What are Bayesian Methods

How did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

P R O B L E M.

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

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

Basic idea: Direct expression of uncertainty about unknown parameters

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

% absolute increase in major bleeds
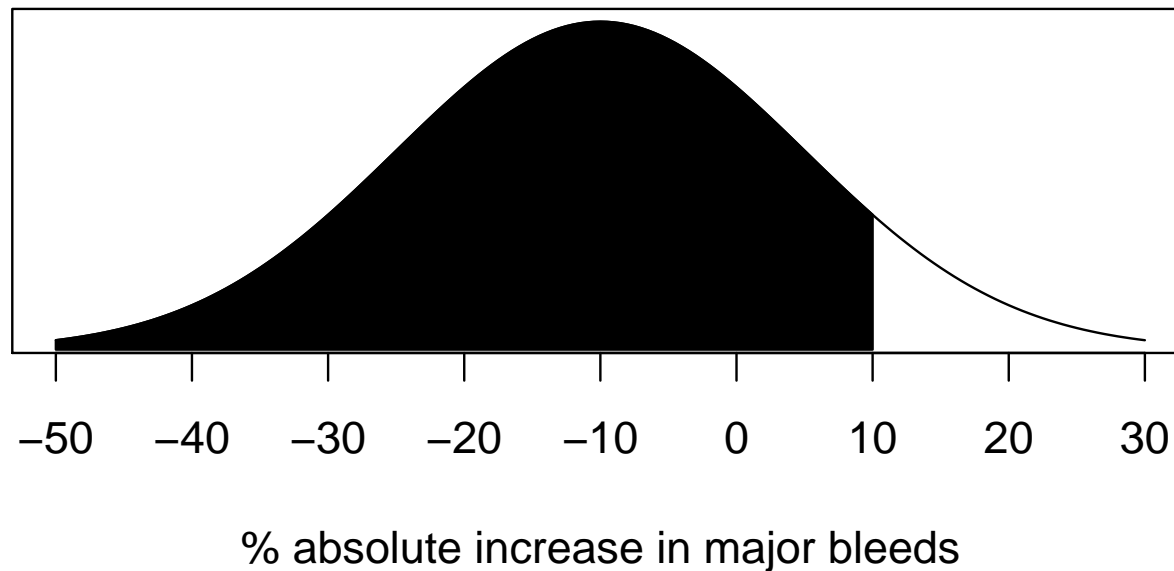
Why a direct probability distribution?

1. Tells us what we want: what are plausible values for the parameter of interest?

2. No *P-values*: just calculate relevant tail areas

3. No (difficult to interpret) *confidence intervals*: just report, say, central area that contains 95% of distribution

4. Fits naturally into decision analysis / cost-effectiveness analysis / project prioritisation

5. There is a procedure for adapting the distribution in the light of additional evidence: i.e. *Bayes theorem* allows us to learn from experience

What are Bayesian methods?

- Bayesian methods have been widely applied in many areas:

  - medicine / epidemiology
  - genetics
  - ecology
  - environmental sciences
  - social and political sciences
  - finance
  - archaeology
  - .....

- Motivations for adopting Bayesian approach vary:

  - natural and coherent way of thinking about science and learning
  - pragmatic choice that is suitable for the problem in hand

Spiegelhalter et al (2004) define a Bayesian approach as

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

They argue that a Bayesian approach is:

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

than traditional methods

Example

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

Conventional analysis

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

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

Bayesian analysis

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

The Bayesian analyst needs to explicitly state

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

and to combine these two sources to produce

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

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

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

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

- point and interval estimates of treatment effect

- point and interval estimates of any function of the parameters

- probability that treatment effect exceeds a clinically relevant value

- prediction of treatment effect in a new patient

- prior information for future trials

- inputs for decision making

- ....

Bayes theorem and its link with Bayesian inference

Axioms of probability

Let A, B be events, and $\{A_i, i = 1, 2, 3, ...\}$ be a set of events. The *probability* of A, $p(A)$, is a number which satisfies:

Axiom 1: $0 \leq p(A) \leq 1$ and $p(A) = 1$ if A is certain.

Axiom 2: If the events A_i are mutually exclusive, $p(\bigcup_i A_i) = \sum_i p(A_i)$

Axiom 3: $p(A \bigcap B) = p(B|A)p(A)$

Bayes' theorem

(provable from axioms)

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

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

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

Example: use of Bayes theorem in diagnostic testing

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

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

Let B be the event that they test positive.

We want $p(A|B)$.

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

"98% specificity" means that $p(B|\overline{A}) = .02$.

Now Bayes theorem says

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

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

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

- Our intuition is poor when processing probabilistic evidence

- The vital issue is *how should this test result change our belief that patient is HIV positive?*

- The disease prevalence can be thought of as a *'prior'* probability ($p = 0.001$)

- Observing a positive result causes us to modify this probability to $p = 0.045$. This is our *'posterior'* probability that patient is HIV positive.

- Bayes theorem applied to *observables* (as in diagnostic testing) is uncontroversial and established

- More controversial is the use of Bayes theorem in general statistical analyses, where *parameters* are the unknown quantities, and their prior distribution needs to be specified — this is **Bayesian inference**

Bayesian inference

Makes fundamental distinction between

- Observable quantities x, i.e. the data

- Unknown quantities θ

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

  $\rightarrow$ parameters are treated as random variables

  $\rightarrow$ in the Bayesian framework, we make probability statements about model parameters

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

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

$$p(x \mid \theta)$$

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

From a Bayesian point of view

- θ is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data
  $\rightarrow$ need to specify a **prior distribution** $p(\theta)$

- x is known so we should condition on it
  $\rightarrow$ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

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

  This is the **posterior distribution**

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

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

Bayesian inference with binary data

Example: Inference on proportions using discrete prior

Assume treatment may have response rate θ 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, $p(x \mid \theta) = \theta^x (1-\theta)^{(1-x)}$

θ	Prior $p(\theta)$	Likelihood $p(x=1 \mid \theta) = \theta$	Likelihood $\times$ prior $p(x=1|\theta)p(\theta)$	Posterior $p(\theta|x=1) = \frac{p(x=1|\theta)p(\theta)}{\sum_j p(x=1|\theta_j)p(\theta_j)}$
.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%.

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 θ, leads to a binomial likelihood

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

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

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

Posterior is then

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

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

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

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

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

Combining this with the binomial likelihood gives a posterior distribution

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

Comments

- When the prior and posterior come from the same family of distributions the prior is said to be **conjugate** to the likelihood

  - Occurs when prior and likelihood have the same 'kernel'

- A Beta(a, b) distribution has

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

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

- a and b are equivalent to observing a priori $a-1$ successes in $a+b-2$ trials $\rightarrow$ can be elicited

- With fixed a and b, as r and n increase, $E(\theta|r, n) \rightarrow r/n$ (the MLE), and the variance tends to zero

  - This is a general phenomenon: as n increases, posterior distribution gets more concentrated and the likelihood dominates the prior

- A Beta(1, 1) is equivalent to Uniform(0, 1)

Example: Drug

- We consider early investigation of a new drug

- Experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible

- We interpreted this as a distribution with mean $= 0.4$, standard deviation 0.1 and showed that a Beta(9.2,13.8) distribution has these properties

- Suppose we now treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses
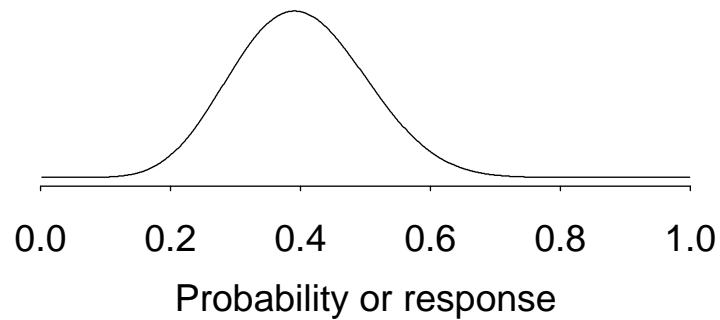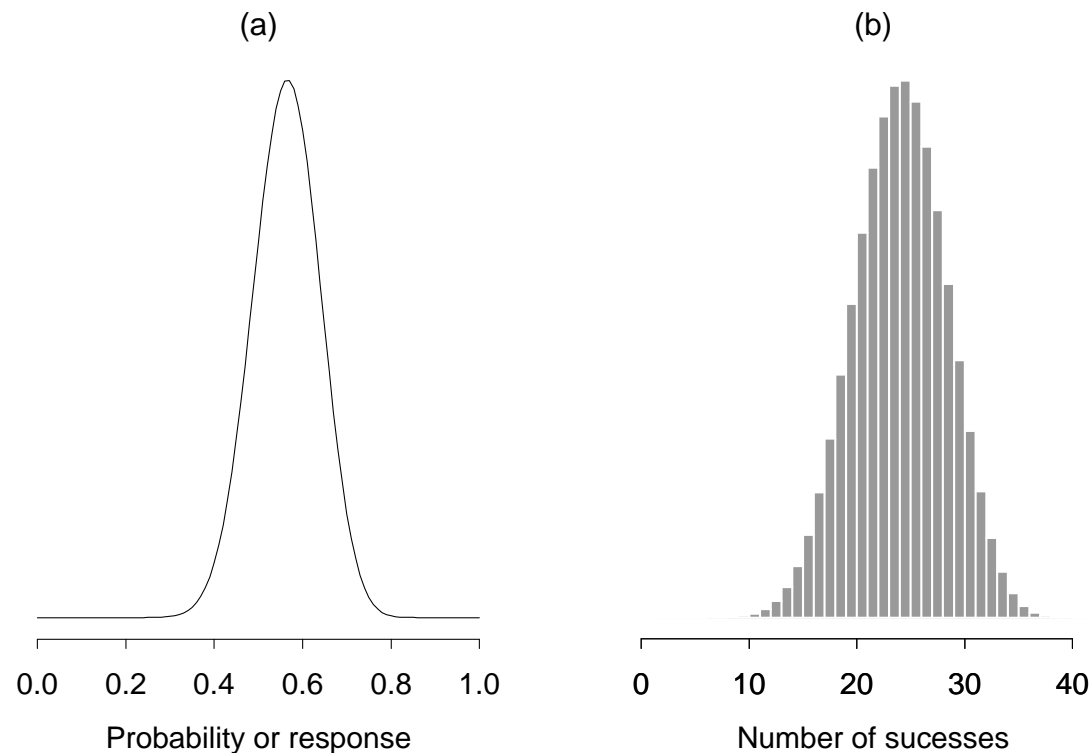
Beta(9.2, 13.8) prior distribution supporting response rates between 0.2 and 0.6
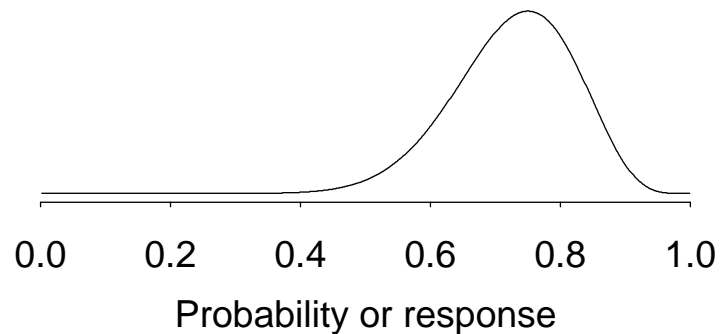
Likelihood arising from a Binomial observation of 15 successes out of 20 cases
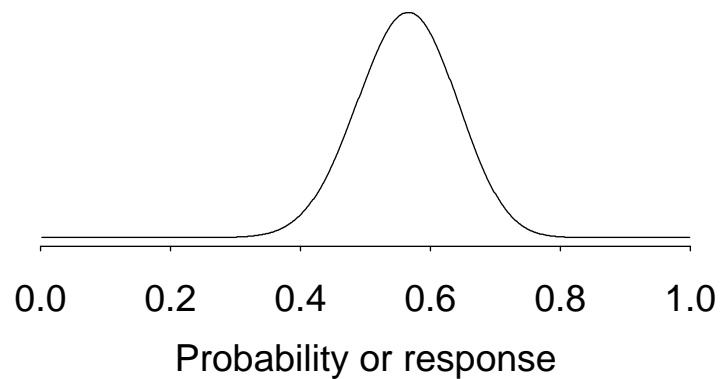
Parameters of the Beta distribution are updated to (a+15, b+20-15) = (24.2, 18.8): mean 24.2/(24.2+18.8) = 0.56

(a)

(b)

Probability or response

Number of sucesses

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

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

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

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

Drug (continued): direct use of data when using Monte-Carlo (MC) methods

Previous slides showed closed-form solution to drug problem

Can also use simulation (Monte Carlo) methods for inference

When using conjugate analysis, we can sample directly from closed-form posterior

As before, assume we have observed 15 successes out of 20 subjects, and wish to predict whether we will get more than 25 successes in next 40 patients.

The model can be written

$$\theta \quad \sim \text{Beta}[a, b] \qquad\qquad \text{prior distribution}$$

$$y \quad \sim \text{Binomial}[\theta, m] \qquad\qquad \text{sampling distribution (likelihood)}$$

$$\theta | y \quad \sim \text{Beta}[a + y, b + m - y] \quad \text{posterior distribution}$$

$$y_{\text{pred}} \sim \text{Binomial}[\theta, n] \qquad\qquad \text{predictive distribution}$$

$$P_{\text{crit}} = P(y_{\text{pred}} \geq n_{\text{crit}}) \qquad\qquad \text{Probability of exceeding critical threshold}$$

Drug (continued): indirect learning about parameters from data using Markov chain Monte-Carlo (MCMC) methods

More generally, we can specify the prior and likelihood separately.

The model can be written

$$\theta \sim \text{Beta}[a, b] \qquad \text{prior distribution}$$

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

$$y_{\text{pred}} \sim \text{Binomial}[\theta, n] \quad \text{predictive distribution}$$

$$P_{\text{crit}} = P(y_{\text{pred}} \geq n_{\text{crit}}) \text{ Probability of exceeding critical threshold}$$

```
# In BUGS syntax:

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

The BUGS program

Bayesian inference Using Gibbs Sampling

- Language for specifying complex Bayesian models
- Constructs object-oriented internal representation of the model
- Simulation from full conditionals using Gibbs sampling
- Current version (WinBUGS 1.4) runs in Windows
- Just Another Gibbs Sampler (JAGS) - C++ Winbugs "clone".

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

Graphical representation of models
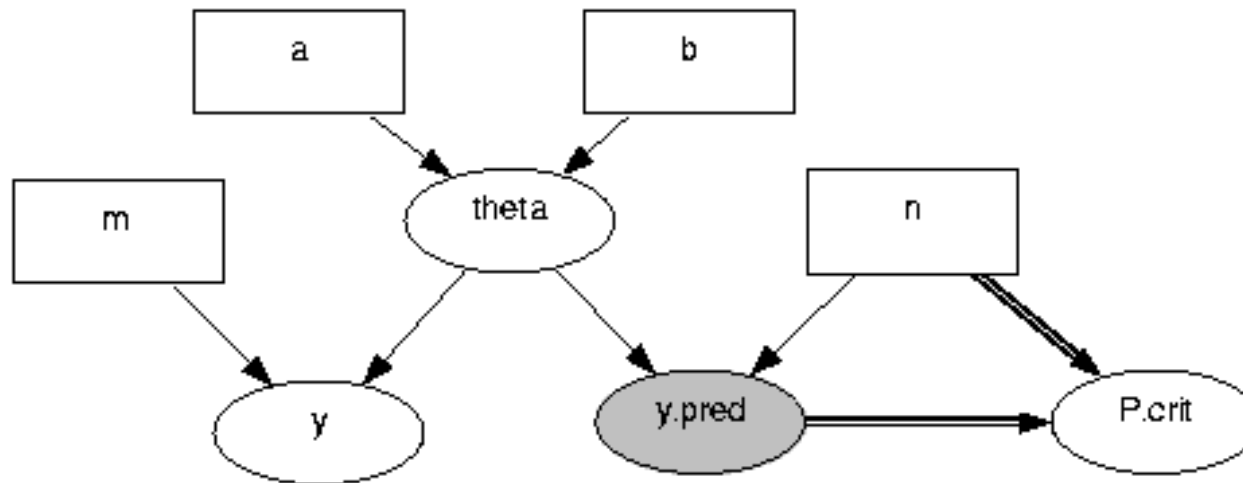
Note that adding data to a model is simply extending the graph.

Data files

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

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

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

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

  y          <- 15
}
```

The `WinBUGS` data formats

`WinBUGS` accepts data files in:

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

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

2. S-Plus format:

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

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

Initial values

- WinBUGS can automatically generate initial values for the MCMC analysis using *gen inits*

- Fine if have informative prior information

- If have fairly 'vague' priors, better to provide reasonable values in an initial-values list

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

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

Running WinBUGS for MCMC analysis (single chain)

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

WinBUGS output
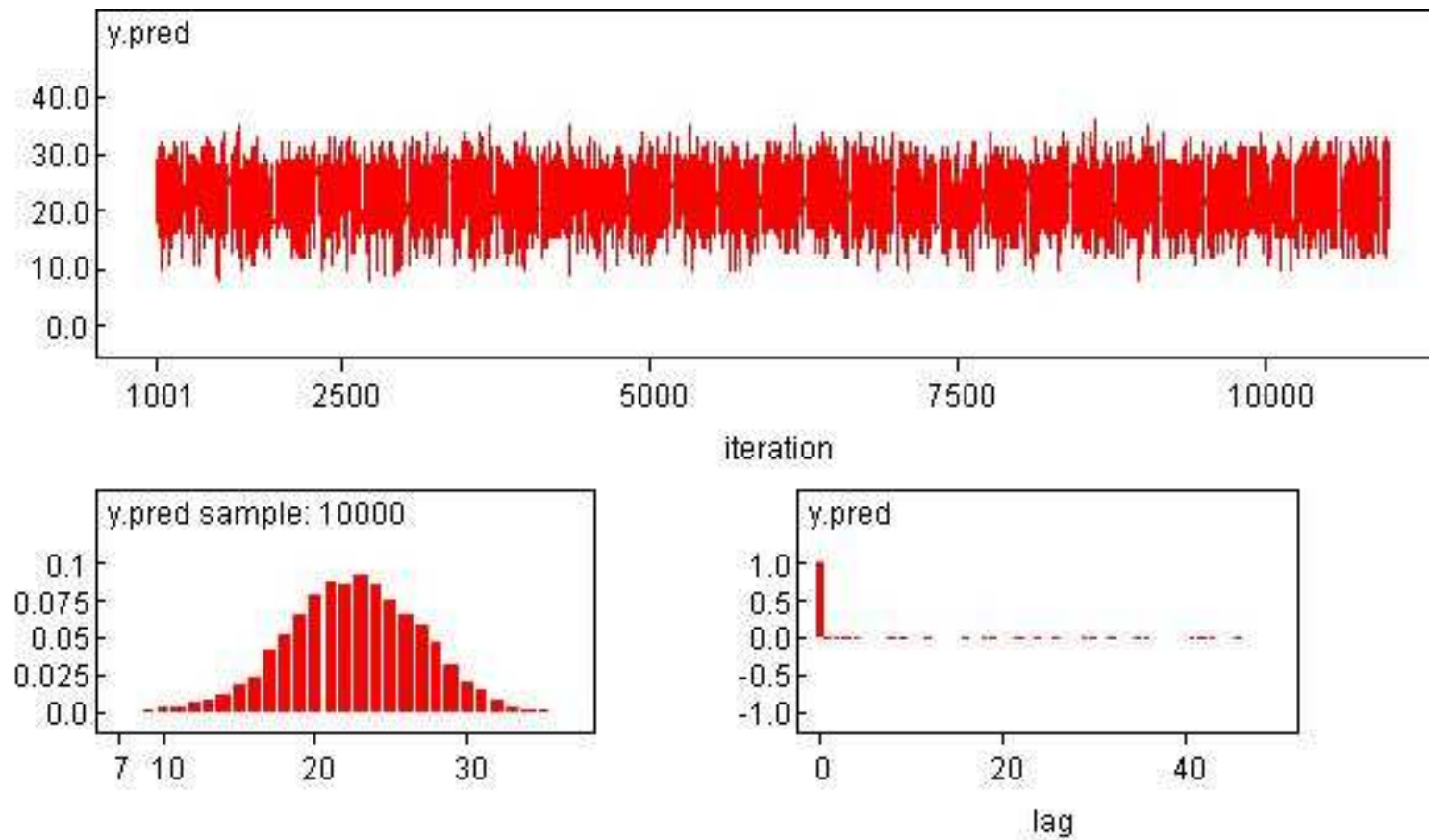
WinBUGS output and exact answers

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

Exact answers from conjugate analysis

- θ: mean 0.563 and standard deviation 0.075

- Y^{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

Summary

For all these examples, we see that

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

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

As $n \to \infty$,

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

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

Choosing prior distributions

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

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

- Conjugate prior distributions are mathematically convenient, but do not exit for all likelihoods, and can be restrictive

- Computations for non-conjugate priors are harder, but possible using MCMC.

Further reading

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

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

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