

BESTA - Center for BioEngineering Statistics
School of BME GaTech



Thursday, April 19, 2007

This handout is adaptation of two chapters from the book (available in July 2007):

Kvam P. and Vidakovic, B. (2007). *Nonparametric Statistics For Engineers and Scientists*, Wiley, Hoboken, NJ, 450pp.

1

Bayesian Statistics

To anyone sympathetic with the current neo-Bernoullian neo-Bayesian Ramseyesque Finettist Savageous movement in statistics, the subject of testing goodness of fit is something of an embarrassment.

– F. J. Anscombe (1962)

1.1 THE BAYESIAN PARADIGM

There are several paradigms for approaching statistical inference, but the two dominant ones are *frequentist* (sometimes called classical or traditional) and *Bayesian*. The overview in the previous chapter covered mainly classical approaches. According to the Bayesian paradigm, the unobservable parameters in a statistical model are treated as random. When no data are available, a *prior distribution* is used to quantify our knowledge about the parameter. When data are available, we can update our prior knowledge using the conditional distribution of parameters, given the data. The transition from the prior to the posterior is possible via the Bayes theorem. Figure 1.1(a) shows a portrait of the Reverend Thomas Bayes whose posthumously published essay gave impetus to alternative statistical approaches (Bayes, 1763). His signature is shown in Figure 1.1(b).

Suppose that before the experiment our prior distribution describing θ is $\pi(\theta)$. The data are coming from the assumed model (likelihood) which depends on the parameter and is denoted by $f(x|\theta)$. Bayes theorem updates the prior



(a)

I am
 My Lord
 Your Lordship's
 most obedient
 humble Servant
 T. Bayes.

(b)

Fig. 1.1 The Reverend Thomas Bayes (1702–1761); (b) Bayes' signature.

$\pi(\theta)$ to the posterior by accounting for the data x ,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}, \quad (1.1)$$

where $m(x)$ is a normalizing constant, $m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$.

Once the data x are available, θ is the only unknown quantity and the posterior distribution $\pi(\theta|x)$ completely describes the uncertainty. There are two key advantages of Bayesian paradigm: (i) once the uncertainty is expressed via the probability distribution and the statistical inference can be automated, it follows a conceptually simple recipe, and (ii) available prior information is coherently incorporated into the statistical model.

1.2 INGREDIENTS FOR BAYESIAN INFERENCE

The *model* for a typical observation X conditional on unknown parameter θ is the density function $f(x|\theta)$. As a function of θ , $f(x|\theta) = L(\theta)$ is called a *likelihood*. The functional form of f is fully specified up to a parameter θ . According to the *likelihood principle*, all experimental information about the data must be contained in this likelihood function.

The parameter θ , with values in the parameter space Θ , is considered a random variable. The random variable θ has a distribution $\pi(\theta)$ called the prior distribution. This prior describes uncertainty about the parameter before data are observed. If the prior for θ is specified up to a parameter τ , $\pi(\theta|\tau)$, τ is called a *hyperparameter*.

Our goal is to start with this prior information and update it using the data to make the best possible estimator of θ . We achieve this through the likelihood function to get $\pi(\theta|x)$, called the *posterior* distribution for θ , given

$X = x$. Accompanying its role as the basis to Bayesian inference, the posterior distribution has been a source for an innumerable accumulation of tacky “butt” jokes by unimaginative statisticians with low-brow sense of humor.

To find $\pi(\theta|x)$, we use Bayes rule to divide *joint* distribution for X and θ ($h(x, \theta) = f(x|\theta)\pi(\theta)$) by the *marginal* distribution $m(x)$, which can be obtained by integrating out parameter θ from the joint distribution $h(x, \theta)$,

$$m(x) = \int_{\Theta} h(x, \theta)d\theta = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta.$$

The marginal distribution is also called the *prior predictive* distribution. Finally we arrive at an expression for the posterior distribution $\pi(\theta|x)$:

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}.$$

The following table summarizes the notation:

Likelihood	$f(x \theta)$
Prior Distribution	$\pi(\theta)$
Joint Distribution	$h(x, \theta) = f(x \theta)\pi(\theta)$
Marginal Distribution	$m(x) = \int_{\Theta} f(x \theta)\pi(\theta)d\theta$
Posterior Distribution	$\pi(\theta x) = f(x \theta)\pi(\theta)/m(x)$

Example 1.1 Normal Likelihood with Normal Prior. The normal likelihood and normal prior combination is important as it is often used in practice. Assume that an observation X is normally distributed with mean θ and known variance σ^2 . The parameter of interest, θ , has normal distribution as well with hyperparameters μ and τ^2 . Starting with our Bayesian model of $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu, \tau^2)$, we will find the marginal and posterior distributions.

The exponent ζ in the joint distribution $h(x, \theta)$ is

$$\zeta = -\frac{1}{2\sigma^2}(x - \theta)^2 - \frac{1}{2\tau^2}(\theta - \mu)^2.$$

After straightforward but somewhat tedious algebra, ζ can be expressed as

$$\zeta = -\frac{1}{2\rho} \left(\theta - \rho \left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2} \right) \right)^2 - \frac{1}{2(\sigma^2 + \tau^2)}(x - \mu)^2,$$

where

$$\rho = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

Recall that $h(x, \theta) = f(x|\theta)\pi(\theta) = \pi(\theta|x)m(x)$, so the marginal distribution

simply resolves to $X \sim \mathcal{N}(\mu, \sigma^2 + \tau^2)$ and the posterior distribution comes out to be

$$\theta|X \sim \mathcal{N}\left(\frac{\tau^2}{\sigma^2 + \tau^2}X + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right).$$

If X_1, X_2, \dots, X_n are observed instead of a single observation X , then the sufficiency of \bar{X} implies that the Bayesian model for θ is the same as for X with σ^2/n in place of σ^2 . In other words, the Bayesian model is

$$\bar{X}|\theta \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right) \text{ and } \theta \sim \mathcal{N}(\mu, \tau^2),$$

producing

$$\theta|\bar{X} \sim \mathcal{N}\left(\frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\bar{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2}\mu, \frac{\frac{\sigma^2}{n}\tau^2}{\frac{\sigma^2}{n} + \tau^2}\right).$$

Notice that the posterior mean

$$\frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\bar{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2}\mu \tag{1.2}$$

is a weighted linear combination of the MLE \bar{X} and the prior mean μ with weights $\lambda = n\tau^2/(\sigma^2 + n\tau^2)$ and $1 - \lambda = \sigma^2/(\sigma^2 + n\tau^2)$. When the sample size n increases, $\lambda \rightarrow 1$, and the influence of the prior mean diminishes. On the other hand when n is small and our prior opinion about μ is strong (i.e., τ^2 is small) the posterior mean is close to the prior mean μ . We will see later several more cases in which the posterior mean is a linear combination of a frequentist estimate and the prior mean.

For instance, suppose 10 observations are coming from $\mathcal{N}(\theta, 100)$. Assume that the prior on θ is $\mathcal{N}(20, 20)$. For the observations $\{2.944, -13.361, 7.143, 16.235, -6.917, 8.580, 12.540, -15.937, -14.409, 5.711\}$ the posterior is $\mathcal{N}(6.835, 6.667)$. The three densities are shown in Figure 1.2.

1.2.1 Quantifying Expert Opinion

Bayesian statistics has become increasingly popular in engineering, and one reason for its increased application is that it allows researchers to input expert opinion as a catalyst in the analysis (through the prior distribution). Expert opinion might consist of subjective inputs from experienced engineers, or perhaps a summary judgment of past research that yielded similar results.

Example 1.2 Prior Elicitation for Reliability Tests. Suppose each of n independent reliability tests a machine reveals either a successful or unsuccessful outcome. If θ represents the reliability of the machine, let X be the number of successful missions the machine experienced in n independent tri-

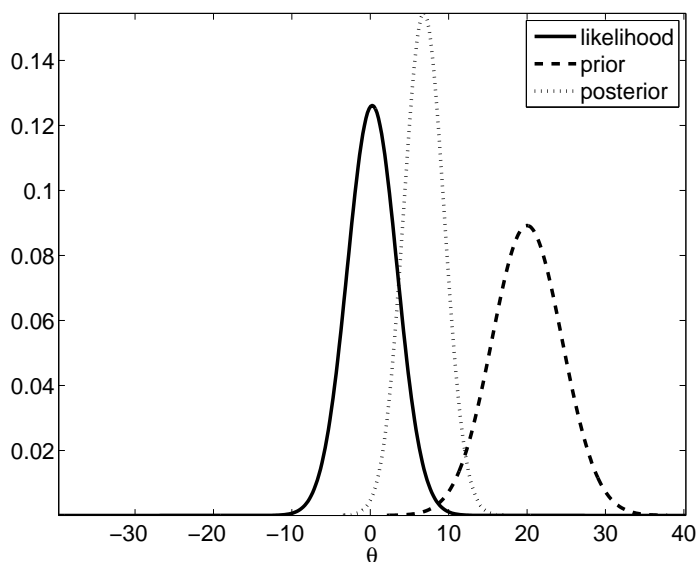


Fig. 1.2 The normal $\mathcal{N}(\theta, 100)$ likelihood, $\mathcal{N}(20, 20)$ prior, and posterior for data $\{2.9441, -13.3618, \dots, 5.7115\}$.

als. X is distributed Binomial with parameters n (known) and θ (unknown). We probably won't expect an expert to quantify their uncertainty about θ directly into a prior distribution $\pi(\theta)$. Perhaps the researcher can elicit information such as the expected value and standard deviation of θ . If we suppose the prior distribution for θ is $\mathcal{Be}(\alpha, \beta)$, where the hyper-parameters α and β are known, then

$$\pi(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \quad 0 \leq \theta \leq 1.$$

With $X|\theta \sim \mathcal{Bin}(n, \theta)$, the joint, marginal, and posterior distributions are

$$\begin{aligned} h(x, \theta) &= \frac{\binom{n}{x}}{B(\alpha, \beta)} \theta^{\alpha+x-1} (1-\theta)^{n-x+\beta-1}, \quad 0 \leq \theta \leq 1, x = 0, 1, \dots, n. \\ m(x) &= \frac{\binom{n}{x} B(x + \alpha, n - x + \beta)}{B(\alpha, \beta)}, \quad x = 0, 1, \dots, n. \\ \pi(\theta|x) &= \frac{1}{B(x + \alpha, n - x + \beta)} \theta^{\alpha+x-1} (1-\theta)^{n-x+\beta-1}, \quad 0 \leq \theta \leq 1. \end{aligned}$$

It is easy to see that the posterior distribution is $\mathcal{Be}(\alpha + x, n - x + \beta)$. Suppose the experts suggest that the previous version of this machine was “reliable 93%

of the time, plus or minus 2%". We might take $\mathbb{E}(\theta) = 0.93$ and insinuate that $\sigma_\theta = 0.04$ (or $\text{Var}(\theta) = 0.0016$), using two-sigma rule as an argument. From the beta distribution,

$$\mathbb{E}\theta = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}\theta = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

We can actually solve for α and β as a function of the expected value μ and variance σ^2 ,

$$\alpha = \mu(\mu - \mu^2 - \sigma^2)/\sigma^2, \quad \text{and} \quad \beta = (1 - \mu)(\mu - \mu^2 - \sigma^2)/\sigma^2.$$

In this example, $(\mu, \sigma^2) = (0.93, 0.0016)$ leads to $\alpha = 36.91$ and $\beta = 2.78$. To update the data X , we will use a $\mathcal{Be}(36.91, 2.78)$ distribution for a prior on θ . Consider the weight given to the expert in this example. If we observe one test only and the machine happened to fail, our posterior distribution is then $\mathcal{Be}(36.91, 3.78)$, which has a mean equal to 0.9071. The MLE for the average reliability is obviously zero, with such precise information elicited from the expert, the posterior is close to the prior. In some cases when you do not trust your expert, this might be unsettling and less informative priors may be a better choice.

1.2.2 Point Estimation

The posterior is the ultimate experimental summary for a Bayesian. The location measures (especially the mean) of the posterior are of great importance. The posterior mean represents the most frequently used Bayes estimator for the parameter. The posterior mode and median are less commonly used alternative Bayes estimators.

An objective way to choose an estimator from the posterior is through a penalty or loss function $L(\hat{\theta}, \theta)$ that describes how we penalize the discrepancy of the estimator $\hat{\theta}$ from the parameter θ . Because the parameter is viewed as a random variable, we seek to minimize *expected* loss, or *posterior risk*:

$$R(\hat{\theta}, x) = \int L(\hat{\theta}, \theta)\pi(\theta|x)d\theta.$$

For example, the estimator based on the common squared-error loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ minimizes $\mathbb{E}((\hat{\theta} - \theta)^2)$, where expectation is taken over the posterior distribution $\pi(\theta|X)$. It's easy to show that the estimator turns out to be the posterior expectation. Similar to squared-error loss, if we use absolute-error loss $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, the Bayes estimator is the posterior median.

The posterior mode maximizes the posterior density the same way MLE is maximizing the likelihood. The *generalized MLE* maximizes $\pi(\theta|X)$. Bayesians prefer the name MAP (maximum aposteriori) estimator or simply posterior mode. The MAP estimator is popular in Bayesian analysis in part because it

is often computationally less demanding than the posterior mean or median. The reason is simple; to find the maximum, the posterior need not to be fully specified because $\operatorname{argmax}_{\theta}\pi(\theta|x) = \operatorname{argmax}_{\theta}f(x|\theta)\pi(\theta)$, that is, one simply maximizes the product of likelihood and the prior.

In general, the posterior mean will fall between the MLE and the the prior mean. This was demonstrated in Example 1.1. As another example, suppose we flipped a coin four times and tails showed up on all 4 occasions. We are interested in estimating probability of heads, θ , in a Bayesian fashion. If the prior is $\mathcal{U}(0, 1)$, the posterior is proportional to $\theta^0(1 - \theta)^4$ which is beta $\mathcal{Be}(1, 5)$. The posterior mean *shrinks* the MLE toward the expected value of the prior ($1/2$) to get $\hat{\theta}_B = 1/(1 + 5) = 1/6$, which is a more reasonable estimator of θ than the MLE.

Example 1.3 Binomial-Beta Conjugate Pair. Suppose $X|\theta \sim \mathcal{Bin}(n, \theta)$. If the prior distribution for θ is $\mathcal{Be}(\alpha, \beta)$, the posterior distribution is $\mathcal{Be}(\alpha + x, n - x + \beta)$. Under squared error loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, the Bayes estimator of θ is the expected value of the posterior

$$\hat{\theta}_B = \frac{\alpha + x}{(\alpha + x)(\beta + n - x)} = \frac{\alpha + x}{\alpha + \beta + n}.$$

This is actually a weighted average of MLE, X/n , and the prior mean $\alpha/(\alpha + \beta)$. Notice that, as n becomes large, the posterior mean is getting close to MLE, because the weight $n/(n + \alpha + \beta)$ tends to 1. On the other hand, when α is large, the posterior mean is close to the prior mean. Large α indicates small prior variance (for fixed β , the variance of $\mathcal{Be}(\alpha, \beta)$ behaves as $O(1/\alpha^2)$) and the prior is concentrated about its mean. Recall the Example 1.2; after one machine trial failure the posterior distribution mean changed from 0.93 (the prior mean) to 0.9071, shrinking only slightly toward the MLE (which is zero).

Example 1.4 Jeremy's IQ. Jeremy, an enthusiastic Georgia Tech student, spoke in class and posed a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean θ and the variance of 80. Prior (and expert) opinion is that the IQ of Georgia Tech students, θ , is a normal random variable, with mean 110 and the variance 120. Jeremy took the test and scored 98. The traditional estimator of θ would be $\hat{\theta} = X = 98$. The posterior is $\mathcal{N}(102.8, 48)$, so the Bayes estimator of Jeremy's IQ score is $\hat{\theta}_B = 102.8$.

Example 1.5 Poisson-Gamma Conjugate Pair. Let X_1, \dots, X_n , given θ are Poisson $\mathcal{P}(\theta)$ with probability mass function

$$f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!} e^{-\theta},$$

and $\theta \sim \mathcal{G}(\alpha, \beta)$ is given by $\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$. Then,

$$\pi(\theta|X_1, \dots, X_n) = \pi(\theta|\sum X_i) \propto \theta^{\sum X_i + \alpha - 1} e^{-(n+\beta)\theta},$$

which is $\mathcal{G}(\sum_i X_i + \alpha, n + \beta)$. The mean is $\mathbb{E}(\theta|X) = (\sum X_i + \alpha)/(n + \beta)$, and it can be represented as a weighted average of the MLE and the prior mean:

$$\mathbb{E}\theta|X = \frac{n}{n + \beta} \frac{\sum X_i}{n} + \frac{\beta}{n + \beta} \frac{\alpha}{\beta}.$$

1.2.3 Conjugate Priors

We have seen two convenient examples for which the posterior distribution remained in the same family as the prior distribution. In such a case, the effect of likelihood is only to “update” the prior parameters and not to change prior’s functional form. We say that such priors are *conjugate* with the likelihood. Conjugacy is popular because of its mathematical convenience; once the conjugate pair likelihood/prior is found, the posterior is calculated with relative ease. In the years BC¹ and pre-MCMC era, conjugate priors have been extensively used (and overused and misused) precisely because of this computational convenience. Nowadays, the general agreement is that simple conjugate analysis is of limited practical value since, given the likelihood, the conjugate prior has limited modeling capability.

There are many univariate and multivariate instances of conjugacy. The following table provides several cases. For practice you may want to workout the posteriors in the table.

1.2.4 Interval Estimation: Credible Sets

Bayesians call interval estimators of model parameters *credible sets*. Naturally, the measure used to assess the credibility of an interval estimator is the posterior distribution. Students learning concepts of classical confidence intervals (CIs) often err by stating that “the probability that the CI interval $[L, U]$ contains parameter θ is $1 - \alpha$ ”. The correct statement seems more convoluted; one needs to generate data from the underlying model many times and for each generated data set to calculate the CI. The proportion of CIs covering the unknown parameter “tends to” $1 - \alpha$. The Bayesian interpretation of a credible set C is arguably more natural: The probability of a parameter belonging to the set C is $1 - \alpha$. A formal definition follows.

Assume the set C is a subset of Θ . Then, C is *credible set* with credibility

¹For some, the BC era signifies *Before Christ*, rather than *Before Computers*.

Table 1.1 Some conjugate pairs. Here \mathbf{X} stands for a sample of size n , X_1, \dots, X_n .

Likelihood	Prior	Posterior
$X \theta \sim \mathcal{N}(\theta, \sigma^2)$	$\theta \sim \mathcal{N}(\mu, \tau^2)$	$\theta X \sim \mathcal{N}\left(\frac{\tau^2}{\sigma^2 + \tau^2}X + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$
$X \theta \sim \mathcal{B}(n, \theta)$	$\theta \sim \mathcal{B}e(\alpha, \beta)$	$\theta X \sim \mathcal{B}e(\alpha + x, n - x + \beta)$
$\mathbf{X} \theta \sim \mathcal{P}(\theta)$	$\theta \sim \mathcal{G}amma(\alpha, \beta)$	$\theta \mathbf{X} \sim \mathcal{G}amma(\sum_i X_i + \alpha, n + \beta)$
$\mathbf{X} \theta \sim \mathcal{NB}(m, \theta)$	$\theta \sim \mathcal{B}e(\alpha, \beta)$	$\theta \mathbf{X} \sim \mathcal{B}e(\alpha + mn, \beta + \sum_{i=1}^n x_i)$
$X \sim \mathcal{G}amma(n/2, 1/(2\theta))$	$\theta \sim \mathcal{IG}(\alpha, \beta)$	$\theta X \sim \mathcal{IG}(n/2 + \alpha, x/2 + \beta)$
$\mathbf{X} \theta \sim \mathcal{U}(0, \theta)$	$\theta \sim \mathcal{P}a(\theta_0, \alpha)$	$\theta \mathbf{X} \sim \mathcal{P}a(\max\{\theta_0, X_1, \dots, X_n\}, \alpha + n)$
$X \theta \sim \mathcal{N}(\mu, \theta)$	$\theta \sim \mathcal{IG}(\alpha, \beta)$	$\theta X \sim \mathcal{IG}(\alpha + 1/2, \beta + (\mu - X)^2/2)$
$X \theta \sim \mathcal{G}amma(\nu, \theta)$	$\theta \sim \mathcal{G}a(\alpha, \beta)$	$\theta X \sim \mathcal{G}amma(\alpha + \nu, \beta + x)$

$(1 - \alpha)100\%$ if

$$P(\theta \in C|X) = \mathbb{E}(I(\theta \in C)|X) = \int_C \pi(\theta|x)d\theta \geq 1 - \alpha.$$

If the posterior is discrete, then the integral is a sum (using the counting measure) and

$$P(\theta \in C|X) = \sum_{\theta_i \in C} \pi(\theta_i|x) \geq 1 - \alpha.$$

This is the definition of a $(1 - \alpha)100\%$ credible set, and for any given posterior distribution such a set is not unique.

For a given credibility level $(1 - \alpha)100\%$, the shortest credible set has obvious appeal. To minimize size, the sets should correspond to highest posterior probability density areas (HPDs).

Definition 1.1 *The $(1 - \alpha)100\%$ HPD credible set for parameter θ is a set C , subset of Θ of the form*

$$C = \{\theta \in \Theta | \pi(\theta|x) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant for which

$$P(\theta \in C|X) \geq 1 - \alpha.$$

Geometrically, if the posterior density is cut by a horizontal line at the height $k(\alpha)$, the set C is projection on the θ axis of the part of line that lies

below the density.

Example 1.6 Jeremy's IQ, Continued. Recall Jeremy, the enthusiastic Georgia Tech student from Example 1.4, who used Bayesian inference in modeling his IQ test scores. For a score $X|\theta$ he was using a $\mathcal{N}(\theta, 80)$ likelihood, while the prior on θ was $\mathcal{N}(110, 120)$. After the score of $X = 98$ was recorded, the resulting posterior was normal $\mathcal{N}(102.8, 48)$.

Here, the MLE is $\hat{\theta} = 98$, and a 95% confidence interval is $[98 - 1.96\sqrt{80}, 98 + 1.96\sqrt{80}] = [80.4692, 115.5308]$. The length of this interval is approximately 35. The Bayesian counterparts are $\hat{\theta} = 102.8$, and $[102.8 - 1.96\sqrt{48}, 102.8 + 1.96\sqrt{48}] = [89.2207, 116.3793]$. The length of 95% credible set is approximately 27. The Bayesian interval is shorter because the posterior variance is smaller than the likelihood variance; this is a consequence of the incorporation of information. The construction of the credible set is illustrated in Figure 1.3.

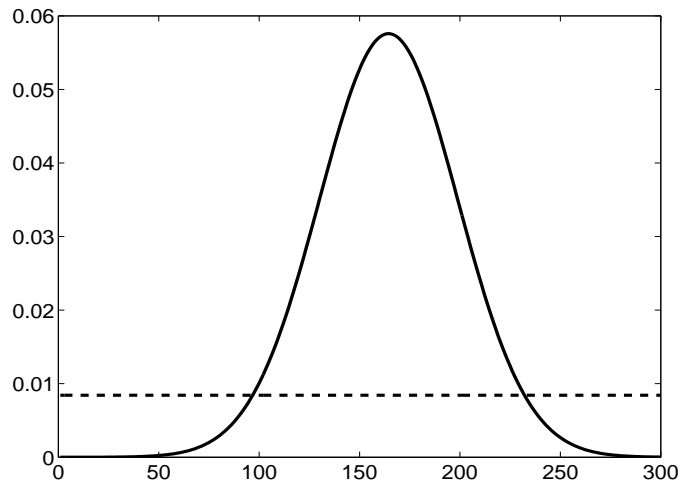


Fig. 1.3 Bayesian credible set based on $\mathcal{N}(102.8, 48)$ density.

1.2.5 Bayesian Testing

Bayesian tests amount to comparison of posterior probabilities of the parameter regions defined by the two hypotheses.

Assume that Θ_0 and Θ_1 are two non-overlapping subsets of the parameter space Θ . We assume that Θ_0 and Θ_1 partition Θ , that is, $\Theta_1 = \Theta_0^c$, although cases in which $\Theta_1 \neq \Theta_0^c$ are easily formulated. Let $\theta \in \Theta_0$ signify the null

hypothesis H_0 and let $\theta \in \Theta_1 = \Theta_0^c$ signify the alternative hypothesis H_1 :

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1.$$

Given the information from the posterior, the hypothesis with higher posterior probability is selected.

Example 1.7 We return again to Jeremy (Examples 1.4 and 1.6) and consider the posterior for the parameter θ , $\mathcal{N}(102.8, 48)$. Jeremy claims he had a bad day and his genuine IQ is at least 105. After all, he is at Georgia Tech! The posterior probability of $\theta \geq 105$ is

$$p_0 = P^{\theta|X}(\theta \geq 105) = P\left(Z \geq \frac{105 - 102.8}{\sqrt{48}}\right) = 1 - \Phi(0.3175) = 0.3754,$$

less than 38%, so his claim is rejected. Posterior odds in favor of H_0 are $0.3754/(1-0.3754)=0.4652$, less than 50%.

We can represent the prior and posterior odds in favor of the hypothesis H_0 , respectively, as

$$\frac{\pi_0}{\pi_1} = \frac{P^\theta(\theta \in \Theta_0)}{P^\theta(\theta \in \Theta_1)} \quad \text{and} \quad \frac{p_0}{p_1} = \frac{P^{\theta|X}(\theta \in \Theta_0)}{P^{\theta|X}(\theta \in \Theta_1)}.$$

The *Bayes factor* in favor of H_0 is the ratio of corresponding posterior to prior odds,

$$B_{01}^\pi(x) = \frac{\frac{P(\theta \in \Theta_0|X)}{P(\theta \in \Theta_1|X)}}{\frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)}} = \frac{p_0/p_1}{\pi_0/\pi_1}. \quad (1.3)$$

When the hypotheses are simple (i.e., $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$), and the prior is just the two point distribution $\pi(\theta_0) = \pi_0$ and $\pi(\theta_1) = \pi_1 = 1 - \pi_0$, then the Bayes factor in favor of H_0 becomes the likelihood ratio:

$$B_{01}^\pi(x) = \frac{\frac{P^{\theta|X}(\theta \in \Theta_0)}{P^{\theta|X}(\theta \in \Theta_1)}}{\frac{P^\theta(\theta \in \Theta_0)}{P^\theta(\theta \in \Theta_1)}} = \frac{f(x|\theta_0)\pi_0}{f(x|\theta_1)\pi_1} / \frac{\pi_0}{\pi_1} = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

If the prior is a mixture of two priors, ξ_0 under H_0 and ξ_1 under H_1 , then the Bayes factor is the ratio of two marginal (prior-predictive) distributions generated by ξ_0 and ξ_1 . Thus, if $\pi(\theta) = \pi_0\xi_0(\theta) + \pi_1\xi_1(\theta)$ then,

$$B_{01}^\pi(x) = \frac{\int_{\Theta_0} f(x|\theta)\pi_0\xi_0(\theta)d\theta}{\int_{\Theta_1} f(x|\theta)\pi_1\xi_1(\theta)d\theta} = \frac{m_0(x)}{m_1(x)}.$$

The Bayes factor measures relative change in prior odds once the evidence

Table 1.2 Treatment of H_0 According to the Value of log-Bayes Factor.

$0 \leq \log B_{10}(x) \leq 0.5$	evidence against H_0 is poor
$0.5 \leq \log B_{10}(x) \leq 1$	evidence against H_0 is substantial
$1 \leq \log B_{10}(x) \leq 2$	evidence against H_0 is strong
$\log B_{10}(x) > 2$	evidence against H_0 is decisive

is collected. Table 1.2 offers practical guidelines for Bayesian testing of hypotheses depending on the value of log-Bayes factor. One could use $B_{01}^\pi(x)$ of course, but then $a \leq \log B_{10}(x) \leq b$ becomes $-b \leq \log B_{01}(x) \leq -a$. Negative values of the log-Bayes factor are handled by using symmetry and changed wording, in an obvious way.

1.2.5.1 Bayesian Testing of Precise Hypotheses Testing precise hypotheses in Bayesian fashion has a considerable body of research. Berger (1985), pp. 148–157, has a comprehensive overview of the problem and provides a wealth of references. See also Berger and Sellke (1984) and Berger and Delampady (1987).

If the priors are continuous, testing precise hypotheses in Bayesian fashion is impossible because with continuous priors and posteriors, the probability of a singleton is 0. Suppose $X|\theta \sim f(x|\theta)$ is observed and we are interested in testing

$$H_0 : \theta = \theta_0 \quad v.s. \quad H_1 : \theta \neq \theta_0.$$

The answer is to have a prior that has a point mass at the value θ_0 with prior weight π_0 and a spread distribution $\xi(\theta)$ which is the prior under H_1 that has prior weight $\pi_1 = 1 - \pi_0$. Thus, the prior is the 2-point mixture

$$\pi(\theta) = \pi_0 \delta_{\theta_0} + \pi_1 \xi(\theta),$$

where δ_{θ_0} is Dirac mass at θ_0 .

The marginal density for X is

$$m(x) = \pi_0 f(x|\theta_0) + \pi_1 \int f(x|\theta) \xi(\theta) d\theta = \pi_0 f(x|\theta_0) + \pi_1 m_1(x).$$

The posterior probability of $\theta = \theta_0$ is

$$\pi(\theta_0|x) = f(x|\theta_0)\pi_0/m(x) = \frac{f(x|\theta_0)\pi_0}{\pi_0 f(x|\theta_0) + \pi_1 m_1(x)} = \left(1 + \frac{\pi_1}{\pi_0} \cdot \frac{m_1(x)}{f(x|\theta_0)}\right)^{-1}.$$

1.2.6 Bayesian Prediction

Statistical prediction fits naturally into the Bayesian framework. Suppose $Y \sim f(y|\theta)$ is to be observed. The posterior predictive distribution of Y , given observed $X = x$ is

$$f(y|x) = \int_{\Theta} f(y|\theta)\pi(\theta|x)d\theta.$$

For example, in the normal distribution example, the predictive distribution of Y , given X_1, \dots, X_n is

$$Y|\bar{X} \sim \mathcal{N}\left(\frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\bar{X} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \tau^2}\mu, \sigma^2 + \frac{\frac{\sigma^2}{n}\tau^2}{\frac{\sigma^2}{n} + \tau^2}\right). \quad (1.4)$$

Example 1.8 Martz and Waller (1985) suggest that Bayesian reliability inference is most helpful in applications where little system failure data exist, but past data from like systems are considered relevant to the present system. They use an example of heat exchanger reliability, where the lifetime X is the failure time for heat exchangers used in refining gasoline. From past research and modeling in this area, it is determined that X has a Weibull distribution with $\kappa = 3.5$. Furthermore, the scale parameter λ is considered to be in the interval $0.5 \leq \lambda \leq 1.5$ with no particular value of λ considered more likely than others.

From this argument, we have

$$\pi(\lambda) = \begin{cases} 1 & 0.5 \leq \lambda \leq 1.5 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x|\lambda) = \kappa\lambda^\kappa x^{\kappa-1} e^{-(x\lambda)^\kappa}$$

where $\kappa = 3.5$. With $n=9$ observed failure times (measured in years of service) at (0.41, 0.58, 0.75, 0.83, 1.00, 1.08, 1.17, 1.25, 1.35), the likelihood is

$$f(x_1, \dots, x_9|\lambda) \propto \lambda^9 \left(\prod_{i=1}^9 x_i^{2.5}\right) e^{-\lambda^{3.5}(\sum x_i^{3.5})},$$

so the sufficient statistic is

$$\sum_{i=1}^9 x_i^{3.5} = 10.16.$$

The resulting posterior distribution is not distributed Weibull (like the likeli-

hood) or uniform (like the prior). It can be expressed as

$$\pi(\lambda|x_1, \dots, x_9) = \begin{cases} (1621.39)\lambda^9 e^{-10.16\lambda^{3.5}} & 0.5 \leq \lambda \leq 1.5 \\ 0 & \text{otherwise,} \end{cases}$$

and has expected value of $\lambda_B = 0.6896$. Figure 1.4(a) shows the posterior density. From the prior distribution, $\mathbb{E}(\lambda) = 1$, so our estimate of λ has decreased in the process of updating the prior with the data.

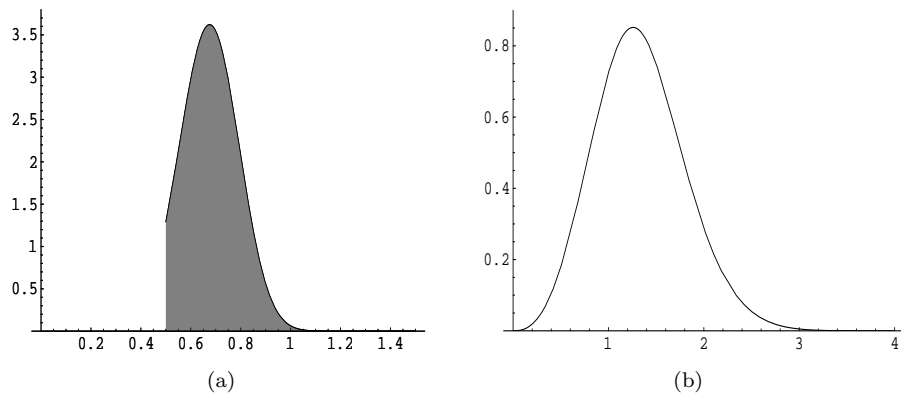


Fig. 1.4 (a) Posterior density for λ ; (b) Posterior predictive density for heat-exchanger lifetime.

Estimation of λ was not the focus of this study; the analysts were interested in predicting future lifetime of a generic (randomly picked) heat exchanger. Using the predictive density from (1.4),

$$f(y|x) = \int_{0.5}^{1.5} \left(3.5\lambda^{3.5} y^{2.5} e^{-(\lambda y)^{3.5}} \right) \left(1621.39\lambda^9 e^{-10.16\lambda^{3.5}} \right) d\lambda.$$

The predictive density is a bit messy, but straightforward to work with. The plot of the density in Figure 1.4(b) shows how uncertainty is gauged for the lifetime of a new heat-exchanger. From $f(y|x)$, we might be interested in predicting early failure by the new item; for example, a 95% lower bound for heat-exchanger lifetime is found by computing the lower 0.05-quantile of $f(y|x)$, which is approximately 0.49.

1.3 BAYESIAN COMPUTATION AND USE OF WINBUGS

If the selection of an adequate prior was the major conceptual and modeling challenge of Bayesian analysis, the major implementational challenge is com-

putation. When the model deviates from the conjugate structure, finding the posterior distribution and the Bayes rule is all but simple. A closed form solution is more an exception than the rule, and even for such exceptions, lucky mathematical coincidences, convenient mixtures, and other tricks are needed to uncover the explicit expression.

If the classical statistics relies on optimization, Bayesian statistics relies on integration. The marginal needed for the posterior is an integral

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta,$$

and the Bayes estimator of $h(\theta)$, with respect to the squared error loss is a ratio of integrals,

$$\delta_{\pi}(x) = \int_{\Theta} h(\theta)\pi(\theta|x)d\theta = \frac{\int_{\Theta} h(\theta)f(x|\theta)\pi(\theta)d\theta}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}.$$

The difficulties in calculating the above Bayes rule come from the facts that (i) the posterior may not be representable in a finite form, and (ii) the integral of $h(\theta)$ does not have a closed form even when the posterior distribution is explicit.

The last two decades of research in Bayesian statistics contributed to broadening the scope of Bayesian models. Models that could not be handled before by a computer are now routinely solved. This is done by *Markov chain Monte Carlo* (MCMC) Methods, and their introduction to the field of statistics revolutionized Bayesian statistics.

The Markov chain Monte Carlo (MCMC) methodology was first applied in statistical physics, (Metropolis et al., 1953). Work by Gelfand and Smith (1990) focused on applications of MCMC to Bayesian models. The principle of MCMC is simple: to sample randomly from a target probability distribution one designs a Markov chain whose stationary distribution is the target distribution. By simulating long runs of such a Markov chain, the target distribution can be well approximated. Various strategies for constructing appropriate Markov chains that simulate form the desired distribution are possible: Metropolis-Hastings, Gibbs sampler, slice sampling, perfect sampling, and many specialized techniques. They are beyond the scope of this text and the interested reader is directed to Robert (2001), Robert and Casella (2004), and Chen, Shao, and Ibrahim (2000), for an overview and a comprehensive treatment.

In the examples that follow we will use WinBUGS for doing Bayesian inference when the models are not conjugate. The Appendix gives a brief introduction to the front-end of WinBUGS. Three volumes of examples are standard addition to the software, in the Examples menu of WinBUGS, see Spiegelhalter, Thomas, Best, and Gilks (1996). It is recommended that you go over some of those in detail because they illustrate functionality and real modeling power of WinBUGS. A wealth of examples on Bayesian modeling strategies

using WinBUGS can be found in the monographs of Congdon (2001, 2003, 2005).

The following example demonstrates a simulational power of WinBUGS, it involves approximating probabilities of complex events and has nothing to do with Bayesian inference.

Example 1.9 Paradox DeMere in WinBUGS. In 1654 the Chevalier de Mere asked Blaise Pascal (1623–1662) the following question: *In playing a game with three dice why the sum 11 is advantageous to sum 12 when both are results of six possible outcomes?* Indeed, there are six favorable triplets for each of the sums **11** and **12**,

11:	(1, 4, 6), (1, 5, 5), (2, 3, 6), (2, 4, 5), (3, 3, 5), (3, 4, 4)
12:	(1, 5, 6), (2, 4, 6), (2, 5, 5), (3, 3, 6), (3, 4, 5), (4, 4, 4)

The solution to the “paradox” deMere is simple. By taking into account all possible permutations of the above triples the sum 11 has 27 favorable permutations while the sum 12 has 25 favorable permutation.

But what if 300 fair dice are rolled and we are interested if the sum 1111 is advantageous to the sum 1112? Exact solution is unappealing, but the probabilities can be well approximated by WinBUGS model `demere1`.

```
model demere1;
{
  for (i in 1:300) {
    dice[i] ~ dcat(p.dice[]);
  }
  is1111 <- equals(sum(dice[]),1111)
  is1112 <- equals(sum(dice[]),1112)
}
```

The data are

```
list(p.dice=c(0.1666666, 0.1666666,
0.1666667, 0.1666667, 0.1666667, 0.1666667) )
```

and the initial values are generated. After five million rolls, WinBUGS outputs `is1111 = 0.0016` and `is1112 = 0.0015`, so the sum of 1111 is advantageous to the sum of 1112.

Example 1.10 Jeremy in WinBUGS. We will calculate a Bayes estimator for Jeremy’s true IQ, θ , using BUGS. Recall, the model in Example 1.4 was $X \sim \mathcal{N}(\theta, 80)$ and $\theta \sim \mathcal{N}(100, 120)$. In WinBUGS we will use the precision parameters $1/120 = 0.00833$ and $1/80 = 0.0125$.

```
#Jeremy in WinBUGS
model{
```

```
x ~ dnorm( theta, tau)
theta ~ dnorm( 110, 0.008333333)
}
#data
list( tau=0.0125, x=98)
#inits
list(theta=100)
```

Here is the summary of MCMC output. The Bayes estimator for θ is rounded to 102.8.

node	mean	sd	MC error	2.5%	median	97.5%
θ	102.8	6.917	0.0214	89.17	102.8	116.3

Because this is a conjugate normal/normal model, the exact posterior distribution, $\mathcal{N}(102.8, 48)$, was easy to find, (see Example 1.4). Note that in simulations, the MCMC approximation, when rounded, coincides with the exact posterior mean. The MCMC variance of θ is $6.917^2 = 47.84489$, close to the exact posterior variance of 48.

1.4 EXERCISES

- A lifetime X (in years) of a particular machine is modeled by an exponential distribution with unknown failure rate parameter θ . The lifetimes of $X_1 = 5$, $X_2 = 6$, and $X_3 = 4$ are observed, and assume that an expert believes that θ should have exponential distribution as well and that, on average θ should be $1/3$.
 - Write down the MLE of θ for those observations.
 - Elicit a prior according to the expert's beliefs.
 - For the prior in (ii), find the posterior. Is the problem conjugate?
 - Find the Bayes estimator $\hat{\theta}_{Bayes}$, and compare it with the MLE estimator from (i). Discuss.
- Suppose $X = (X_1, \dots, X_n)$ is a sample from $\mathcal{U}(0, \theta)$. Let θ have Pareto $\mathcal{Pa}(\theta_0, \alpha)$ distribution. Show that the posterior distribution is $\mathcal{Pa}(\max\{\theta_0, x_1, \dots, x_n\}, \alpha + n)$.
- Let $X \sim \mathcal{G}(n/2, 2\theta)$, so that X/θ is χ_n^2 . Let $\theta \sim \mathcal{IG}(\alpha, \beta)$. Show that the posterior is $\mathcal{IG}(n/2 + \alpha, (x/2 + \beta^{-1})^{-1})$.
- If $X = (X_1, \dots, X_n)$ is a sample from $\mathcal{NB}(m, \theta)$ and $\theta \sim \mathcal{Be}(\alpha, \beta)$, show that the posterior for θ is beta $\mathcal{Be}(\alpha + mn, \beta + \sum_{i=1}^n x_i)$.
- In Example 1.5 on p. 7, show that the marginal distribution is negative binomial.

6. What is the Bayes factor B_{01}^π in Jeremy's case? Test H_0 is using the Bayes factor and wording from the Table 1.2. Argue that the evidence against H_0 is poor.
7. Assume $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \pi(\theta) = 1$. Consider testing $H_0 : \theta \leq \theta_0$ v.s. $H_1 : \theta > \theta_0$. Show that $p_0 = P^{\theta|X}(\theta \leq \theta_0)$ is equal to the classical p -value.
8. Show that the Bayes factor is $B_{01}^\pi(x) = f(x|\theta_0)/m_1(x)$.
9. Show that

$$p_0 = \pi(\theta_0|x) \geq \left[1 + \frac{\pi_1}{\pi_0} \cdot \frac{r(x)}{f(x|\theta_0)} \right]^{-1},$$

where $r(x) = \sup_{\theta \neq \theta_0} f(x|\theta)$. Usually, $r(x) = f(x|\hat{\theta}_{MLE})$, where $\hat{\theta}_{MLE}$ is MLE estimator of θ . The Bayes factor $B_{01}^\pi(x)$ is bounded from below:

$$B_{01}^\pi(x) \geq \frac{f(x|\theta_0)}{r(x)} = \frac{f(x|\theta_0)}{f(x|\hat{\theta}_{MLE})}.$$

10. Suppose $X = -2$ was observed from the population distributed as $N(0, 1/\theta)$ and one wishes to estimate the parameter θ . (Here θ is the reciprocal of variance σ^2 and is called the *precision parameter*. The precision parameter is used in WinBUGS to parameterize the normal distribution). A classical estimator of θ (e.g., the MLE) does exist, but one may be disturbed to estimate $1/\sigma^2$ based on a single observation. Suppose the analyst believes that the prior on θ is $\mathcal{Gamma}(1/2, 3)$.
 - (i) What is the MLE of θ ?
 - (ii) Find the posterior distribution and the Bayes estimator of θ . If the prior on θ is $\mathcal{Gamma}(\alpha, \beta)$, represent the Bayes estimator as weighted average (sum of weights = 1) of the prior mean and the MLE.
 - (iii) Find a 95% HPD Credible set for θ .
 - (iv) Test the hypothesis $H_0 : \theta \leq 1/4$ versus $H_1 : \theta > 1/4$.
11. *The Lindley (1957) Paradox.* Suppose $\bar{y}|\theta \sim N(\theta, 1/n)$. We wish to test $H_0 : \theta = 0$ versus the two sided alternative. Suppose a Bayesian puts the prior $P(\theta = 0) = P(\theta \neq 0) = 1/2$, and in the case of the alternative, the $1/2$ is uniformly spread over the interval $[-M/2, M/2]$. Suppose $n = 40,000$ and $\bar{y} = 0.01$ are observed, so $\sqrt{n} \bar{y} = 2$. The classical statistician rejects H_0 at level $\alpha = 0.05$. Show that posterior odds in favor of H_0 are 11 if $M = 1$, indicating that a Bayesian statistician strongly favors H_0 , according to Table 1.2.
12. This exercise concerning Bayesian binary regression with a probit model using WinBUGS is borrowed from David Madigan's Bayesian Course Site. Finney (1947) describes a binary regression problem with data of

size $n = 39$, two continuous predictors x_1 and x_2 , and a binary response y . Here are the data in BUGS-ready format:

```
list(n=39,x1=c(3.7,3.5,1.25,0.75,0.8,0.7,0.6,1.1,0.9,0.9,0.8,0.55,0.6,1.4,
0.75,2.3,3.2,0.85,1.7,1.8,0.4,0.95,1.35,1.5,1.6,0.6,1.8,0.95,1.9,1.6,2.7,
2.35,1.1,1.1,1.2,0.8,0.95,0.75,1.3),
x2=c(0.825,1.09,2.5,1.5,3.2,3.5,0.75,1.7,0.75,0.45,0.57,2.75,3.0,2.33,3.75,
1.64,1.6,1.415,1.06,1.8,2.0,1.36,1.35,1.36,1.78,1.5,1.5,1.9,0.95,0.4,0.75,
0.03,1.83,2.2,2.0,3.33,1.9,1.9,1.625),
y=c(1,1,1,1,1,0,0,0,0,0,0,0,1,1,1,1,0,1,0,0,0,0,1,0,1,0,1,0,1,0,0,1,1,
1,0,0,1))
```

The objective is to build a predictive model that predicts y from x_1 and x_2 . Proposed approach is the probit model: $P(y = 1|x_1, x_2) = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ where Φ is the standard normal CDF.

(i) Use WinBUGS to compute posterior distributions for β_0, β_1 and β_2 using diffuse normal priors for each.

(ii) Suppose instead of the diffuse normal prior for β_i , $i = 0, 1, 2$, you use a normal prior with mean zero and variance v_i , and assume the v_i s are independently exponentially distributed with some hyperparameter γ . Fit this model using BUGS. How different are the two posterior distributions from this exercise?

13. The following WinBUGS code flips a coin, the outcome H is coded by 1 and tails by 0. Mimic the following code to simulate a rolling of a fair die.

```
#coin.bug:
model coin;
{
flip12 ~ dcat(p.coin[])
coin <- flip12 - 1
}
#coin.dat:
list(p.coin=c(0.5, 0.5))
# just generate initials
```

14. The highly publicized (recent TV reports) *in vitro fertilization* success cases for women in their late fifties all involve donor's egg. If the egg is the woman's own, the story is quite different.

In vitro fertilization (IVF), one of the assisted reproductive technology (ART) procedures, involves extracting a woman's eggs, fertilizing the eggs in the laboratory, and then transferring the resulting embryos into the woman's uterus through the cervix. Fertilization involves a specialized technique known as intracytoplasmic sperm injection (ICSI).

The table shows the live-birth success rate per transfer rate from the recipients' eggs, stratified by age of recipient. The data are for year 1999, published by US - Centers for Disease Control and Prevention (CDC): (<http://www.cdc.gov/reproductivehealth/ART99/index99.htm>)

Age (x)	24	25	26	27	28	29	30	31
Percentage (y)	38.7	38.6	38.9	41.4	39.7	41.1	38.7	37.6
Age (x)	32	33	34	35	36	37	38	39
Percentage(y)	36.3	36.9	35.7	33.8	33.2	30.1	27.8	22.7
Age (x)	40	41	42	43	44	45	46	
Percentage(y)	21.3	15.4	11.2	9.2	5.4	3.0	1.6	

Assume the change-point regression model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, \tau \\ y_i &= \gamma_0 + \gamma_1 x_i + \epsilon_i, \quad i = \tau + 1, \dots, n \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

- (i) Propose priors (with possibly hyperpriors) on σ^2 , β_0 , β_1 , γ_0 , and γ_1 .
- (ii) Take discrete uniform prior on τ . Write a program for this exercise in WinBUGS.

15. Is the cloning of humans moral? Recent Gallup Poll estimates that about 88% Americans opposed cloning humans. Results are based on telephone interviews with a randomly selected national sample of $n = 1000$ adults, aged 18 and older, conducted May 2-4, 2004. In these 1000 interviews, 882 adults opposed cloning humans.

- (i) Write WinBUGS program to estimate the proportion p of people opposed to cloning humans. Use a non-informative prior for p .
- (ii) Test the hypothesis that $p \leq 0.87$.
- (iii) Pretend that the original poll had $n = 1062$ adults, i.e., results for 62 adults are missing. Estimate the number of people opposed to cloning among the 62 missing in the poll.

Hint:

```
model {
  anticlons ~ dbin(prob,npolled) ;
  lessthan87 <- step(prob-0.87)
  anticlons.missing ~ dbin(prob,nmissing)
  prob ~ dbeta(1,1)}
```

Data

```
list(anticlons=882,npolled= 1000, nmissing=62)
```

References

- Anscombe, F. J. (1962), "Tests of Goodness of Fit," *Journal of the Royal Statistical Society (B)*, 25, 81-94.
- Bayes, T. (1763), "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society, London*, 53, 370-418.
- Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, New York: Springer-Verlag.
- Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypothesis," *Statistical Science*, 2, 317-352.
- Berger, J. O., and Selke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p -values and Evidence (with Discussion)", *Journal of American Statistical Association*, 82, 112-122.
- Chen, M.-H., Shao, Q.-M., and Ibrahim, J. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer Verlag.
- Congdon, P. (2001), *Bayesian Statistical Modelling*, Hoboken, NJ: Wiley.
- Congdon, P. (2003), *Applied Bayesian Models*, Hoboken, NJ: Wiley.
- Congdon, P. (2005), *Bayesian Models for Categorical Data*, Hoboken, NJ: Wiley.
- Finney, D. J. (1947), "The Estimation from Individual Records of the Relationship Between Dose and Quantal Response," *Biometrika*, 34, 320-334.
- Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of American Statistical Association*, 85, 398-409.
- Lindley, D. V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187-192.
- Madigan, D. <http://stat.rutgers.edu/madigan/bayes02/>. A Web Site for Course on Bayesian Statistics.
- Martz, H., and Waller, R. (1985), *Bayesian Reliability Analysis*, New York: Wiley.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21, 1087-1092.
- Robert, C. (2001), *The Bayesian Choice: From Decision-Theoretic Motivations to Computational Implementation*, Second Edition, New York: Springer Verlag.
- Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Second Edition, New York: Springer Verlag.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996), "BUGS Examples Volume 1," Version 0.5. Cambridge: Medical Research Council Biostatistics Unit (PDF).

WinBUGS



Beware: MCMC sampling can be dangerous! (Disclaimer in WinBUGS User Manual)

BUGS is freely available software for constructing Bayesian statistical models and evaluating them using MCMC methodology.

BUGS and WINBUGS are distributed freely and are the result of many years of development by a team of statisticians and programmers at the Medical research Council Biostatistics Research Unit in Cambridge (BUGS and WinBUGS), and from recently by a team at University of Helsinki (OpenBUGS) see the project pages: <http://www.mrc-bsu.cam.ac.uk/bugs/> and <http://mathstat.helsinki.fi/openbugs/>.

Models are represented by a flexible language, and there is also a graphical feature, DOODLEBUGS, that allows users to specify their models as directed graphs. For complex models the DOODLEBUGS can be very useful. As of April 2007, the latest version of WinBUGS is 1.4.1 and OpenBUGS 3.0.

We start the introduction to WinBUGS with a simple regression example. Consider the model

$$\begin{aligned}
 y_i | \mu_i, \tau &\sim \mathcal{N}(\mu_i, \tau), \quad i = 1, \dots, n \\
 \mu_i &= \alpha + \beta(x_i - \bar{x}), \\
 \alpha &\sim \mathcal{N}(0, 10^{-4}) \\
 \beta &\sim \mathcal{N}(0, 10^{-4}) \\
 \tau &\sim \text{Gamma}(0.001, 0.001).
 \end{aligned}$$

The scale in normal distributions here is parameterized in terms of a *precision* parameter τ which is the reciprocal of variance, $\tau = 1/\sigma^2$. Natural distributions for the precision parameters are Gamma and small values of the precision reflect the flatness (noninformativeness) of the priors. Assume that (x, y) -pairs $(1, 1)$, $(2, 3)$, $(3, 3)$, $(4, 3)$, and $(5, 5)$ are observed.

Estimators in classical, Least Square regression of y on $x - \bar{x}$, are given in the following table.

Coef	LSEstimate	SE Coef	t	p
ALPHA	3.0000	0.3266	9.19	0.003
BETA	0.8000	0.2309	3.46	0.041
S = 0.730297	R-Sq = 80.0%	R-Sq(adj) = 73.3%		

How about Bayesian estimators? We will find the estimators by MCMC calculations as means on the simulated posteriors. Assume that the initial values of parameters are $\alpha_0 = 0.1$, $\beta_0 = 0.6$, and $\tau = 1$. Start BUGS and input the following code in [File > New].

```
# A simple regression
model{
  for (i in 1:N) {
    Y[i] ~ dnorm(mu[i],tau);
    mu[i] <- alpha + beta * (x[i] - x.bar);
  }
  x.bar <- mean(x[]);
  alpha ~ dnorm(0, 0.0001);
  beta ~ dnorm(0, 0.0001);
  tau ~ dgamma(0.001, 0.001);
  sigma <- 1.0/sqrt(tau);
}
#-----
#these are observations
list( x=c(1,2,3,4,5), Y=c(1,3,3,3,5), N=5);
#-----
#the initials
list(alpha = 0.1, beta = 0.6, tau = 1);
```

Next, highlight the line in the code that contains word “model.” Go to the **Model** menu and open **Specification**. The **Specification Tool** window will pop-out. If your model is highlighted, you may **check model** in the specification tool window. If the model is correct, the response on the lower bar of the BUGS window should be: **model is syntactically correct**. Next, highlight the “list” statement in the data-part of your code. In the Specification Tool window select **load data**. If the data are in correct format, you should receive response on the bottom bar of BUGS window: **data loaded**. You will need to compile your model on order to activate **inits**-buttons. Select **compile** in the Specification Tool window. The response should be: **model compiled**, and the buttons **load inits** and **gen inits** become active. Finally, highlight the “list” statement in the initials-part of your code and in the Specification Tool window select **load inits**. The response should be: **model is initialized**, and this finishes reading in the model. If the response is **initial values loaded but this or other chain contain uninitialized**

variables, click on the **gen inits** button. The response should be: **initial values generated, model initialized.**

Now, you are ready to Burn-in some simulations and at the same time check that the program is working. In the **Model** menu, choose **Update...** and open **Update Tool** to check if your model updates.

From the **Inference** menu, open **Samples....** A window titled **Sample Monitor Tool** will pop out. In the **node** sub-window input the names of the variables you want to monitor. In this case, the variables are **alpha**, **beta**, and **tau**. If you correctly input the variable the **set** button becomes active and you should set the variable. Do this for all 3 variables of interest. In fact, **sigma** as transformation of **tau** is available, as well.

Now choose **alpha** from the subwindow in **Sample Monitor Tool**. All of the buttons (**clear**, **set**, **trace**, **history**, **density**, **stats**, **coda**, **quantiles**, **bgr diag**, **auto cor**) are now active. Return to **Update Tool** and select the desired number of simulations, say 10000, in the **updates** subwindow. Press the **update** button.

Return to **Sample Monitor Tool** and check **trace** for the part of MC trace for α , **history** for the complete trace, **density** for a density estimator of α , etc. For example, pressing **stats** button will produce something like the following table

	mean	sd	MCError	val2.5pc	median	val97.5pc	start	sample
alpha	3.003	0.549	0.003614	1.977	3.004	4.057	10000	20001

The mean 3.003 is the Bayes estimator (as the mean from the sample from the posterior for α). There are two precision outputs, **sd** and **MCError**. The former is an estimator of the standard deviation of the posterior and can be improved by increasing the sample size but not the number of simulations. The later one is the error of simulation and can be improved by additional simulations. The 95% credible set is bounded by **val2.5pc** and **val97.5pc**, which are the 0.025 and 0.975 (empirical) quantiles from the posterior. The empirical median of the posterior is given by **median**. The outputs **start** and **sample** show the starting index for the simulations (after burn-in) and the available number of simulations.

For all parameters a comparative table is

	mean	sd	MCError	val2.5pc	median	val97.5pc	start	sample
alpha	3.003	0.549	0.003614	1.977	3.004	4.057	10000	20001
beta	0.7994	0.3768	0.002897	0.07088	0.7988	1.534	10000	20001
tau	1.875	1.521	0.01574	0.1399	1.471	5.851	10000	20001
sigma	1.006	0.7153	0.009742	0.4134	0.8244	2.674	10000	20001

If you want to save the trace for α in a file and process it in MATLAB, say, select **coda** and the data window will open with an information window as well. Keep the data window active and select **Save As** from the **File**

menu. Save the α s in `alphas.txt` where it will be ready to be imported to MATLAB.

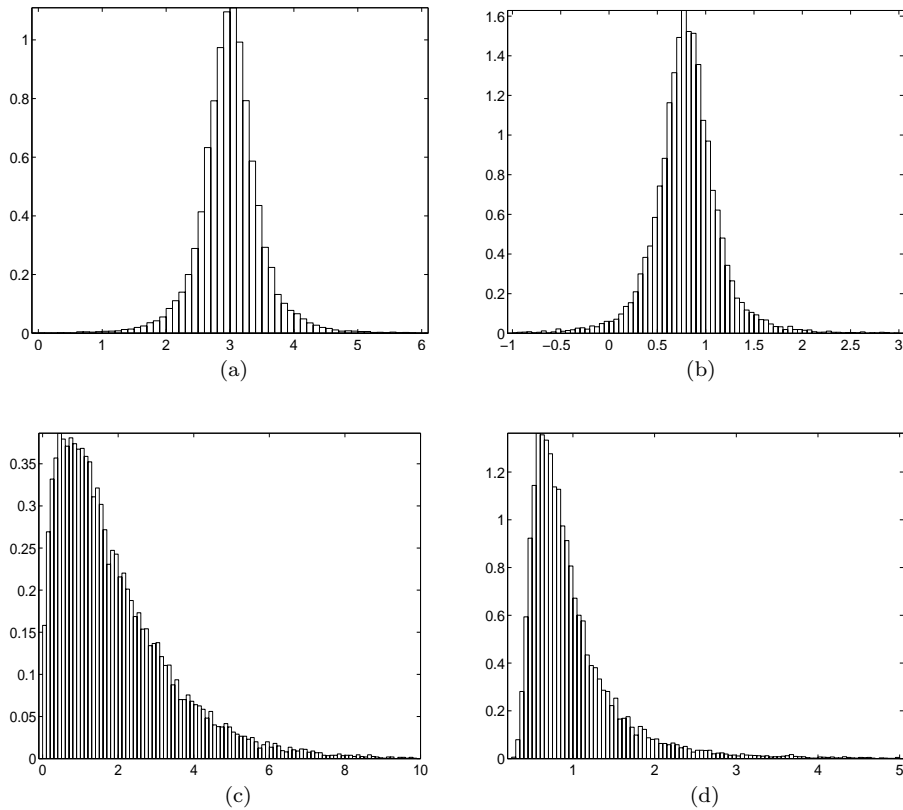


Fig. 1.5 Traces of the four parameters from simple example: (a) α , (b) β , (c) τ , and (d) σ from WinBUGS. Data are plotted in MATLAB after being exported from WinBUGS.

Kevin Murphy lead the project for communication between WinBUGS and MATLAB:

<http://www.cs.ubc.ca/~murphyk/Software/MATBUGS/matbugs.html>.

His suite MATBUGS, maintained by several researchers, communicates with WinBUGS directly from MATLAB.

Built-in Functions and Common Distributions in BUGS

This section contains two tables: one with the list of built-in functions and the second with the list of available distributions.

The first-time WinBUGS user may be disappointed by the selection of built in functions – the set is minimal but sufficient. The full list of distributions in WinBUGS can be found in **Help>WinBUGS User Manual** under **The BUGS language: stochastic nodes>Distributions**. BUGS also allows for construction of distributions for which are not in default list. In Table 1.4 a list of important continuous and discrete distributions, with their BUGS syntax and parametrization, is provided. BUGS has the capability to define custom distributions, both as likelihood or as a prior, via the so called *zero-Poisson device*.

Table 1.3 Built-in Functions in WinBUGS

BUGS Code	function
<code>abs(y)</code>	$ y $
<code>cloglog(y)</code>	$\ln(-\ln(1-y))$
<code>cos(y)</code>	$\cos(y)$
<code>equals(y, z)</code>	1 if $y = z$; 0 otherwise
<code>exp(y)</code>	$\exp(y)$
<code>inprod(y, z)</code>	$\sum_i y_i z_i$
<code>inverse(y)</code>	y^{-1} for symmetric positive-definite matrix y
<code>log(y)</code>	$\ln(y)$
<code>logfact(y)</code>	$\ln(y!)$
<code>loggam(y)</code>	$\ln(\Gamma(y))$
<code>logit(y)</code>	$\ln(y/(1-y))$
<code>max(y, z)</code>	y if $y > z$; z otherwise
<code>mean(y)</code>	$n^{-1} \sum_i y_i$, $n = \dim(y)$
<code>min(y, z)</code>	y if $y < z$; z otherwise
<code>phi(y)</code>	standard normal cdf $\Phi(y)$
<code>pow(y, z)</code>	y^z
<code>sin(y)</code>	$\sin(y)$
<code>sqrt(y)</code>	\sqrt{y}
<code>rank(v, s)</code>	number of components of v less than or equal to v_s
<code>ranked(v, s)</code>	the s th smallest component of v
<code>round(y)</code>	nearest integer to y
<code>sd(v)</code>	standard deviation of components of y ($n-1$ in denom.)
<code>step(y)</code>	1 if $y \geq 0$; 0 otherwise
<code>sum(y)</code>	$\sum_i y_i$
<code>trunc(y)</code>	greatest integer less than or equal to y

Table 1.4 Built-in distributions with BUGS names and their parametrizations.

Distribution	BUGS Code	Density
Bernoulli	<code>x ~ dbern(p)</code>	$p^x(1-p)^{1-x}, x = 0, 1; 0 \leq p \leq 1$
Binomial	<code>x ~ dbin(p, n)</code>	$\binom{n}{x} p^x (1-p)^{n-x}, x = 0, \dots, n; 0 \leq p \leq 1$
Categorical	<code>x ~ dcat(p[])</code>	$p[x], x = 1, 2, \dots, \dim(p)$
Poisson	<code>x ~ dpois(lambda)</code>	$\frac{\lambda^x}{x!} \exp\{-\lambda\}, x = 0, 1, 2, \dots, \lambda > 0$
Beta	<code>x ~ dbeta(a, b)</code>	$\frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, 0 = x \leq 1, a, b > -1$
Chi-square	<code>x ~ dchisqr(k)</code>	$\frac{x^{k/2-1} \exp\{-x/2\}}{2^{k/2} \Gamma(k/2)}, x \geq 0, k > 0$
Double Exponential	<code>x ~ ddexp(mu, tau)</code>	$\frac{\tau}{2} \exp\{-\tau x-\mu \}, x \in \mathbb{R}, \tau > 0, \mu \in \mathbb{R}$
Exponential	<code>x ~ dexp(lambda)</code>	$\lambda \exp\{-\lambda x\}, x \geq 0, \lambda \geq 0$
Flat	<code>x ~ dflat()</code>	constant; not a proper density
Gamma	<code>x ~ dgamma(a, b)</code>	$\frac{b^a x^{a-1}}{\Gamma(a)} \exp(-bx), x, a, b > 0$
Normal	<code>x ~ dnorm(mu, tau)</code>	$\sqrt{\tau/(2\pi)} \exp\{-\frac{\tau}{2}(x-\mu)^2\}, x, \mu \in \mathbb{R}, \tau > 0$
Pareto	<code>x ~ dpar(alpha, c)</code>	$\alpha c^\alpha x^{-(\alpha+1)}, x > c$
Student-t	<code>x ~ dt(mu, tau, k)</code>	$\frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \sqrt{\frac{\tau}{k\pi}} [1 + \frac{\tau}{k}(x-\mu)^2]^{-(k+1)/2}, x \in \mathbb{R}, k \geq 2$
Uniform	<code>x ~ dunif(a, b)</code>	$\frac{1}{b-a}, a \leq x \leq b$
Weibull	<code>x ~ dweib(v, lambda)</code>	$v \lambda x^{v-1} \exp\{-\lambda x^v\}, x, v, \lambda > 0,$
Multinomial	<code>x[] ~ dmulti(p[], N)</code>	$\frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_i^{x_i}, \sum_i x_i = N, 0 < p_i < 1, \sum_i p_i = 1$
Dirichlet	<code>p[] ~ ddirch(alpha[])</code>	$\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i p_i^{\alpha_i-1}, 0 < p_i < 1, \sum_i p_i = 1$
Multivariate Normal	<code>x[] ~ dnmnorm(mu[], T[,])</code>	$(2\pi)^{-d/2} T ^{1/2} \exp\{-1/2(x-\mu)'T(x-\mu)\}, x \in \mathbb{R}^d$
Multivariate Student-t	<code>x[] ~ dmt(mu[], T[,], k)</code>	$\frac{\Gamma((k+d)/2)}{\Gamma(k/2)} \frac{ T ^{1/2}}{k^{d/2} \pi^{d/2}} [1 + \frac{1}{k}(x-\mu)'T(x-\mu)]^{-(k+d)/2}, x \in \mathbb{R}^d, k \geq 2$
Wishart	<code>x[,] ~ dwish(R[,], J, k)</code>	$ R ^{k/2} x ^{(k-p-1)/2} \exp\{-1/2T\tau(Rx)\}$