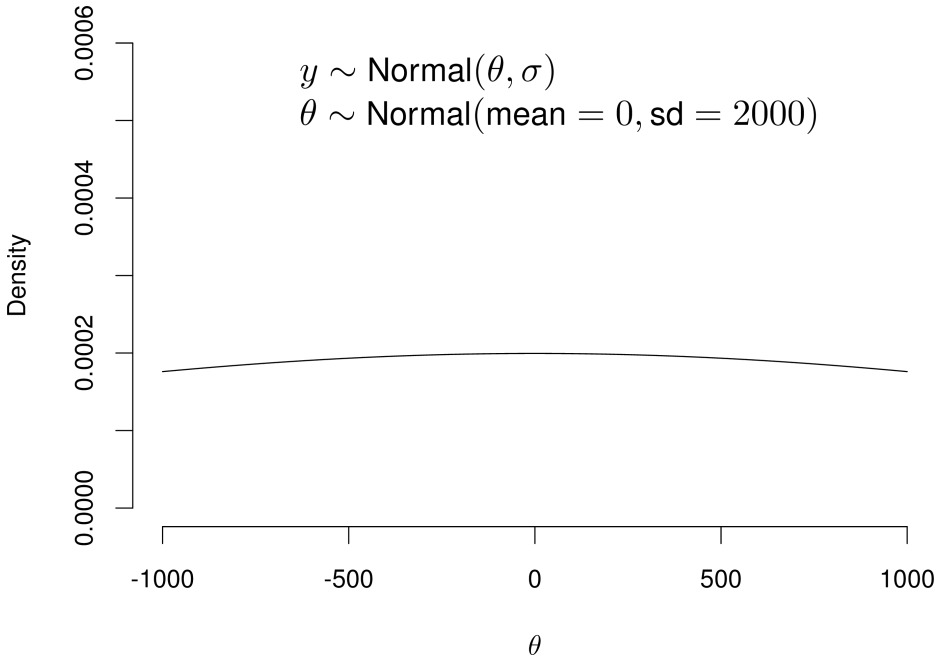
  – express uncertainty by probability statements about unknowns

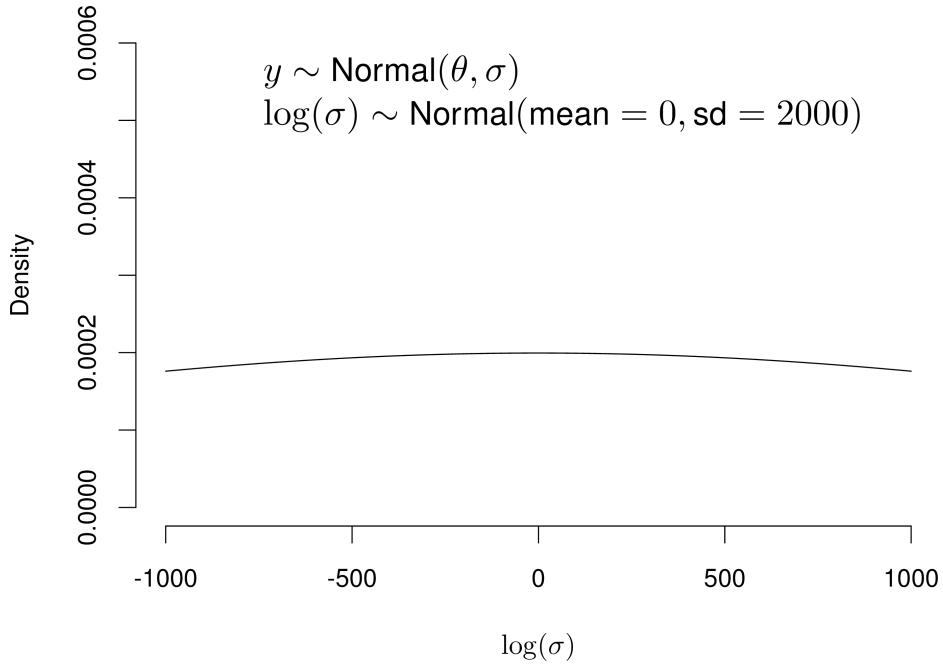
## Challenges

- Harder to implement than classical "frequentist" methods
- Extra source of information (the prior) – can be tricky to specify...

# Choice of prior distributions

- **"Vague" priors**: typically distributions with big variances (*on a suitable scale*!), eg `mu ~ dnorm(0, 0.00001)` (**NB**: precision=0.00001 $\Rightarrow$ variance=100000!)

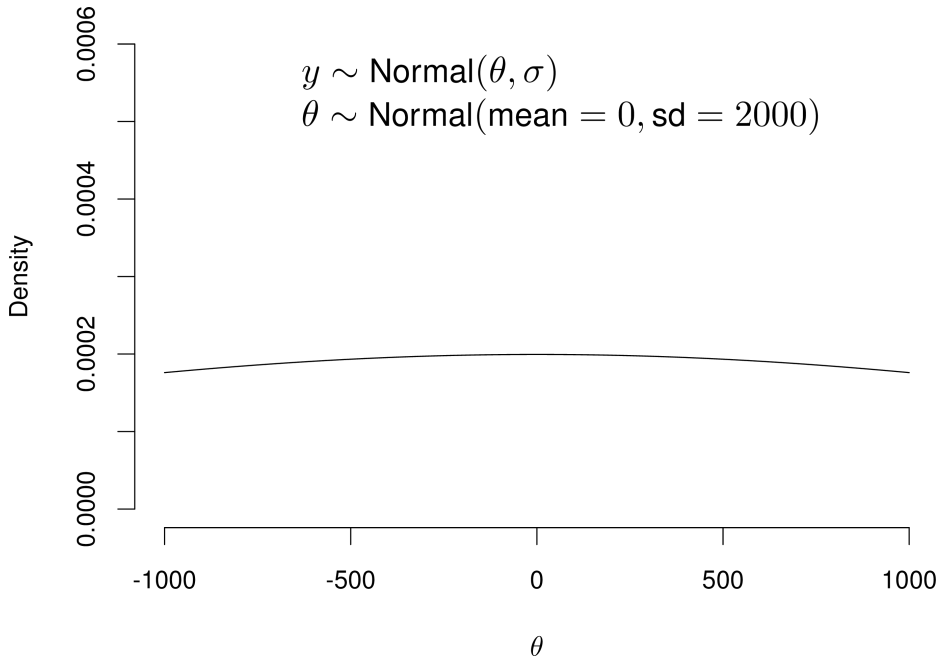


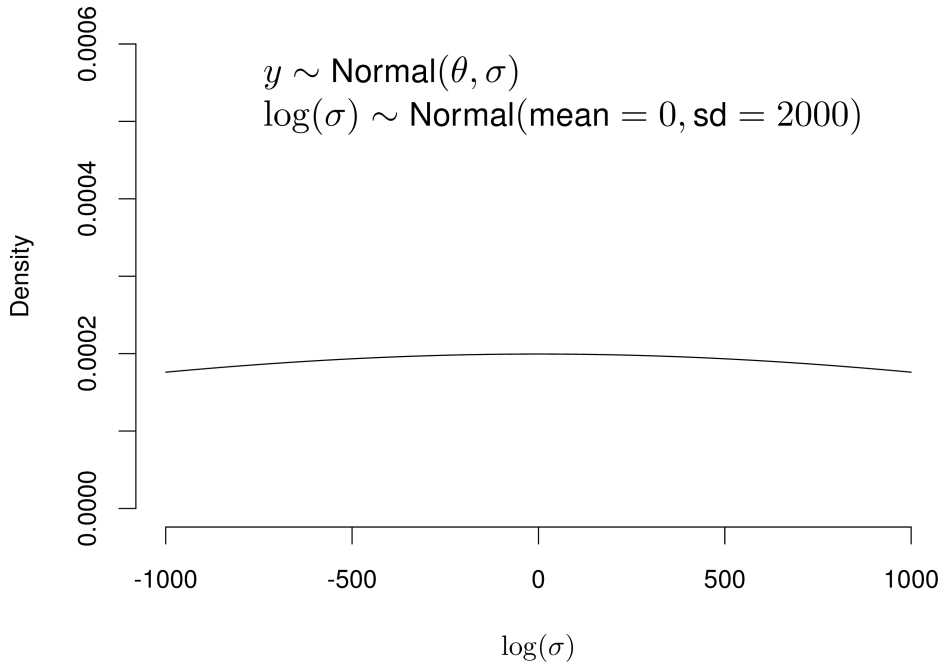




- **NB**: Beware of the implications of your prior – are you assuming too much (unrealistic) variance?

# Choice of prior distributions

- **"Vague" priors**: typically distributions with big variances (*on a suitable scale*!), eg `mu ~ dnorm(0, 0.00001)` (**NB**: precision=0.00001 $\Rightarrow$ variance=100000!)




- **NB**: Beware of the implications of your prior – are you assuming too much (unrealistic) variance?
- More recent proposal: **Penalised Complexity** (PC) Priors
  - Use "default" distributional assumptions; penalise deviations from (= added complexity in comparison to) a simpler, base model – can be very hard to think about and construct
  - (**Very** technical) paper

# PC Priors (**optional**)

- Regularise inference while not forcing too strong information
- Penalise departure from a "base" model (eg parameter = some fixed value)
  - Prior tends to favour the base model $\Rightarrow$ need fairly strong evidence to move away from it
  - Distance between the **base** model $g(\xi)$ and an **alternative**, more complex model $f(\xi)$ is measured by

$$d(f,g) = \sqrt{2\text{KLD}(f,g)} \qquad \text{with} \qquad \text{KLD}(f,g) = \int f(\xi)\log\left(\frac{f(\xi)}{g(\xi)}\right) d\xi$$

# PC Priors (**optional**)

- Regularise inference while not forcing too strong information
- Penalise departure from a "base" model (eg parameter = some fixed value)
  - Prior tends to favour the base model $\Rightarrow$ need fairly strong evidence to move away from it
  - Distance between the **base** model $g(\xi)$ and an **alternative**, more complex model $f(\xi)$ is measured by

$$d(f,g) = \sqrt{2\text{KLD}(f,g)} \qquad \text{with} \qquad \text{KLD}(f,g) = \int f(\xi)\log\left(\frac{f(\xi)}{g(\xi)}\right) d\xi$$

- Penalisation done at a constant rate

$$p(d) = \lambda\exp(-\lambda d) \sim \text{Exponential}(\lambda) \qquad \Rightarrow \qquad p(\xi) = \lambda e^{-\lambda d(\xi)} \left|\frac{\partial d(\xi)}{\partial \xi}\right|$$

- PC prior defined using probability statements on the model parameters (in the appropriate scale) to determine the value of $\lambda$ using "reasonable" information

Example | Proof

PC prior for a **precision** $\tau = \sigma^{-2}$

- Base model: $\sigma = 0$
- Set $\Pr(\sigma > \sigma_0) = \alpha$, for some constants $\sigma_0$ and $\alpha$
- This implies

$$p(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda\tau^{-1/2}\right) \sim \text{type-2 Gumbel}$$

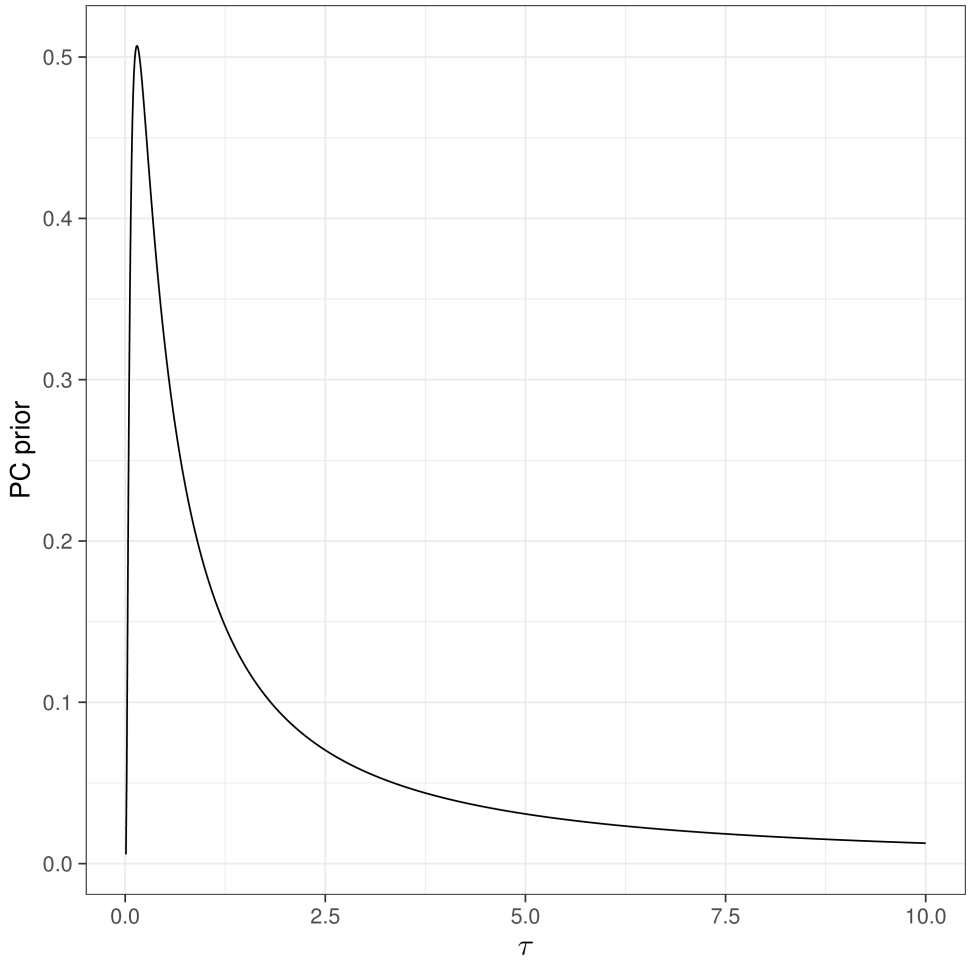
with

$$\lambda = -\frac{\log(\alpha)}{\sigma_0}$$

- **NB**: The regularising constraint and the actual prior may be defined on **different scales**!
  - In this case, the resulting prior for the standard deviation is

$$p(\sigma) \sim \text{Exponential}(\lambda)$$

eg: setting $\sigma_0 = 2$ and $\alpha = 0.1$ gives this



Example | Proof

Consider the two competing models for some parameter $\theta$ (or data $y$) as a function of a **precision** $\tau$

$$g(\tau) \sim \text{Normal}(0, \tau = \tau_0 \to \infty) \quad \text{and} \quad f(\tau) \sim \text{Normal}(0, \tau), \tau \in (0, \infty)$$

Then

- $\text{KLD}(f, g) = \frac{1}{2}\frac{\tau_0}{\tau}\left[1 + \frac{\tau}{\tau_0}\log\left(\frac{\tau}{\tau_0}\right) - \frac{\tau}{\tau_0}\right] \to \frac{1}{2}\frac{\tau_0}{\tau}$ if $\tau << \tau_0$
- $d(\tau) = \sqrt{2\text{KLD}(f, g)} = \sqrt{\frac{\tau_0}{\tau}} = \tau_0^{1/2}\tau^{-1/2}$

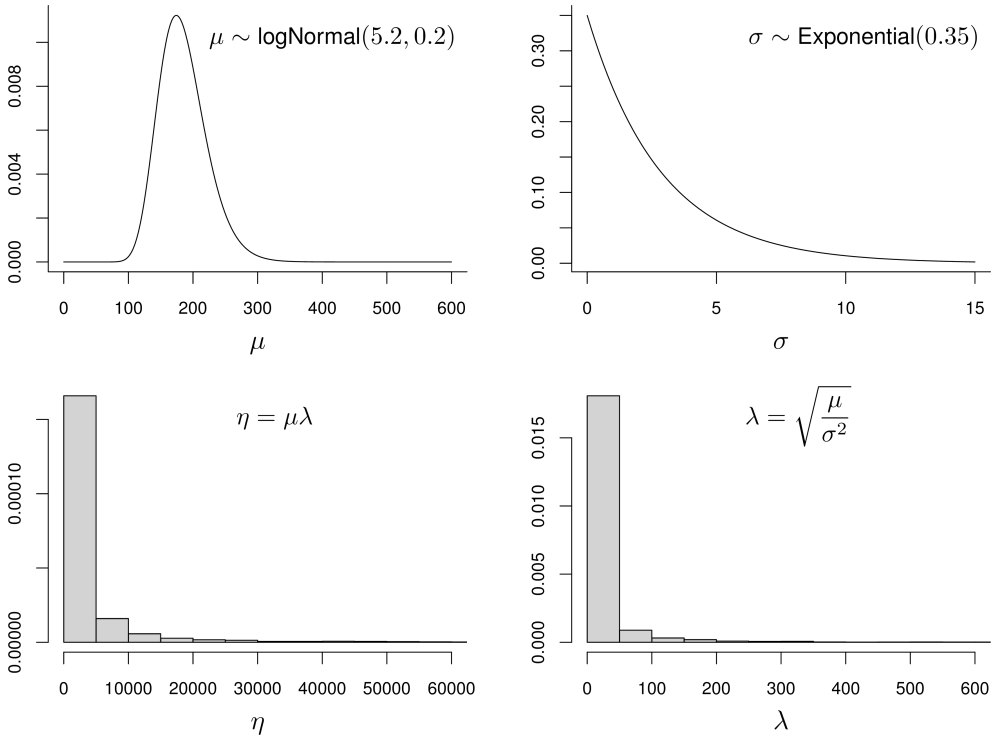
Assuming $p(d) = \lambda\exp(-\lambda d)$ then

- $\left|\frac{\partial d(\tau)}{\partial \tau}\right| = \left|-\frac{1}{2}\tau^{-3/2}\right| = \frac{1}{2}\tau^{-3/2} \quad \Rightarrow \quad p(\tau) = \lambda\exp\left[-\lambda d(\tau)\right]\left|\frac{\partial d(\tau)}{\partial \tau}\right| = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda\tau^{-1/2}\right)$
- $\left|\frac{\partial \tau}{\partial \sigma}\right| = \left|\frac{\partial \sigma^{-2}}{\partial \sigma}\right| = |-2\sigma^{-3}| = 2\sigma^{-3} \quad \Rightarrow \quad p(\sigma = \tau^{-1/2}) = \frac{\lambda}{2}\sigma^3\exp\left(-\lambda\sigma\right)\left|\frac{\partial \tau}{\partial \sigma}\right| = \lambda\exp(-\lambda\sigma)$

# Choice of prior distributions

**Informative priors**: from previous studies, or *elicited* from experts

- May be difficult to derive, so present sensitivity analysis to different choices
- If exact choice of vague prior is influential $\Rightarrow$ need more data or informative prior
- Put prior on "**natural** scale" (as opposed to "*original* scale"!) parameters
  - Easier to include genuine information – **more on this later!**
  - For instance: place a prior on $\theta \sim \text{Gamma}(\eta, \lambda)$ ...




```
> # Simulates from priors for mu and sigma
> mu=rlnorm(10000,5.2,.2)
> sigma=rexp(10000,.35)
> # Check interval estimates
> quantile(mu,c(.025,.975))
```

```
    2.5%    97.5%
122.7655 270.3207
```

```
> quantile(sigma,c(.025,.975))
```

```
      2.5%       97.5%
0.07482428 10.83392815
```

```
> # Simulates from priors for lambda and eta
> lambda=sqrt(mu/sigma^2)
> eta=mu*lambda
```

Suppose $\theta$ is the true underlying success rate (proportion) of a drug

Assume a Beta$(a, b)$ prior distribution for $\theta$

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

If we observe $r$ successes out of $n$ trials, the Binomial distribution means that

$$\text{Likelihood} \propto \theta^{r}(1-\theta)^{n-r}$$

Then by Bayes theorem

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

**Beta** prior + **Binomial** data = **Beta** posterior distribution

**For example...**

One of the Covid vaccines has been approved by the FDA on the back of a Bayesian modelling procedure based on a Binomial-Beta model)

See Lecture 2

So: $\begin{cases} \theta \sim \text{Beta}(0, 0) \\ y_0 \sim \text{Binomial}(\theta, n_0) \end{cases} \quad \Rightarrow \quad \theta \mid y_0, n_0 \sim \text{Beta}(y_0, n_0 - y_0)$

- Intuition: a $\text{Beta}(0, 0)$ prior essentially implies you have truly no knowledge whatsoever about the parameter $\theta$ (even **less** than 0 successes in 0 trials!)

See Lecture 2

So: $\begin{cases} \theta \sim \text{Beta}(0,0) \\ y_0 \sim \text{Binomial}(\theta, n_0) \end{cases} \Rightarrow \quad \theta \mid y_0, n_0 \sim \text{Beta}(y_0, n_0 - y_0)$

- Intuition: a $\text{Beta}(0,0)$ prior essentially implies you have truly no knowledge whatsoever about the parameter $\theta$ (even **less** than 0 successes in 0 trials!)

- **BUT**: this is an **improper** prior, because it does not integrate/sum to 1 (which is a fundamental property of probability distributions)

$$\theta \sim \text{Beta}(0,0) \Rightarrow \int_0^1 p(\theta \mid \alpha = 0, \beta = 0) d\theta \propto \int_0^1 \frac{1}{\theta(1-\theta)} d\theta \to \infty$$

- It is possible that when using improper priors, the posterior also does not integrate to 1, which means you **cannot** make probabilistic assessment of your output – in that case, Bayesian inference is not valid

See Lecture 2

So: $\begin{cases} \theta \sim \text{Beta}(0,0) \\ y_0 \sim \text{Binomial}(\theta, n_0) \end{cases} \Rightarrow \quad \theta \mid y_0, n_0 \sim \text{Beta}(y_0, n_0 - y_0)$

- Intuition: a $\text{Beta}(0,0)$ prior essentially implies you have truly no knowledge whatsoever about the parameter $\theta$ (even **less** than 0 successes in 0 trials!)

- **BUT**: this is an **improper** prior, because it does not integrate/sum to 1 (which is a fundamental property of probability distributions)

$$\theta \sim \text{Beta}(0,0) \Rightarrow \int_0^1 p(\theta \mid \alpha = 0, \beta = 0) d\theta \propto \int_0^1 \frac{1}{\theta(1-\theta)} d\theta \to \infty$$

- It is possible that when using improper priors, the posterior also does not integrate to 1, which means you **cannot** make probabilistic assessment of your output – in that case, Bayesian inference is not valid

- It is still OK to consider this intuition and set up to validate the idea that a Beta prior can be formed to encode a thought experiment with $y_0$ "successes" out of $n_0$ "trials"

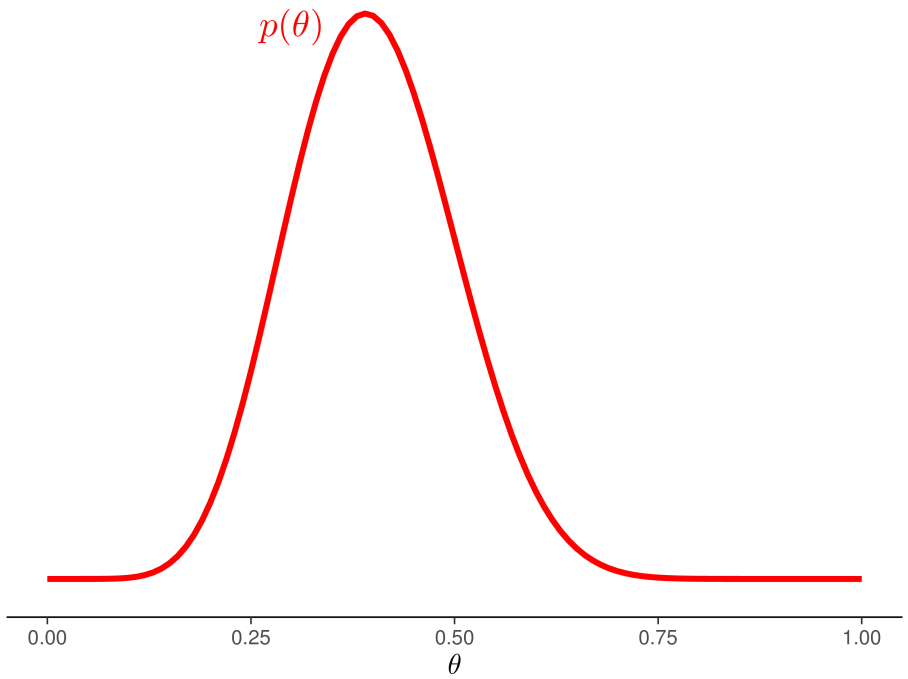
## Recap from last lecture

- Consider a drug to be given for relief of chronic pain
- Experience with similar compounds has suggested that annual response rates between 0.2 and 0.6 could be feasible
- Interpret this as a distribution with mean = 0.4, standard deviation 0.1
- $\rightarrow$ Beta(9.2,13.8) **prior** distribution

We actually do the study and **observe** 15 successes out of 20 patients

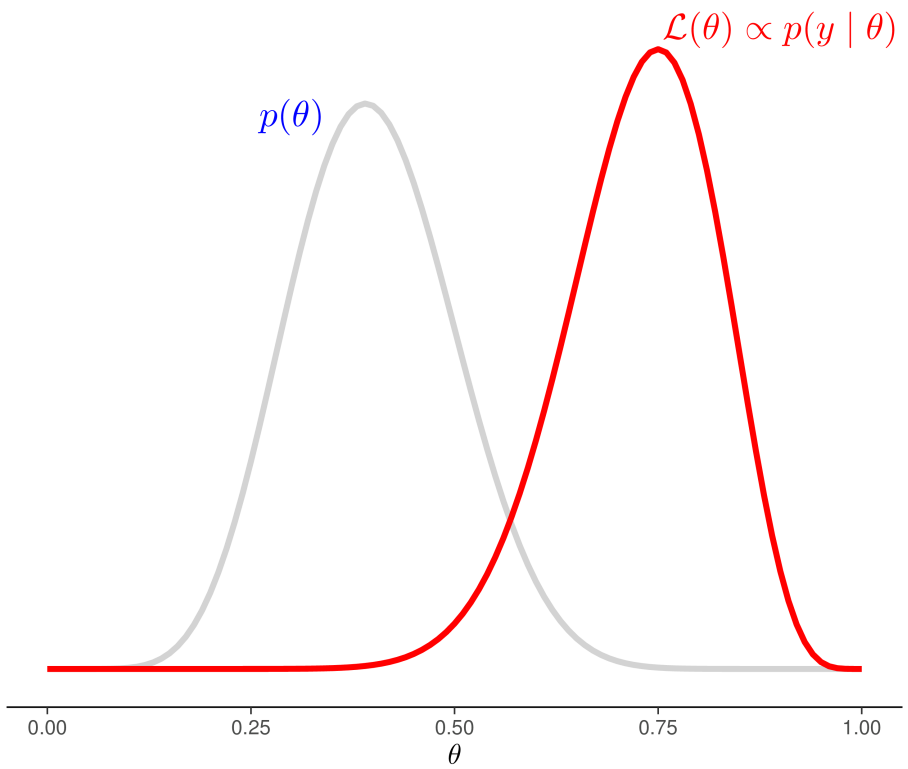
- Predict whether $>$ 25 successes in next 40 patients

## Prior distribution




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

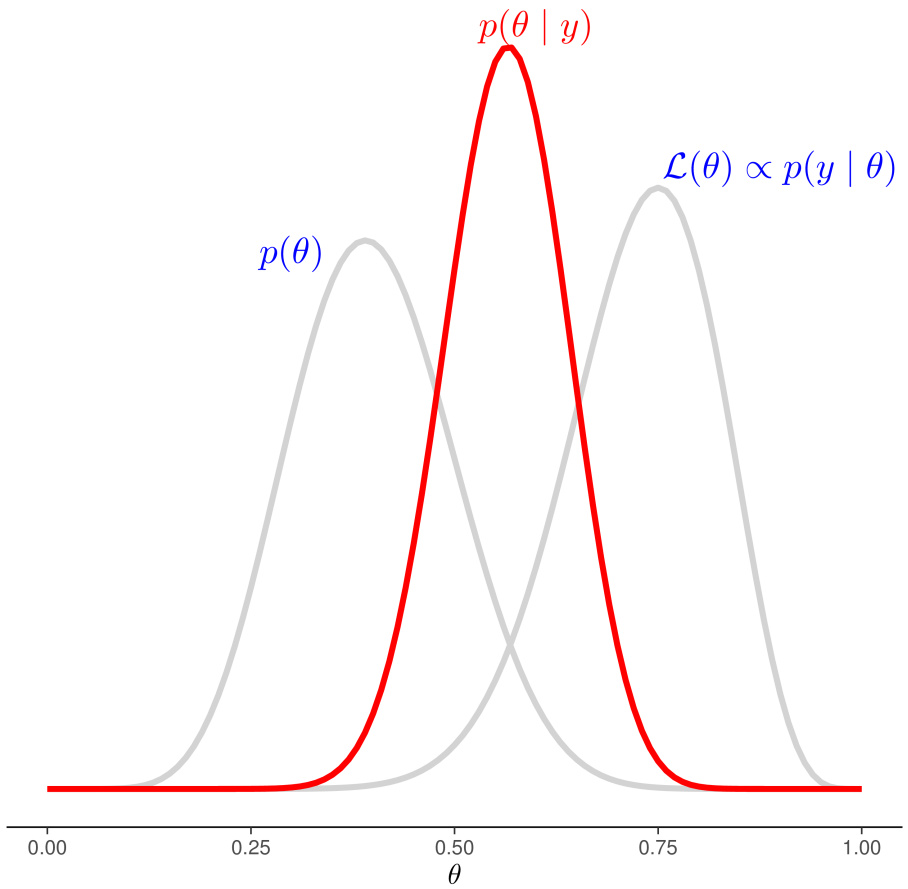
## Likelihood




Likelihood arising from a Binomial observation of 15 responders out of 20 patients given the drug

$\mathcal{L}(\theta) = \theta^{15}(1-\theta)^{(20-15)} \Rightarrow$ MLE = 15/20 = 0.75

## Posterior distribution




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

# Conjugate distributions

This is a case of **conjugate analysis**, when the posterior distribution is in the same **family** as the prior distribution

Other examples:

- Gamma prior for **rate** parameter of a Poisson likelihood (for count data, e.g. number of people arriving at emergency department)
- Normal prior for mean of a Normal likelihood
- Gamma prior for precision (1/variance) of a Normal likelihood

Advantage: don't need simulation to determine posterior

... But real situations usually more complex $\Rightarrow$ software like BUGS needed

# Drug (continued)

## Learning from data using Markov chain Monte-Carlo (MCMC) methods

Assume we observed 15 successes out of 20 subjects, and wish to predict whether we will get $> 25$ successes in next 40 patients

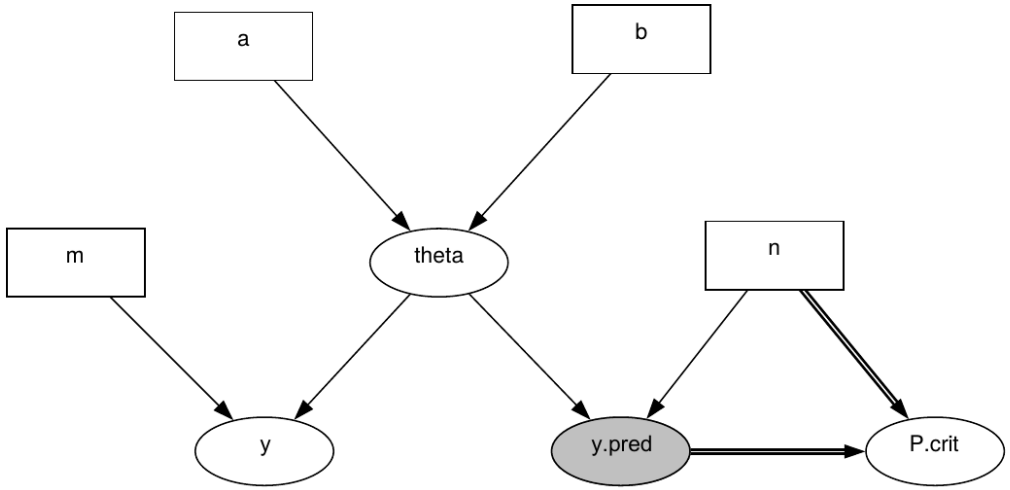
The model can be written

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

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

# (Equivalent to BUGS code)

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




- "Parent" nodes (start of arrow) generate "child" nodes (end of arrow)
  – data `y` generated by model with parameter `theta`
  – parameter `theta` generated by its "parents" `a,b`
- BUGS samples from **posterior** of `theta` – formed by combining data `y` (**likelihood**) and **prior** parameters `a,b`
- Evidence flows **up and down** arrows: Knowing about child `y` tells you about parent `theta`, just as information on `theta` used to predict child `y.pred`

## Gibbs sampling

- Uses a **Markov chain** (type of random walk – distribution for the next simulated value depends only on current value)
- Give **initial values** to parameters $\theta_1, \ldots, \theta_P$
- **Update** parameter values by repeatedly sampling from **full-conditional** posterior distribution of

$$(\theta_1 \mid \text{current } \theta_p : p \neq 1)$$

$$(\theta_2 \mid \text{current } \theta_p : p \neq 2)$$

$$\ldots$$

$$(\theta_P \mid \text{current } \theta_p : p \neq P)$$

then repeat the cycle until **convergence** (more on this later!)

- Should converge to sampling from **joint posterior** of all unknown quantities $\theta_p$ of interest
- Summarize marginal posterior of $\theta_p$ using converged sample:
  - Use sample mean as estimate of posterior mean
  - Draw smoothed **histogram** to estimate shape of posterior

## Gibbs sampling

**(Convenient) Example: semi-conjugated Normal model**

Assume

- $y_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, with $i = 1, \ldots, n \Rightarrow$ observed data
- $\mu \mid \sigma^2 \sim \text{Normal}(\mu_0, \sigma_0^2)$ $\left(\text{e.g. } \sigma_0^2 = \frac{\sigma^2}{\kappa}\right)$ and $\tau = \frac{1}{\sigma^2} \sim \text{Gamma}(\alpha_0, \beta_0)$

  for fixed $\mu_0, \sigma_0^2, \alpha_0, \beta_0$

This implies that:

- **Conditionally on** $\sigma^2$, $\mu$ has a **conjugate prior** (Normal)
- **Marginally**, $\tau$ has a **conjugate prior** (Gamma)

## Gibbs sampling

**(Convenient) Example: semi-conjugated Normal model**

Assume

- $y_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, with $i = 1, \dots, n \Rightarrow$ observed data
- $\mu \mid \sigma^2 \sim \text{Normal}(\mu_0, \sigma_0^2)$ $\left(\text{e.g. } \sigma_0^2 = \dfrac{\sigma^2}{\kappa}\right)$ and $\tau = \dfrac{1}{\sigma^2} \sim \text{Gamma}(\alpha_0, \beta_0)$

  for fixed $\mu_0, \sigma_0^2, \alpha_0, \beta_0$

This implies that:

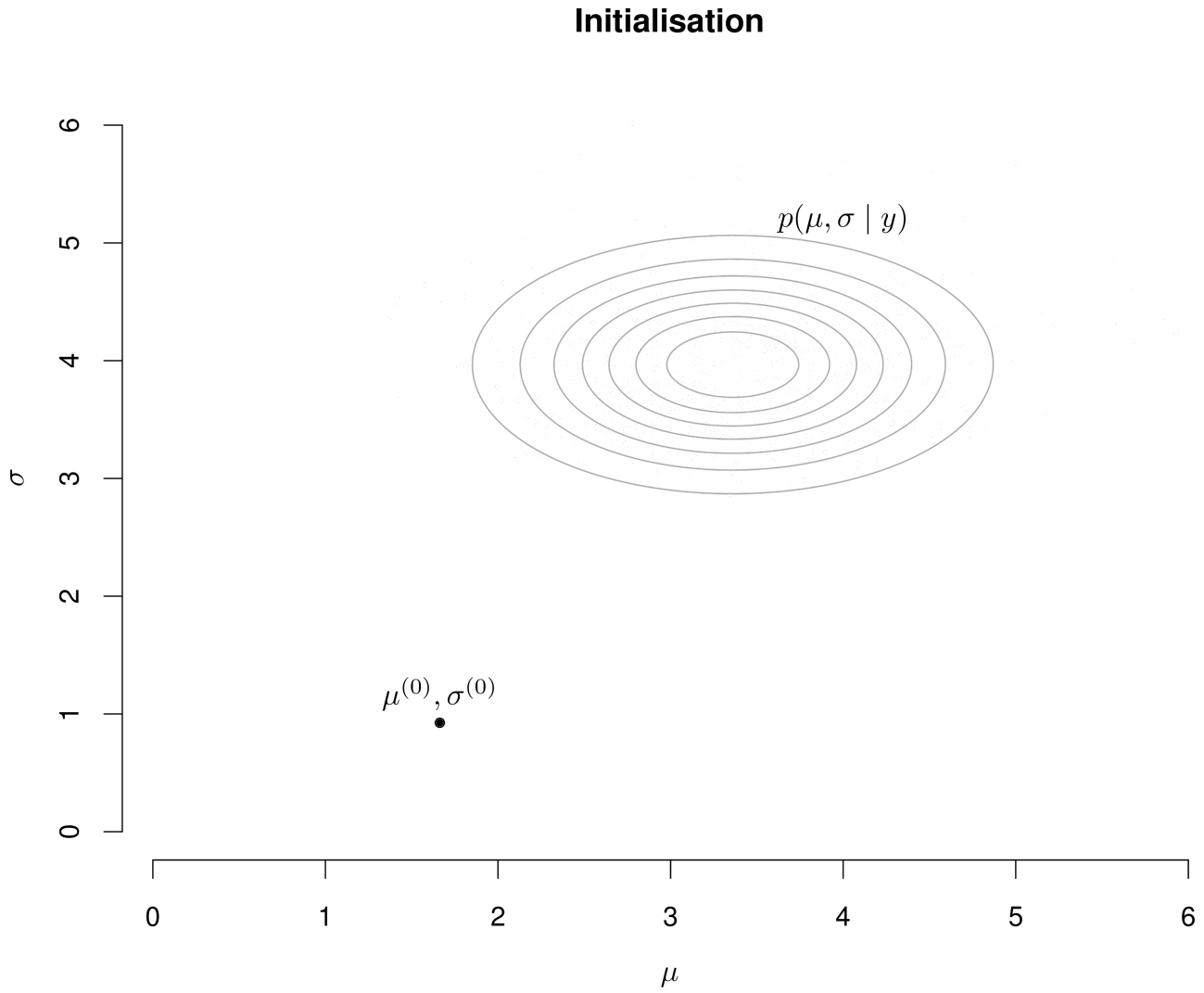
- **Conditionally on** $\sigma^2$, $\mu$ has a **conjugate prior** (Normal)
- **Marginally**, $\tau$ has a **conjugate prior** (Gamma)

Can prove that under these assumptions

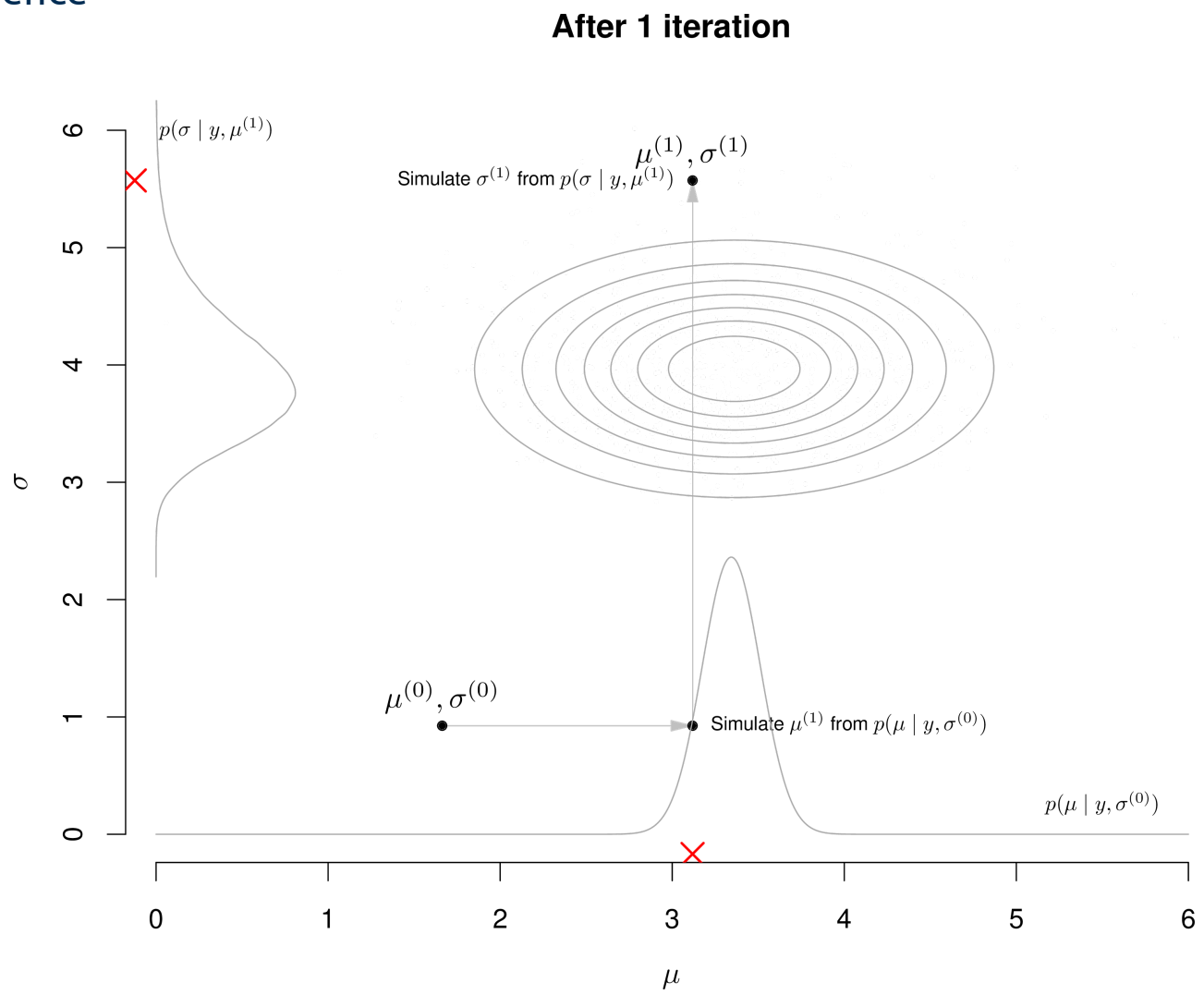
$$\mu \mid \sigma^2, \boldsymbol{y} \sim \text{Normal}(\mu_1, \sigma_1^2) \quad \text{with: } \mu_1 = \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}\right) \quad \text{and} \quad \sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

$$\tau \mid \mu, \boldsymbol{y} \sim \text{Gamma}(\alpha_1, \beta_1) \quad \text{with: } \alpha_1 = \alpha_0 + \frac{n}{2} \quad \text{and} \quad \beta_1 = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2$$

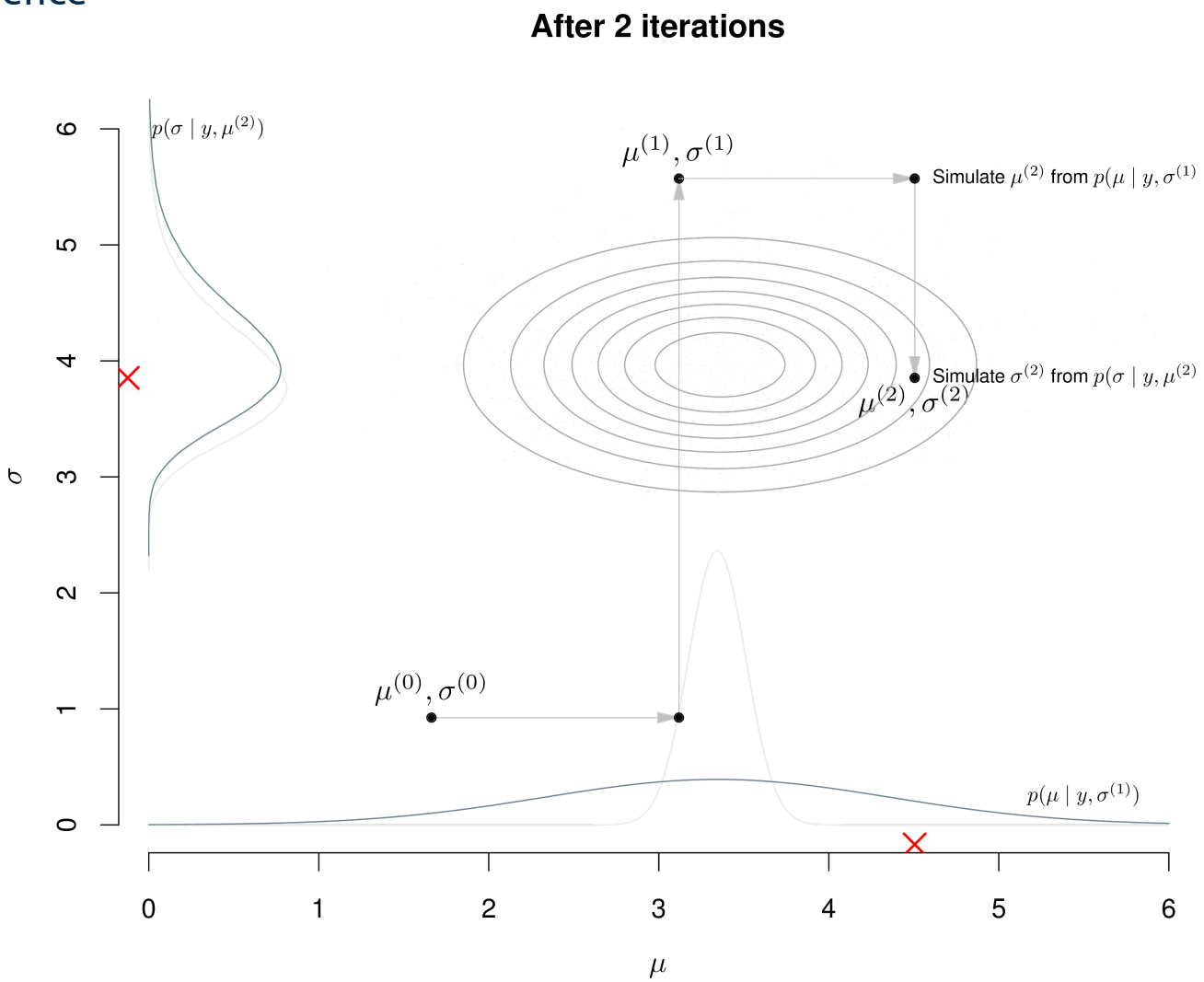
## Gibbs sampling – convergence



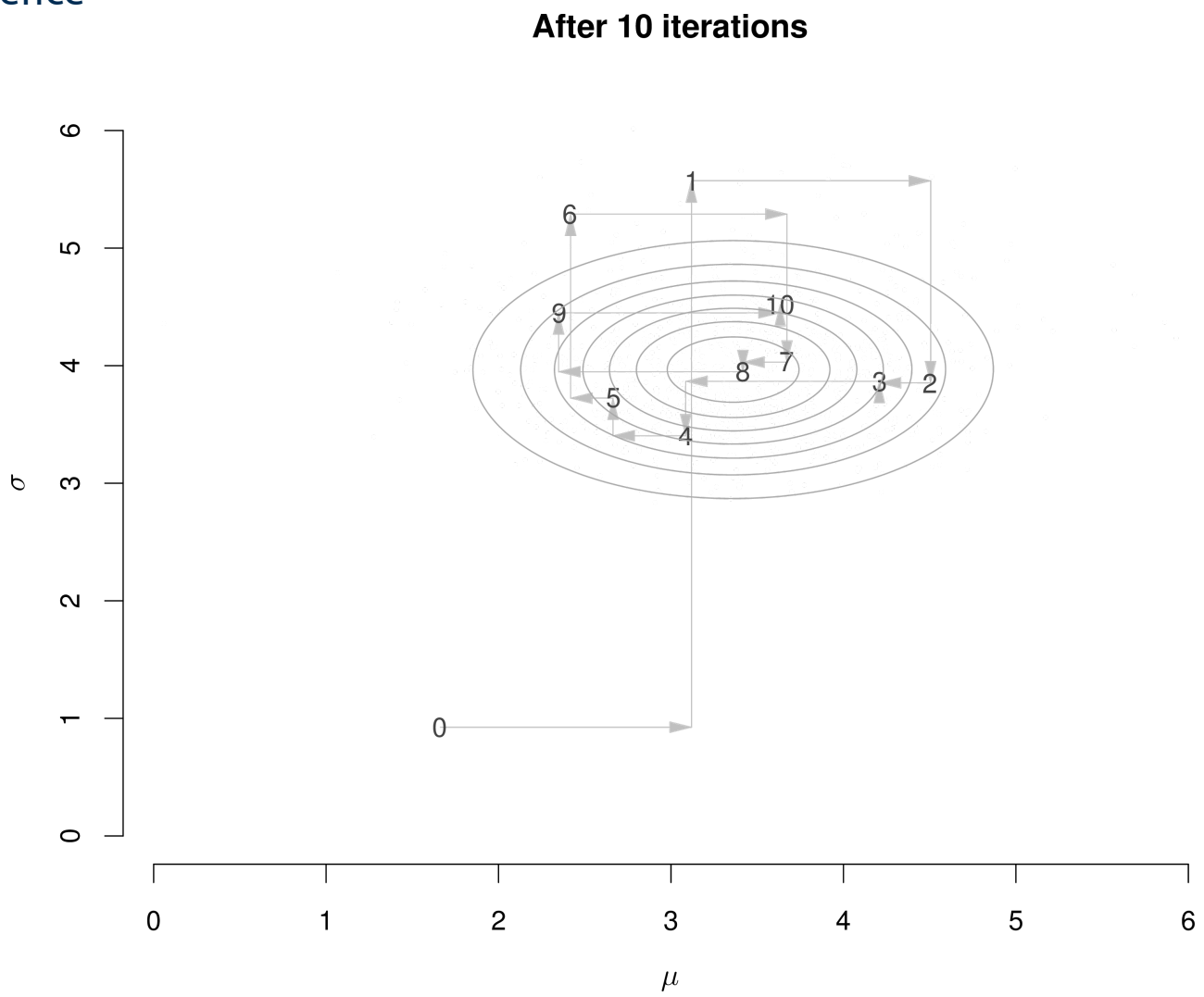
## Gibbs sampling – convergence



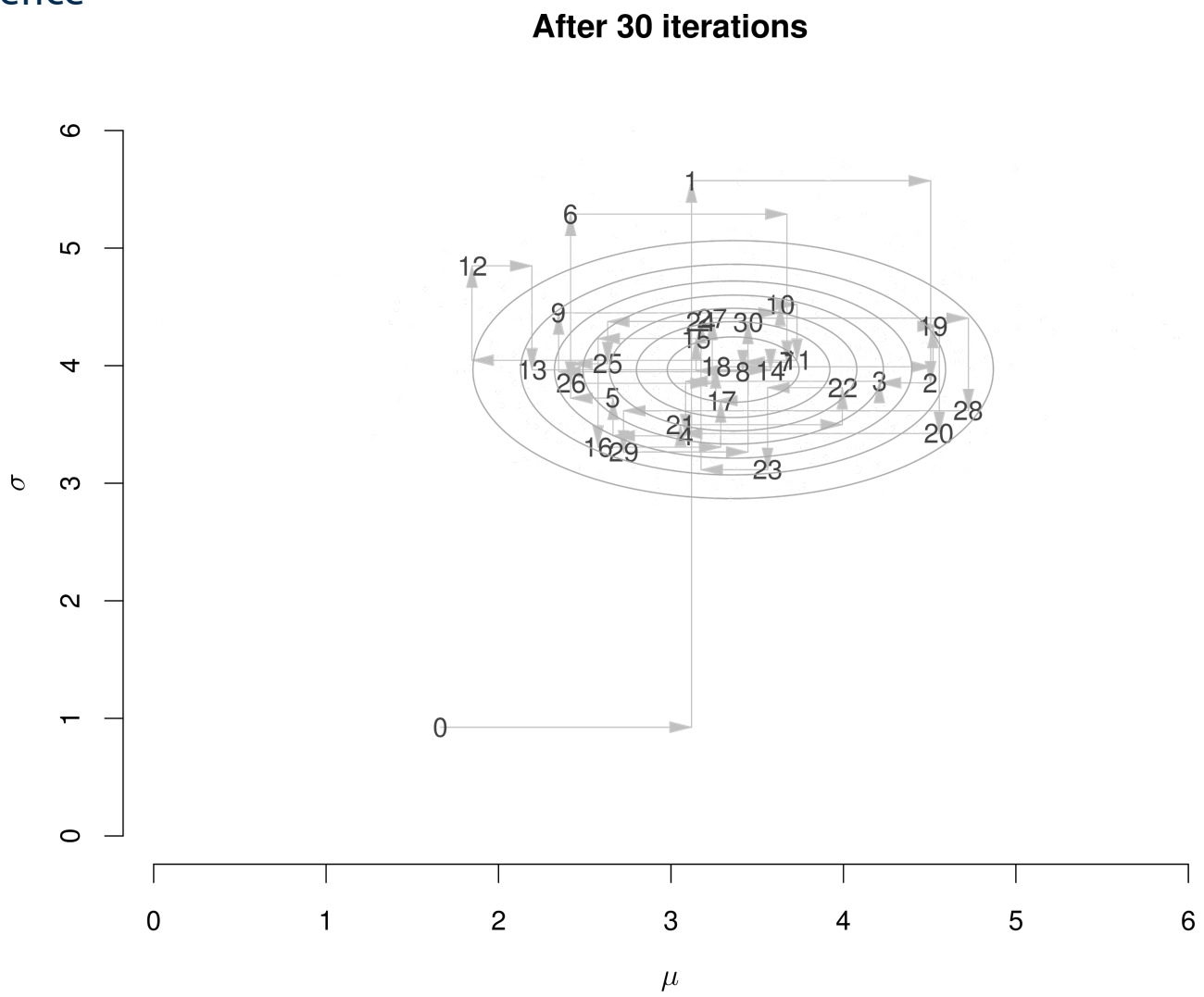
## Gibbs sampling – convergence



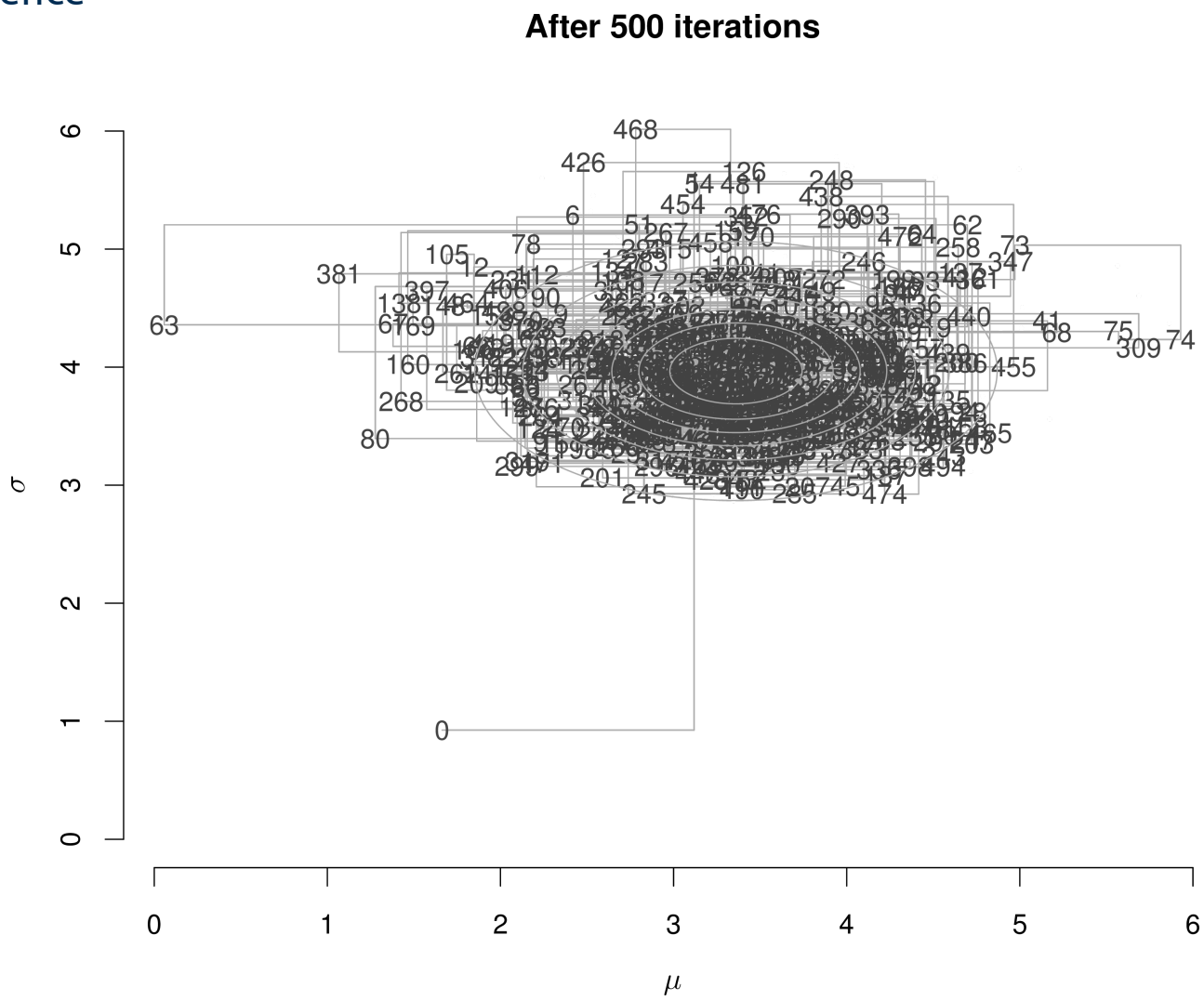
## Gibbs sampling – convergence



## Gibbs sampling – convergence

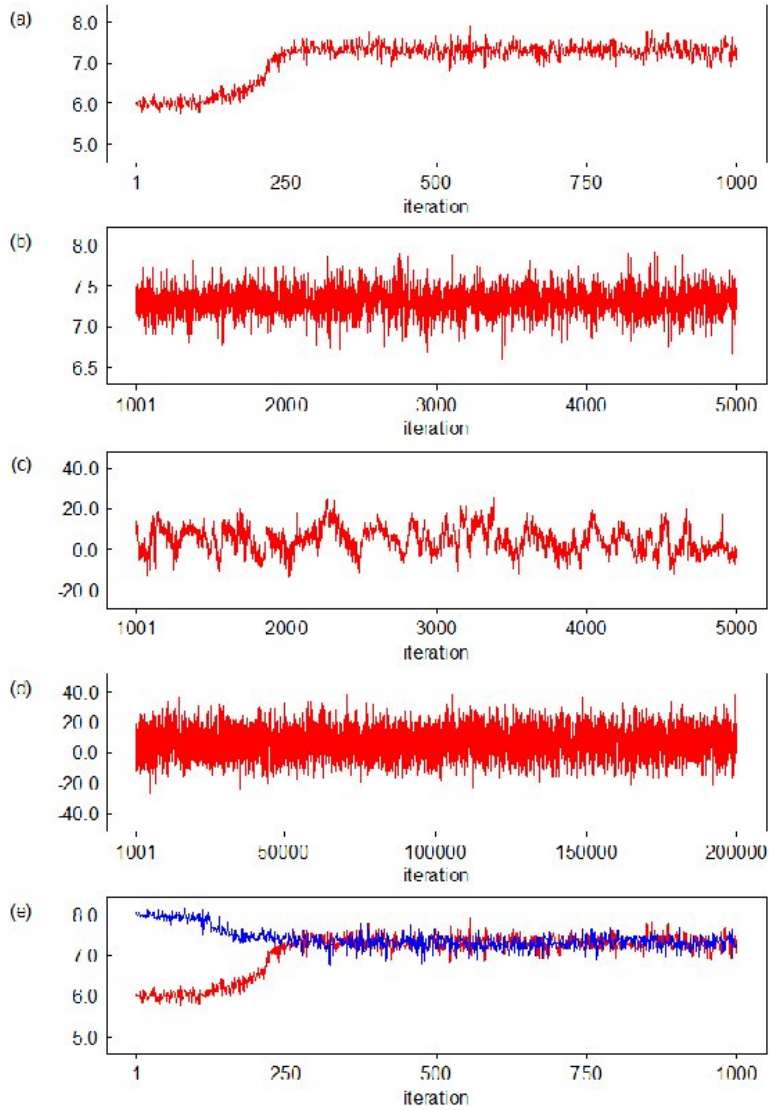


## Gibbs sampling – convergence



## Gibbs sampling – convergence

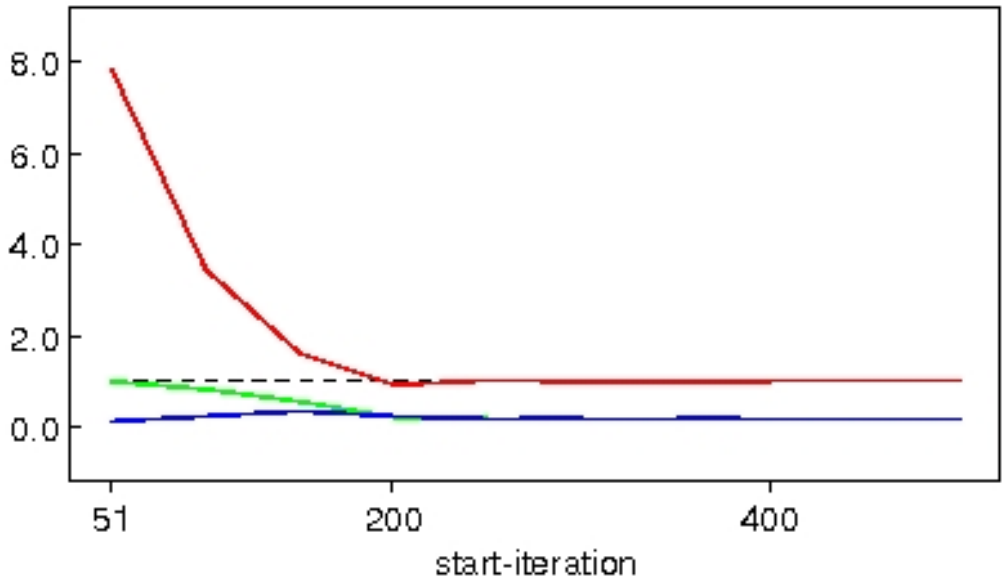
# Checking convergence: monitoring samples




- Convergence (or lack of) can be apparent from one chain
  - Want **fat hairy caterpillars** (b) – not twisting snakes
- Chain may have converged but be slow to **mix** (c)
  - Run chain for longer (d) to get more precise estimates
- One chain may get "stuck" in some area due to extreme initial value (a)
  - Run multiple chains from different initial values (e): check all end up in same place
- Parameterisation may make a big difference (see manual)

# Formal convergence tests

Formal diagnostics exist to check if multiple chains end up in essentially same place, eg Brooks-Gelman-Rubin (often referred to as Potential Scale Reduction, PSR) statistic




- Based on ratio of between to within variances of multiple chains: (ANOVA)
- OpenBUGS produces plots of
  - Average 80% interval within-chains (blue) and pooled 80% interval between-chains (green)
  - Ratio green/blue should converge to 1 (red) as iterations increase

- coda and R2OpenBUGS packages for R contain many other diagnostics
- **NB** This is only a heuristic measure – more recent work suggests alternative ways
  - https://avehtari.github.io/rhat_ess/rhat_ess.html
  - http://cknudson.com/Presentations/BayesComp2020.pdf

# How many iterations after convergence?

- How many significant figures do you need in your estimates: **your decision**
- Easiest strategy: run chains until the posterior summaries of interest don't change
- Monte Carlo Standard Error (MCSE) says how accurate the posterior mean is
- **Autocorrelated** samples need to be longer to get the same accuracy, compared to independent samples
  - Some theory (Raftery & Lewis, see BUGS Book for further details) suggests that to get 95% posterior quantiles with true 94.5-95.5% coverage need MCSE/posterior SD < 0.01, or **effective sample size** > 4000

# Supplying data to BUGS

1. Rectangular format – traditional "spreadsheet"-shaped data

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

2. R / S-Plus style "lists"

```
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))
```

List format more flexible, can specify constants alongside data – useful for complex multilevel models with variables of different lengths

(**NB** If using R interfaces to BUGS, just supply data in R list object – see later...)

## Drugs example

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

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

Alternatively, put all data and constants into model description:

```
model{
  theta ~ dbeta(9.2, 13.8)       # prior distribution
  y ~ dbin(theta, 20)            # sampling distribution
  y.pred ~ dbin(theta, 40)       # predictive distribution
  P.crit <- step(y.pred - 24.5) # =1 if y.pred >= ncrit,
                                 # =0 otherwise
  y <- 15                        # observed successes
}
```

# Initial values

- BUGS simulates from posterior – combination of prior and evidence from data – by MCMC
- Posterior **unknown** to start with – need to **initialize** simulation
- BUGS can automatically generate initial values for the simulation using `gen.inits` – simulates from the prior
  – We have seen this in the practical for forward sampling
- Fine if have informative prior information
- If have fairly "vague" priors, better to provide reasonable values in an initial-values list

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

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

# OpenBUGS output and exact answers

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

Exact answers from conjugate analysis:

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

MCMC results are within Monte Carlo error of the true values

Next lecture