INTRODUCTION TO BAYESIAN ANALYSIS

Arto Luoma
University of Tampere, Finland

Autumn 2014

Who was Thomas Bayes?

Thomas Bayes (1701-1761) was an English philosopher and Presbyterian minister. In his later years he took a deep interest in probability. He suggested a solution to a problem of inverse probability. What do we know about the probability of success if the number of successes is recorded in a binomial experiment? Richard Price discovered Bayes' essay and published it posthumously. He believed that Bayes' Theorem helped prove the existence of God.

Bayesian paradigm:

posterior information = prior information + data information

Bayesian paradigm:

posterior information = prior information + data information

More formally:

$$p(\theta|y) \propto p(\theta)p(y|\theta),$$

where $\propto$ is a symbol for proportionality, θ is an unknown parameter, y is data, and $p(\theta)$, $p(\theta|y)$ and $p(y|\theta)$ are the density functions of the prior, posterior and sampling distributions, respectively.

Bayesian paradigm:

posterior information = prior information + data information

More formally:

$$p(\theta|y) \propto p(\theta)p(y|\theta),$$

where $\propto$ is a symbol for proportionality, θ is an unknown parameter, y is data, and $p(\theta)$, $p(\theta|y)$ and $p(y|\theta)$ are the density functions of the prior, posterior and sampling distributions, respectively.

In Bayesian inference, the unknown parameter θ is considered stochastic, unlike in classical inference. The distributions $p(\theta)$ and $p(\theta|y)$ express uncertainty about the exact value of θ. The density of data, $p(y|\theta)$, provides information from the data. It is called a likelihood function when considered a function of θ.

Software for Bayesian Statistics

In this course we use the R and BUGS programming languages. BUGS stands for **B**ayesian inference **U**sing **G**ibbs **S**ampling. Gibbs sampling was the computational technique first adopted for Bayesian analysis. The goal of the BUGS project is to separate the ”knowledge base” from the ”inference machine” used to draw conclusions. BUGS language is able to describe complex models using very limited syntax.

Software for Bayesian Statistics

In this course we use the R and BUGS programming languages. BUGS stands for **B**ayesian inference **U**sing **G**ibbs **S**ampling. Gibbs sampling was the computational technique first adopted for Bayesian analysis. The goal of the BUGS project is to separate the "knowledge base" from the "inference machine" used to draw conclusions. BUGS language is able to describe complex models using very limited syntax.

There are three widely used BUGS implementations: WinBUGS, OpenBUGS and JAGS. Both WinBUGS and OpenBUGS have a Windows GUI. Further, each engine can be controlled from R. In this course we introduce rjags, the R interface to JAGS.

Contents of the course

Basic concepts

Single-parameter models

Hypothesis testing

Simple multiparameter models

Markov chains

MCMC methods

Model checking and comparison

Hierarchical and regression models

Categorical data

Basic concepts

Bayes' theorem

Let $A_1, A_2, ..., A_k$ be events that partition the sample space Ω, (i.e. $\Omega = A_1 \cup A_2 \cup ... \cup A_k$ and $A_i \cap A_j = \emptyset$ when $i \neq j$) and let B an event on that space for which $\Pr(B) > 0$. Then Bayes' theorem is

$$\Pr(A_j|B) = \frac{\Pr(A_j)\Pr(B|A_j)}{\sum_{j=1}^{k}\Pr(A_j)\Pr(B|A_j)}.$$

Bayes' theorem

Let $A_1, A_2, ..., A_k$ be events that partition the sample space Ω, (i.e. $\Omega = A_1 \cup A_2 \cup ... \cup A_k$ and $A_i \cap A_j = \emptyset$ when $i \neq j$) and let B an event on that space for which $\Pr(B) > 0$. Then Bayes' theorem is

$$\Pr(A_j|B) = \frac{\Pr(A_j)\Pr(B|A_j)}{\sum_{j=1}^{k} \Pr(A_j)\Pr(B|A_j)}.$$

This formula can be used to reverse conditional probabilities. If one knows the probabilities of the events A_j and the conditional probabilities $\Pr(B|A_j)$, $j = 1, ..., k$, the formula can be used to compute the conditinal probabilites $\Pr(A_j|B)$.

Example (Diagnostic tests)

A disease occurs with prevalence γ in population, and θ indicates that an individual has the disease. Hence $\Pr(\theta = 1) = \gamma$, $\Pr(\theta = 0) = 1 - \gamma$. A diagnostic test gives a result Y, whose distribution function is $F_1(y)$ for a diseased individual and $F_0(y)$ otherwise. The most common type of test declares that a person is diseased if $Y > y_0$, where y_0 is fixed on the basis of past data.

Example (Diagnostic tests)

A disease occurs with prevalence γ in population, and θ indicates that an individual has the disease. Hence $\Pr(\theta = 1) = \gamma$, $\Pr(\theta = 0) = 1 - \gamma$. A diagnostic test gives a result Y, whose distribution function is $F_1(y)$ for a diseased individual and $F_0(y)$ otherwise. The most common type of test declares that a person is diseased if $Y > y_0$, where y_0 is fixed on the basis of past data. The probability that a person is diseased, given a positive test result, is

$$\begin{aligned} & \Pr(\theta = 1 | Y > y_0) \\ = & \frac{\gamma[1 - F_1(y_0)]}{\gamma[1 - F_1(y_0)] + (1 - \gamma)[1 - F_0(y_0)]}. \end{aligned}$$

Example (Diagnostic tests)

A disease occurs with prevalence γ in population, and θ indicates that an individual has the disease. Hence $\Pr(\theta = 1) = \gamma$, $\Pr(\theta = 0) = 1 - \gamma$. A diagnostic test gives a result Y, whose distribution function is $F_1(y)$ for a diseased individual and $F_0(y)$ otherwise. The most common type of test declares that a person is diseased if $Y > y_0$, where y_0 is fixed on the basis of past data. The probability that a person is diseased, given a positive test result, is

$$\begin{aligned} & \Pr(\theta = 1 | Y > y_0) \\ = & \frac{\gamma[1 - F_1(y_0)]}{\gamma[1 - F_1(y_0)] + (1 - \gamma)[1 - F_0(y_0)]}. \end{aligned}$$

This is sometimes called the *positive predictive value* of test. Its *sensitivity* and *specifity* are $1 - F_1(y_0)$ and $F_0(y_0)$.

(Example from Davison, 2003).

Prior and posterior distributions

In a more general case, θ can take a finite number of values, labelled $1, ..., k$. We can assign to these values probabilites $p_1, ..., p_k$ which express our beliefs about θ before we have access to the data. The data y are assumed to be the observed value of a (multidimensional) random variable Y, and $p(y|\theta)$ the density of y given θ (the likelihood function).

Prior and posterior distributions

In a more general case, θ can take a finite number of values, labelled $1, ..., k$. We can assign to these values probabilites $p_1, ..., p_k$ which express our beliefs about θ before we have access to the data. The data y are assumed to be the observed value of a (multidimensional) random variable Y, and $p(y|\theta)$ the density of y given θ (the likelihood function). Then the conditional probabilites

$$\Pr(\theta = j|Y = y) = \frac{p_j p(y|\theta = j)}{\sum_{i=1}^{k} p_i p(y|\theta = i)}, \quad j = 1, ..., k,$$

summarize our beliefs about θ after we have observed Y.

Prior and posterior distributions

In a more general case, θ can take a finite number of values, labelled $1, ..., k$. We can assign to these values probabilites $p_1, ..., p_k$ which express our beliefs about θ before we have access to the data. The data y are assumed to be the observed value of a (multidimensional) random variable Y, and $p(y|\theta)$ the density of y given θ (the likelihood function). Then the conditional probabilites

$$\Pr(\theta = j|Y = y) = \frac{p_j p(y|\theta = j)}{\sum_{i=1}^{k} p_i p(y|\theta = i)}, \quad j = 1, ..., k,$$

summarize our beliefs about θ after we have observed Y.

The unconditional probabilities $p_1, ..., p_k$ are called *prior probablities* and $\Pr(\theta = 1|Y = y), ..., \Pr(\theta = k|Y = y)$ are called *posterior probabilites* of θ.

Prior and posterior distributions (2)

When θ can get values continuosly on some interval, we can express our beliefs about it with a *prior density* $p(\theta)$. After we have obtained the data y, our beliefs about θ are contained in the conditional density,

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta}, \tag{1}$$

called *posterior density*.

Prior and posterior distributions (2)

When θ can get values continuosly on some interval, we can express our beliefs about it with a *prior density* $p(\theta)$. After we have obtained the data y, our beliefs about θ are contained in the conditional density,

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta}, \tag{1}$$

called *posterior density*.

Since θ is integrated out in the denominator, it can be considered as a constant with respect to θ. Therefore, the Bayes' formula in (1) is often written as

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{2}$$

which denotes that $p(\theta|y)$ is proportional to $p(\theta)p(y|\theta)$.

Example 1 (Introducing a New Drug in the Market)

A drug company would like to introduce a drug to reduce acid indigestion. It is desirable to estimate θ, the proportion of the market share that this drug will capture. The company interviews n people and Y of them say that they will buy the drug. In the non-Bayesian analysis $\theta \in [0, 1]$ and $Y \sim \text{Bin}(n, \theta)$.

Example 1 (Introducing a New Drug in the Market)

A drug company would like to introduce a drug to reduce acid indigestion. It is desirable to estimate θ, the proportion of the market share that this drug will capture. The company interviews n people and Y of them say that they will buy the drug. In the non-Bayesian analysis $\theta \in [0, 1]$ and $Y \sim \text{Bin}(n, \theta)$.

We know that $\hat{\theta} = Y/n$ is a very good estimator of θ. It is unbiased, consistent and minimum variance unbiased. Moreover, it is also the maximum likelihood estimator (MLE), and thus asymptotically normal.

Example 1 (Introducing a New Drug in the Market)

A drug company would like to introduce a drug to reduce acid indigestion. It is desirable to estimate θ, the proportion of the market share that this drug will capture. The company interviews n people and Y of them say that they will buy the drug. In the non-Bayesian analysis $\theta \in [0, 1]$ and $Y \sim \text{Bin}(n, \theta)$.

We know that $\hat{\theta} = Y/n$ is a very good estimator of θ. It is unbiased, consistent and minimum variance unbiased. Moreover, it is also the maximum likelihood estimator (MLE), and thus asymptotically normal.

A Bayesian may look at the past performance of new drugs of this type. If in the past new drugs tend to capture a proportion between say .05 and .15 of the market, and if all values in between are assumed equally likely, then $\theta \sim \text{Unif}(.05, .15)$.

(Example from Rohatgi, 2003).

Example 1 (continued)

Thus, the prior distribution is given by

$$p(\theta) = \begin{cases} 1/(0.15 - 0.05) = 10, & 0.05 \leq \theta \leq 0.15 \\ 0, & \text{otherwise.} \end{cases}$$

Example 1 (continued)

Thus, the prior distribution is given by

$$p(\theta) = \begin{cases} 1/(0.15 - 0.05) = 10, & 0.05 \leq \theta \leq 0.15 \\ 0, & \text{otherwise.} \end{cases}$$

and the likelihood function by

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Example 1 (continued)

Thus, the prior distribution is given by

$$p(\theta) = \begin{cases} 1/(0.15 - 0.05) = 10, & 0.05 \leq \theta \leq 0.15 \\ 0, & \text{otherwise.} \end{cases}$$

and the likelihood function by

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

The posterior distribution is

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta} = \begin{cases} \frac{\theta^y(1-\theta)^{n-y}}{\int_{0.05}^{0.15} \theta^y(1-\theta)^{n-y}d\theta} & 0.05 \leq \theta \leq 0.15 \\ 0, & \text{otherwise.} \end{cases}$$

Example 1 (continued)

Suppose that the sample size is $n = 100$ and $y = 20$ say that they will use the drug. Then the following BUGS code can be used to simulate the posterior distribution.

```
model{
  theta ~ dunif(0.05,0.15)
  y ~ dbin(theta,n)
}
```

Suppose that this is the contents of file Acid.txt at the home directory. Then JAGS can be called from R as follows:

```
acid <- list(n=100,y=20)
acid.jag <- jags.model("Acid1.txt",acid)
acid.coda <- coda.samples(acid.jag,"theta",10000)
hist(acid.coda[[1]][,"theta"],main="",xlab=expression(theta))
```

Example 1 (continued)

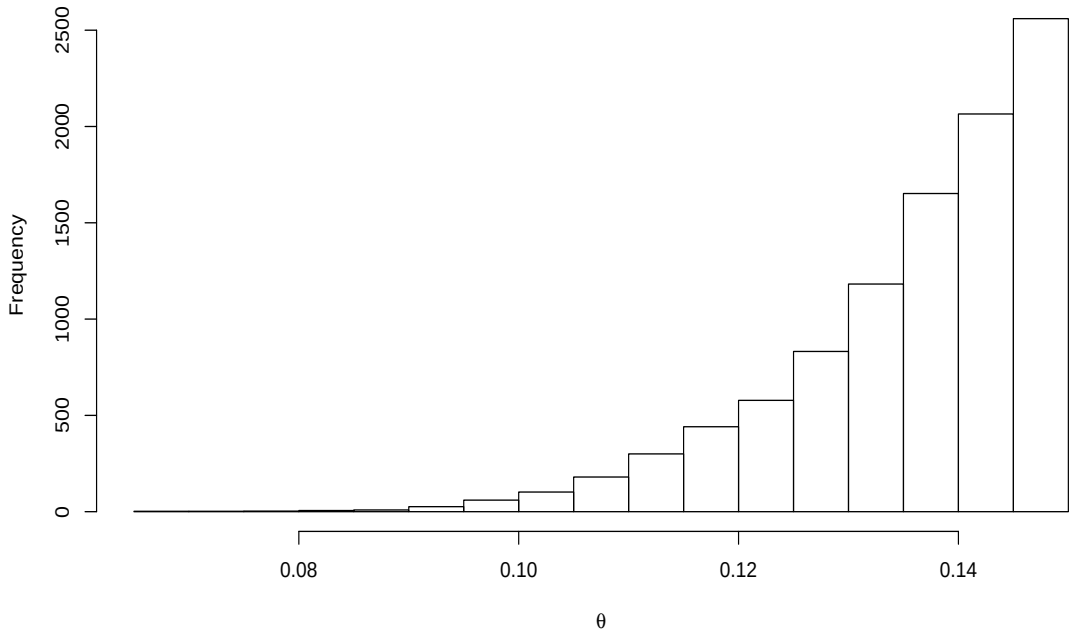


Figure 1: Market share of a new drug: Simulations from the posterior distribution of θ.

Example 2 (Diseased White Pine Trees.)

White pine is one of the best known species of pines in the northeastern United States and Canada. White pine is susceptible to blister rust, which develops cankers on the bark. These cankers swell, resulting in death of twigs and small trees. A forester wishes to estimate the average number of diseased pine trees per acre in a forest.

Example 2 (Diseased White Pine Trees.)

White pine is one of the best known species of pines in the northeastern United States and Canada. White pine is susceptible to blister rust, which develops cankers on the bark. These cankers swell, resulting in death of twigs and small trees. A forester wishes to estimate the average number of diseased pine trees per acre in a forest.

The number of diseased trees per acre can be modeled by a Poisson distribution with mean θ. Since θ changes from area to area, the forester believes that $\theta \sim \text{Exp}(\lambda)$. Thus,

$$p(\theta) = (1/\lambda)e^{-\theta/\lambda}, \quad \text{if } \theta > 0, \text{and 0 elsewhere}$$

Example 2 (Diseased White Pine Trees.)

White pine is one of the best known species of pines in the northeastern United States and Canada. White pine is susceptible to blister rust, which develops cankers on the bark. These cankers swell, resulting in death of twigs and small trees. A forester wishes to estimate the average number of diseased pine trees per acre in a forest.

The number of diseased trees per acre can be modeled by a Poisson distribution with mean θ. Since θ changes from area to area, the forester believes that $\theta \sim \text{Exp}(\lambda)$. Thus,

$$p(\theta) = (1/\lambda)e^{-\theta/\lambda}, \quad \text{if } \theta > 0, \text{and 0 elsewhere}$$

The forester takes a random sample of size n from n different one-acre plots.

(Example from Rohatgi, 2003).

Example 2 (continued)

The likelihood function is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod y_i!} e^{-n\theta}.$$

Example 2 (continued)

The likelihood function is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod y_i!} e^{-n\theta}.$$

Consequently, the posterior distribution is

$$p(\theta|y) = \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}}{\int_0^\infty \theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}}.$$

We see that this is a Gamma-distribution with parameters $\alpha = \sum_{i=1}^{n} y_i + 1$ and $\beta = n + 1/\lambda$.

Example 2 (continued)

The likelihood function is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \frac{\theta^{\sum_{i=1}^{n} y_i}}{\prod y_i!} e^{-n\theta}.$$

Consequently, the posterior distribution is

$$p(\theta|y) = \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}}{\int_0^\infty \theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}}.$$

We see that this is a Gamma-distribution with parameters $\alpha = \sum_{i=1}^{n} y_i + 1$ and $\beta = n + 1/\lambda$. Thus,

$$p(\theta|y) = \frac{(n+1/\lambda)^{\sum_{i=1}^{n} y_i + 1}}{\Gamma(\sum_{i=1}^{n} y_i + 1)} \theta^{\sum_{i=1}^{n} y_i} e^{-\theta(n+1/\lambda)}.$$

Statistical decision theory

The outcome of a Bayesian analysis is the posterior distribution, which combines the prior information and the information from data. However, sometimes we may want to summarize the posterior information with a scalar, for example the mean, median or mode of the posterior distribution. In the following, we show how the use of scalar estimator can be justified using statistical decision theory.

Let $L(\theta, \hat{\theta})$ denote the loss function which gives the cost of using $\hat{\theta} = \hat{\theta}(y)$ as an estimate for θ. We define that $\hat{\theta}$ is a *Bayes estimate* of θ if it minimizes the *posterior expected loss*

$$\mathsf{E}[L(\theta, \hat{\theta})|y] = \int L(\theta, \hat{\theta}) p(\theta|y) d\theta.$$

Statistical decision theory (continued)

On the other hand, the expectation of the loss function over the sampling distribution of y is called *risk function*:

$$R_{\hat{\theta}}(\theta) = \mathsf{E}[L(\theta, \hat{\theta})|\theta] = \int L(\theta, \hat{\theta}) p(y|\theta) dy.$$

Statistical decision theory (continued)

On the other hand, the expectation of the loss function over the sampling distribution of y is called *risk function*:

$$R_{\hat{\theta}}(\theta) = \mathsf{E}[L(\theta, \hat{\theta})|\theta] = \int L(\theta, \hat{\theta}) p(y|\theta) dy.$$

Further, the expectation of the risk function over the prior distribution of θ,

$$\mathsf{E}[R_{\hat{\theta}}(\theta)] = \int R_{\hat{\theta}}(\theta) p(\theta) d\theta,$$

is called *Bayes risk*.

Statistical decision theory (continued)

By changing the order of integration one can see that the Bayes risk

$$\begin{aligned}\int R_{\hat{\theta}}(\theta)p(\theta)d\theta &= \int p(\theta)\int L(\theta,\hat{\theta})p(y|\theta)dyd\theta \\ &= \int p(y)\int L(\theta,\hat{\theta})p(\theta|y)d\theta dy \qquad (3)\end{aligned}$$

is minimized when the inner integral in (3) is minimized for each y, that is, when a Bayes estimator is used.

Statistical decision theory (continued)

By changing the order of integration one can see that the Bayes risk

$$\begin{aligned}\int R_{\hat{\theta}}(\theta)p(\theta)d\theta &= \int p(\theta)\int L(\theta,\hat{\theta})p(y|\theta)dyd\theta \\ &= \int p(y)\int L(\theta,\hat{\theta})p(\theta|y)d\theta dy \qquad (3)\end{aligned}$$

is minimized when the inner integral in (3) is minimized for each y, that is, when a Bayes estimator is used.

In the following, we introduce the Bayes estimators for three simple loss functions.

Bayes estimators: zero-one loss function

Zero-one loss:

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{when } |\hat{\theta} - \theta| < a \\ 1 & \text{when } |\hat{\theta} - \theta| \geq a. \end{cases}$$

Bayes estimators: zero-one loss function

Zero-one loss:

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{when } |\hat{\theta} - \theta| < a \\ 1 & \text{when } |\hat{\theta} - \theta| \geq a. \end{cases}$$

We should minimize

$$\begin{aligned} \int_{-\infty}^{\infty} L(\theta, \hat{\theta}) p(\theta|y) d\theta &= \int_{-\infty}^{\hat{\theta}-a} p(\theta|y) d\theta + \int_{\hat{\theta}+a}^{\infty} p(\theta|y) d\theta \\ &= 1 - \int_{\hat{\theta}-a}^{\hat{\theta}+a} p(\theta|y) d\theta, \end{aligned}$$

Bayes estimators: zero-one loss function

Zero-one loss:

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{when } |\hat{\theta} - \theta| < a \\ 1 & \text{when } |\hat{\theta} - \theta| \geq a. \end{cases}$$

We should minimize

$$\begin{aligned} \int_{-\infty}^{\infty} L(\theta, \hat{\theta}) p(\theta|y) d\theta &= \int_{-\infty}^{\hat{\theta}-a} p(\theta|y) d\theta + \int_{\hat{\theta}+a}^{\infty} p(\theta|y) d\theta \\ &= 1 - \int_{\hat{\theta}-a}^{\hat{\theta}+a} p(\theta|y) d\theta, \end{aligned}$$

or maximize

$$\int_{\hat{\theta}-a}^{\hat{\theta}+a} p(\theta|y) d\theta.$$

Bayes estimators: absolute error loss and quadratic loss function

If $p(\theta|y)$ is unimodal, maximization is achieved by choosing $\hat{\theta}$ to be the midpoint of the interval of length $2a$ for which $p(\theta|y)$ has the same value at both ends. If we let $a \to 0$, then $\hat{\theta}$ tends to the mode of the posterior distribution. This equals the MLE if $p(\theta)$ is 'flat'.

Bayes estimators: absolute error loss and quadratic loss function

If $p(\theta|y)$ is unimodal, maximization is achieved by choosing $\hat{\theta}$ to be the midpoint of the interval of length $2a$ for which $p(\theta|y)$ has the same value at both ends. If we let $a \to 0$, then $\hat{\theta}$ tends to the mode of the posterior distribution. This equals the MLE if $p(\theta)$ is 'flat'.

Absolute error loss: $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$. In general, if X is a random variable, then the expectation $\mathsf{E}(|X - d|)$ is minimized by choosing d to be the median of the distribution of X. Thus, the Bayes estimate of θ is the posterior median.

Bayes estimators: absolute error loss and quadratic loss function

If $p(\theta|y)$ is unimodal, maximization is achieved by choosing $\hat{\theta}$ to be the midpoint of the interval of length $2a$ for which $p(\theta|y)$ has the same value at both ends. If we let $a \to 0$, then $\hat{\theta}$ tends to the mode of the posterior distribution. This equals the MLE if $p(\theta)$ is 'flat'.

Absolute error loss: $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$. In general, if X is a random variable, then the expectation $\mathsf{E}(|X - d|)$ is minimized by choosing d to be the median of the distribution of X. Thus, the Bayes estimate of θ is the posterior median.

Quadratic loss function: $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$. In general, if X is a random variable, then the expectation $\mathsf{E}[(X - d)^2]$ is minimized by choosing d to be the mean of the distribution of X. Thus, the Bayes estimate of θ is the posterior mean.

Bayes estimators: Example 1 (cont)

We continue our example of the market share of a new drug. Using R, we can compute the posterior mean and median estimates, and various posterior intervals:

```
summary(acid.coda)

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
     0.1357622      0.0121584      0.0001216      0.0002253

2. Quantiles for each variable:

  2.5%    25%    50%    75%  97.5%
0.1050 0.1294 0.1390 0.1453 0.1496
```

Bayes estimators: Example 1 (cont)

From Figure 1 we see that the posterior mode is 0.15.

Bayes estimators: Example 1 (cont)

From Figure 1 we see that the posterior mode is 0.15.

If we use Beta(α, β), whose density is

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{when } 0 < \theta < 1,$$

Bayes estimators: Example 1 (cont)

From Figure 1 we see that the posterior mode is 0.15.

If we use Beta(α, β), whose density is

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{when } 0 < \theta < 1,$$

as a prior, then the posterior is

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}.$$

We see immediately that the posterior distribution is Beta$(\alpha + y, \beta + n - y)$.

Bayes estimators: Example 1 (cont)

From Figure 1 we see that the posterior mode is 0.15.

If we use Beta(α, β), whose density is

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{when } 0 < \theta < 1,$$

as a prior, then the posterior is

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}.$$

We see immediately that the posterior distribution is Beta$(\alpha + y, \beta + n - y)$.

The posterior mean (Bayes estimator with quadratic loss) is $(\alpha + y)/(\alpha + \beta + n)$. The mode (Bayes estimator with zero-one loss when $a \to 0$) is $(\alpha + y - 1)/(\alpha + \beta + n - 2)$, provided that the distribution is unimodal.

Bayes estimators: Example 2 (cont)

We now continue our example of estimating the proportion of diseased trees. We derived that the posterior distribution is $\text{Gamma}(\sum_{i=1}^{n} y_i + 1, n + 1/\lambda)$. Thus, the Bayes estimator with a quadratic loss function is the mean of this distribution, $(\sum_{i=1}^{n} y_i + 1)/(n + 1/\lambda)$. However, the mean and mode of a gamma distribution do not exist in closed form.

Bayes estimators: Example 2 (cont)

We now continue our example of estimating the proportion of diseased trees. We derived that the posterior distribution is $\text{Gamma}(\sum_{i=1}^{n} y_i + 1, n + 1/\lambda)$. Thus, the Bayes estimator with a quadratic loss function is the mean of this distribution, $(\sum_{i=1}^{n} y_i + 1)/(n + 1/\lambda)$. However, the mean and mode of a gamma distribution do not exist in closed form.

Note that the classical estimate for θ is the sample mean $\bar{y}$.

Conjugate prior distribution

Computations can often be facilitated using *conjugate prior distributions*. We say that a prior is conjugate for the likelihood if the prior and posterior distributions belong to the same family. There are conjugate distributions for the exponential family of sampling distributions.

Conjugate prior distribution

Computations can often be facilitated using *conjugate prior distributions*. We say that a prior is conjugate for the likelihood if the prior and posterior distributions belong to the same family. There are conjugate distributions for the exponential family of sampling distributions.

Conjugate priors can be formed with the following simple steps:

1. Write the likelihood function.
2. Remove the factors that do not depend on θ.
3. Replace the expressions which depend on data with parameters. Also the sample size n should be replaced.
4. Now you have the kernel of the conjugate prior. You can complement it with the normalizing constant.
5. In order to obtain the standard parametrization it may be necessary to reparametrize.

Example: Poisson likelihood

Let $y = (y_1, ...y_n)$ be a sample from $\text{Poi}(\theta)$. Then the likelihood is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \propto \theta^{\sum y_i} e^{-n\theta}.$$

Example: Poisson likelihood

Let $y = (y_1, ...y_n)$ be a sample from Poi(θ). Then the likelihood is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \propto \theta^{\sum y_i} e^{-n\theta}.$$

By replacing $\sum y_i$ and n, which depend on the data, with the parameters α_1 and α_2, we obtain the conjugate prior

$$p(\theta) \propto \theta^{\alpha_1} e^{-\alpha_2 \theta},$$

which is Gamma($\alpha_1 + 1, \alpha_2$) distribution. If we reparametrize this distribution so that $\alpha = \alpha_1 + 1$ and $\beta = \alpha_2$ we obtain the prior Gamma(α, β).

Example: Uniform likelihood

Assume that $y = (y_1, ..., y_n)$ is a random sample from Unif$(0, \theta)$. The the density of a single observation y_i is

$$p(y_i|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq y_i \leq \theta, \\ 0, & \text{otherwise}, \end{cases}$$

Example: Uniform likelihood

Assume that $y = (y_1, ..., y_n)$ is a random sample from Unif$(0, \theta)$. The the density of a single observation y_i is

$$p(y_i|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq y_i \leq \theta, \\ 0, & \text{otherwise}, \end{cases}$$

and the likelihood of θ is

$$\begin{aligned} p(y|\theta) &= \begin{cases} \frac{1}{\theta^n}, & 0 \leq y_{(1)} \leq \ldots \leq y_{(n)} \leq \theta, \\ 0, & \text{otherwise}, \end{cases} \\ &= \frac{1}{\theta^n} I_{\{y_{(n)} \leq \theta\}}(y)\ I_{\{y_{(1)} \geq 0\}}(y), \end{aligned}$$

where $I_A(y)$ denotes an indicator function obtaining value 1 when $y \in A$ and 0 otherwise.

Example: Uniform likelihood (cont)

Now, by removing the factor $I_{\{y_{(1)} \geq 0\}}(y)$, which does not depend on θ, and replacing n and $y_{(n)}$ with parameters we obtain

$$\begin{aligned} p(\theta) &\propto \frac{1}{\theta^{\alpha}} I_{\{\theta \geq \beta\}}(\theta) \\ &= \begin{cases} \frac{1}{\theta^{\alpha}}, & \text{when } \theta \geq \beta, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

This is the kernel of the Pareto distribution.

Example: Uniform likelihood (cont)

Now, by removing the factor $I_{\{y_{(1)} \geq 0\}}(y)$, which does not depend on θ, and replacing n and $y_{(n)}$ with parameters we obtain

$$\begin{aligned} p(\theta) &\propto \frac{1}{\theta^{\alpha}} I_{\{\theta \geq \beta\}}(\theta) \\ &= \begin{cases} \frac{1}{\theta^{\alpha}}, & \text{when } \theta \geq \beta, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

This is the kernel of the Pareto distribution.The posterior distribution

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \begin{cases} \frac{1}{\theta^{n+\alpha}}, & \text{when } \theta \geq \max(\beta, y_{(n)}) \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

is also a Pareto distribution.

Noninformative prior distribution

When there is no prior information available on the estimated parameters, noninformative priors can be used. They can also be used to find out how an informative prior affects the outcome of the inference.

Noninformative prior distribution

When there is no prior information available on the estimated parameters, noninformative priors can be used. They can also be used to find out how an informative prior affects the outcome of the inference.

The uniform distribution $p(\theta) \propto 1$ is often used as a noninformative prior. However, this is not fully unproblematic. If the uniform distribution is restricted to an interval, it is not, in fact, noninformative. For example, the prior $\text{Unif}(0, 1)$, contains the information that θ is in the interval $[0.2, 0.4]$ with probability 0.2. This information content becomes obvious when a parametric transformation is made. The distribution of the transformed parameter is no more uniform.

Noninformative prior distribution (cont)

Another problem arises if the parameter can obtain values in an infinite interval. In such a case there is no proper uniform distribution. However, one can use an improper uniform prior distribution. Then the posterior is proportional to the likelihood.

Noninformative prior distribution (cont)

Another problem arises if the parameter can obtain values in an infinite interval. In such a case there is no proper uniform distribution. However, one can use an improper uniform prior distribution. Then the posterior is proportional to the likelihood.

Some parameters, for example scale parameteres and variances, can obtain only positive values. Such variables are often given the improper prior $p(\theta) \propto 1/\theta$, which implies that $\log(\theta)$ has a uniform prior.

Noninformative prior distribution (cont)

Another problem arises if the parameter can obtain values in an infinite interval. In such a case there is no proper uniform distribution. However, one can use an improper uniform prior distribution. Then the posterior is proportional to the likelihood.

Some parameters, for example scale parameteres and variances, can obtain only positive values. Such variables are often given the improper prior $p(\theta) \propto 1/\theta$, which implies that $\log(\theta)$ has a uniform prior.

Jeffreys has suggested giving a uniform prior for such a transformation of θ that its Fisher information is a constant. Jeffreys' prior is defined as $p(\theta) \propto \mathcal{I}(\theta)^{\frac{1}{2}}$, where $\mathcal{I}(\theta)$ is the Fisher information of θ. That this definition is invariant to parametrization, can be seen as follows:

Noninformative prior distribution (cont)

Let $\phi = h(\theta)$ be a regular, monotonic transformation of θ, and its inverse transformation $\theta = h^{-1}(\phi)$. Then the Fisher information of ϕ is

$$\begin{aligned}
\mathcal{I}(\phi) =& \mathsf{E}\left[\left(\frac{d\log p(y|\phi)}{d\phi}\right)^2 \middle| \phi\right] \\
=& \mathsf{E}\left[\left(\frac{d\log p(y|\theta = h^{-1}(\phi))}{d\theta}\right)^2 \middle| \phi\right] \left|\frac{d\theta}{d\phi}\right|^2 \\
=& \mathcal{I}(\theta)\left|\frac{d\theta}{d\phi}\right|^2 .
\end{aligned}$$

Noninformative prior distribution (cont)

Let $\phi = h(\theta)$ be a regular, monotonic transformation of θ, and its inverse transformation $\theta = h^{-1}(\phi)$. Then the Fisher information of ϕ is

$$\begin{aligned}
\mathcal{I}(\phi) =& \mathsf{E}\left[\left(\frac{d\log p(y|\phi)}{d\phi}\right)^2 \middle| \phi\right] \\
=& \mathsf{E}\left[\left(\frac{d\log p(y|\theta = h^{-1}(\phi))}{d\theta}\right)^2 \middle| \phi\right] \left|\frac{d\theta}{d\phi}\right|^2 \\
=& \mathcal{I}(\theta)\left|\frac{d\theta}{d\phi}\right|^2 .
\end{aligned}$$

Thus, $\mathcal{I}(\phi)^{\frac{1}{2}} = \mathcal{I}(\Theta)^{\frac{1}{2}}\left|\frac{d\theta}{d\phi}\right|$.

Noninformative prior distribution (cont)

Let $\phi = h(\theta)$ be a regular, monotonic transformation of θ, and its inverse transformation $\theta = h^{-1}(\phi)$. Then the Fisher information of ϕ is

$$\begin{aligned}
\mathcal{I}(\phi) =& \mathsf{E}\left[\left(\frac{d\log p(y|\phi)}{d\phi}\right)^2 \middle| \phi\right] \\
=& \mathsf{E}\left[\left(\frac{d\log p(y|\theta = h^{-1}(\phi))}{d\theta}\right)^2 \middle| \phi\right] \left|\frac{d\theta}{d\phi}\right|^2 \\
=& \mathcal{I}(\theta)\left|\frac{d\theta}{d\phi}\right|^2 .
\end{aligned}$$

Thus, $\mathcal{I}(\phi)^{\frac{1}{2}} = \mathcal{I}(\Theta)^{\frac{1}{2}}\left|\frac{d\theta}{d\phi}\right|$.

On the other hand, $p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = \mathcal{I}(\Theta)^{\frac{1}{2}}\left|\frac{d\theta}{d\phi}\right|$, as required.

Jeffreys' prior: Examples

Binomial distribution

The Fisher information of the binomial distribution parameter θ is $\mathcal{I}(\theta) = n/[(\theta(1-\theta)]$. Thus, the Jeffreys prior is $p(\theta) \propto [\theta(1-\theta)]^{-1/2}$, which is the Beta(1/2,1/2) distribution.

Jeffreys' prior: Examples

Binomial distribution

The Fisher information of the binomial distribution parameter θ is $\mathcal{I}(\theta) = n/[(\theta(1-\theta)]$. Thus, the Jeffreys prior is $p(\theta) \propto [\theta(1-\theta)]^{-1/2}$, which is the Beta(1/2,1/2) distribution.

The mean of the normal distribution

The Fisher information for the mean θ of the normal distribution is $\mathcal{I}(\theta) = n/\sigma^2$. This is independent of θ, so that Jeffreys' prior is constant, $p(\theta) \propto 1$.

Jeffreys' prior: Examples

Binomial distribution

The Fisher information of the binomial distribution parameter θ is $\mathcal{I}(\theta) = n/[(\theta(1-\theta)]$. Thus, the Jeffreys prior is $p(\theta) \propto [\theta(1-\theta)]^{-1/2}$, which is the Beta(1/2,1/2) distribution.

The mean of the normal distribution

The Fisher information for the mean θ of the normal distribution is $\mathcal{I}(\theta) = n/\sigma^2$. This is independent of θ, so that Jeffreys' prior is constant, $p(\theta) \propto 1$.

The variance of the normal distribution

Assume that the variance θ of the normal distribution $N(\mu, \theta)$ is unknown. Then its Fisher information is $\mathcal{I}(\theta) = n/(2\theta^2)$, and Jeffreys' prior $p(\theta) \propto 1/\theta$.

Posterior intervals

Whe have seen that it is possible to summarize posterior information using point estimators. However, posterior regions and intervals are usually more useful. We define that a set C is a posterior region of level $1-\alpha$ for θ if the posterior probability of θ belonging to C is $1-\alpha$:

$$\Pr(\theta \in C|y) = \int_C p(\theta|y)d\theta = 1-\alpha.$$

Posterior intervals

Whe have seen that it is possible to summarize posterior information using point estimators. However, posterior regions and intervals are usually more useful. We define that a set C is a posterior region of level $1-\alpha$ for θ if the posterior probability of θ belonging to C is $1-\alpha$:

$$\Pr(\theta \in C|y) = \int_C p(\theta|y)d\theta = 1-\alpha.$$

In the case of scalar parameters one can use posterior intervals (credible intervals). An *equi-tailed posterior inteval* is defined using quantiles of the posterior. Thus, (θ_L, θ_U) is an $100(1-\alpha)\%$ interval if $\Pr(\theta < \theta_L|y) = \Pr(\theta > \theta_U|y) = \alpha/2$.

Posterior intervals

Whe have seen that it is possible to summarize posterior information using point estimators. However, posterior regions and intervals are usually more useful. We define that a set C is a posterior region of level $1-\alpha$ for θ if the posterior probability of θ belonging to C is $1-\alpha$:

$$\Pr(\theta \in C|y) = \int_C p(\theta|y)d\theta = 1-\alpha.$$

In the case of scalar parameters one can use posterior intervals (credible intervals). An *equi-tailed posterior inteval* is defined using quantiles of the posterior. Thus, (θ_L, θ_U) is an $100(1-\alpha)\%$ interval if $\Pr(\theta < \theta_L|y) = \Pr(\theta > \theta_U|y) = \alpha/2$.

An advantage of this type of interval is that it is invariant with respect to one-to-one parameter transformations. Further, it is easy to compute.

Posterior intervals (cont)

A posterior region is said to be a *highest posterior density region* (HPD region) if the posterior density is larger in all points of the region than in any point outside the region. This type of region has the smallest possible volume. In a scalar case, an HPD interval has the smallest length. On the other hand, the bounds of the interval are not invariant with respect to parameter transformations, and it is not always easy to determine them.

Posterior intervals (cont)

A posterior region is said to be a *highest posterior density region* (HPD region) if the posterior density is larger in all points of the region than in any point outside the region. This type of region has the smallest possible volume. In a scalar case, an HPD interval has the smallest length. On the other hand, the bounds of the interval are not invariant with respect to parameter transformations, and it is not always easy to determine them.

Example. Cardiac surgery data. Table 1 shows mortality rates for cardiac surgery on babies at 12 hospitals. If one wishes to estimate the mortality rate in hospital A, denoted as θ_A, the simpliest approach is to assume that the number of deaths y is binomially distributed with parameters n and θ_A where n is the number of operations in A. Then the MLE is $\hat{\theta}_A = 0$, which sounds too optimistic.

Posterior intervals (cont)

If we give a uniform prior for θ_A, then the posterior distribution is Beta(1,48), with posterior mean 1/49. The 95% HPD interval is (0,6.05)% and equi-tailed interval (0.05,7.30)%. Figure 2 shows the posterior density. Another approach would use the total numbers of deaths and operations in all hospitals.

Table 1: Mortality rates y/n from cardiac surgery in 12 hospitals (Spiegelhalter et. al, *BUGS 0.5 Examples Volume 1*, Cambridge: MRC Biostatistics Unit, 1996). The numbers of deaths y out of n operations.

A	0/47	B	18/148	C	8/119	D	46/810
E	8/211	F	13/196	G	9/148	H	31/215
I	14/207	J	8/97	K	29/256	L	24/360

Posterior intervals (cont)

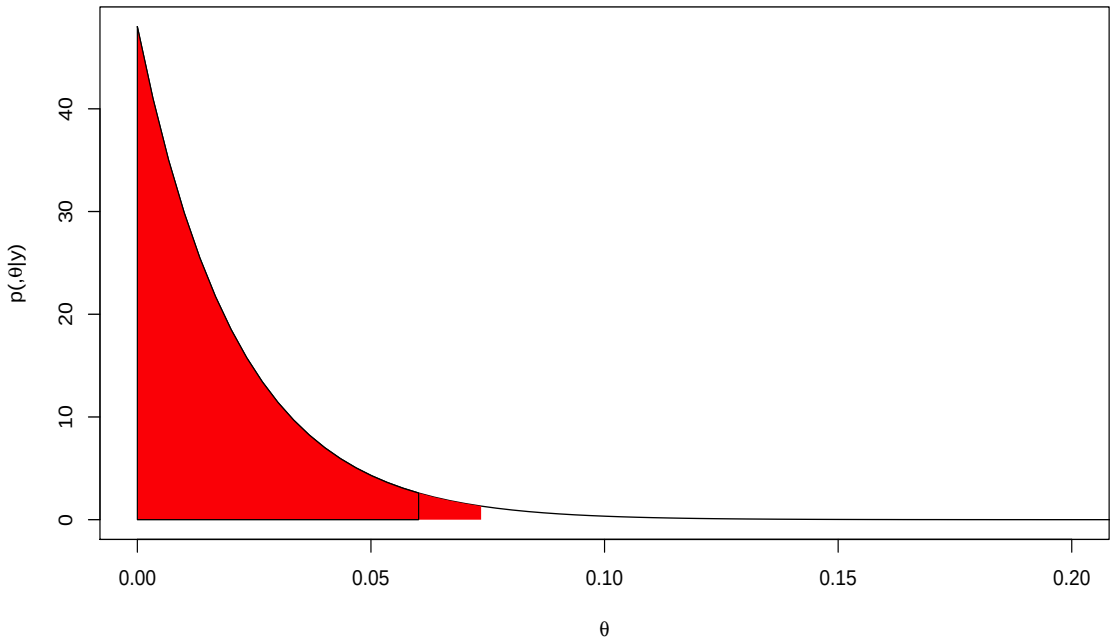


Figure 2: Posterior density of θ_A when the prior is uniform. The 95% HPD interval is indicated with vertical lines and 95% equi-tailed interval with red colour.

Posterior intervals (cont)

The following BUGS and R codes can be used to compute the equi-tailed and HPD intervals:

```
model{
  theta ~ dbeta(1,1)
  y ~ dbin(theta,n)
}

hospital <- list(n=47,y=0)
hospital.jag <- jags.model("Hospital.txt",hospital)
hospital.coda <- coda.samples(hospital.jag,"theta",10000)
summary(hospital.coda)
HPDinterval(hospital.coda)

#Compare with exact upper limit of HPD interval:
qbeta(0.95,1,48)
[1] 0.06050341
```

Posterior predictive distribution

If we wish to predict a new observation $\tilde{y}$ on the basis of the sample $y = (y_1, ...y_n)$, we may use its *posterior predictive distribution*. This is defined to be the conditional distribution of $\tilde{y}$ given y:

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

where $p(\tilde{y}|y, \theta)$ is the density of the *predictive distribution*.

Posterior predictive distribution

If we wish to predict a new observation $\tilde{y}$ on the basis of the sample $y = (y_1, ...y_n)$, we may use its *posterior predictive distribution*. This is defined to be the conditional distribution of $\tilde{y}$ given y:

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

where $p(\tilde{y}|y, \theta)$ is the density of the *predictive distribution*.

It is easy to simulate the posterior predictive distribution. First, draw simulations $\theta_1, ..., \theta_L$ from the posterior $p(\theta|y)$, then, for each i, draw $\tilde{y}_i$ from the predictive distribution $p(\tilde{y}|y, \theta_i)$.

Posterior predictive distribution: Example

Assume that we have a coin with unknown probability θ of a head. If there occurs y heads among the first n tosses what is the probability of a head on the next throw?

Posterior predictive distribution: Example

Assume that we have a coin with unknown probability θ of a head. If there occurs y heads among the first n tosses what is the probability of a head on the next throw?

Let $\tilde{y} = 1$ ($\tilde{y} = 0$) indicate the event that the next throw is a head (tail). If the prior of θ is Beta(α, β), then

$$\begin{aligned} p(\tilde{y}|y) &= \int_0^1 p(\tilde{y}|y,\theta)p(\theta|y)d\theta \\ &= \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} \frac{\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}}{B(\alpha+y,\beta+n-y)} d\theta \\ &= \frac{B(\alpha+y+\tilde{y},\beta+n-y-\tilde{y}+1)}{B(\alpha+y,\beta+n-y)} \\ &= \frac{(\alpha+y)^{\tilde{y}}(\beta+n-y)^{1-\tilde{y}}}{\alpha+\beta+n}. \end{aligned}$$

Posterior predictive distribution: Example (cont)

Thus, $\Pr(\tilde{y} = 1|y) = (\alpha + y)/(\alpha + \beta + n)$. This tends to the sample proportion y/n as $n \to \infty$, so that the role of the prior information vanishes. If $n = 10$ and $y = 4$ and prior parameters $\alpha = \beta = 0.5$ (Jeffreys' prior), the posterior predictive distribution can be simulated with BUGS as follows:

```
model{
  theta ~ dbeta(alpha,beta)
  y ~ dbin(theta,n)
  ynew ~ dbern(theta)
}
```

```
coin <- list(n=10,y=4,alpha=0.5,beta=0.5)
coin.jag <- jags.model("Coin.txt",coin)
coin.coda <- coda.samples(coin.jag,c("theta","ynew"),10000)
summary(coin.coda)
```

Single-parameter models

Normal distribution with known variance

Next we will consider some simple single-parameter models. Let us first assume that $y = (y_1, ...y_n)$ is a sample from a normal distribution unknown mean θ and known variance σ^2. The likelihood is then

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2} \\ &\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2} \\ &\propto e^{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2}. \end{aligned}$$

Normal distribution with known variance

Next we will consider some simple single-parameter models. Let us first assume that $y = (y_1, ...y_n)$ is a sample from a normal distribution unknown mean θ and known variance σ^2. The likelihood is then

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2} \\ &\propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2} \\ &\propto e^{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2}. \end{aligned}$$

By replacing σ^2/n with τ_0^2, and $\bar{y}$ with μ_0, we find a conjugate prior

$$p(\theta) \propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2},$$

which is $N(\mu_0, \tau_0^2)$.

Normal distribution with known variance (cont)

With this prior the posterior becomes

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta)p(y|\theta) \\
&\propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2} e^{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left(\theta^2 - 2\frac{\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\theta\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right\},
\end{aligned}
$$

Normal distribution with known variance (cont)

With this prior the posterior becomes

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2} e^{-\frac{n}{2\sigma^2}(\theta-\bar{y})^2} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\left(\theta^2 - 2\frac{\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\theta\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right\}, \end{aligned}$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \quad \text{and} \quad \tau_n^2 = \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)^{-1}.$$

Normal distribution with known variance (cont)

Thus, the posterior distribution is $N(\mu_n, \tau_n^2)$.

Normal distribution with known variance (cont)

Thus, the posterior distribution is $N(\mu_n, \tau_n^2)$.

The inverse of variance is called *precision*. We see that

posterior precision = prior precision + data precision

where the prior precision is $1/\tau_0^2$ and data precision n/σ^2 (the inverse of the variance of the sample mean).

Normal distribution with known variance (cont)

Thus, the posterior distribution is $N(\mu_n, \tau_n^2)$.

The inverse of variance is called *precision*. We see that

posterior precision = prior precision + data precision

where the prior precision is $1/\tau_0^2$ and data precision n/σ^2 (the inverse of the variance of the sample mean).

The posterior mean is a weighted average of the prior mean μ_0 and sample mean $\bar{y}$ where the weights are the corresponding precisions. When $n \to \infty$ (or when $\tau_0^2 \to \infty$), the role of the prior information vanishes. Thus, for large values of n, approximately $\theta|y \sim N(\bar{y}, \sigma^2/n)$.

Normal distribution with known variance (cont)

Next, we determine the posterior predictive distribution of a new observation $\tilde{y}$. The joint posterior distribution of θ and $\tilde{y}$ is

$$\begin{aligned} p(\theta, \tilde{y}|y) &= p(\theta|y)p(\tilde{y}|y,\theta) \\ &\propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2 - \frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right\}. \end{aligned}$$

Normal distribution with known variance (cont)

Next, we determine the posterior predictive distribution of a new observation $\tilde{y}$. The joint posterior distribution of θ and $\tilde{y}$ is

$$\begin{aligned} p(\theta, \tilde{y}|y) &= p(\theta|y)p(\tilde{y}|y,\theta) \\ &\propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2 - \frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right\}. \end{aligned}$$

Since the exponent is a quadratic function of θ and $\tilde{y}$, their joint distribution is bivariate normal. Consequently, the marginal distribution $p(\tilde{y}|y)$ is univariate normal, and it is sufficient to determine its mean and variance.

Normal distribution with known variance (cont)

Using the rules of iterated mean and variance, we obtain that

$$\mathsf{E}(\tilde{y}|y) = \mathsf{E}[\mathsf{E}(\tilde{y}|y,\theta)|y] = \mathsf{E}[\theta|y] = \mu_n,$$

and

Normal distribution with known variance (cont)

Using the rules of iterated mean and variance, we obtain that

$$\mathsf{E}(\tilde{y}|y) = \mathsf{E}[\mathsf{E}(\tilde{y}|y,\theta)|y] = \mathsf{E}[\theta|y] = \mu_n,$$

and

$$\begin{aligned}\mathsf{Var}(\tilde{y}|y) &= \mathsf{E}[\mathsf{Var}(\tilde{y}|y,\theta)|y] + \mathsf{Var}[\mathsf{E}(\tilde{y}|y,\theta)|y] \\ &= \mathsf{E}[\sigma^2|y] + \mathsf{Var}[\theta|Y] \\ &= \sigma^2 + \tau_n^2.\end{aligned}$$

Normal distribution with known variance (cont)

Using the rules of iterated mean and variance, we obtain that

$$\mathsf{E}(\tilde{y}|y) = \mathsf{E}[\mathsf{E}(\tilde{y}|y,\theta)|y] = \mathsf{E}[\theta|y] = \mu_n,$$

and

$$\begin{aligned}\mathsf{Var}(\tilde{y}|y) &= \mathsf{E}[\mathsf{Var}(\tilde{y}|y,\theta)|y] + \mathsf{Var}[\mathsf{E}(\tilde{y}|y,\theta)|y] \\ &= \mathsf{E}[\sigma^2|y] + \mathsf{Var}[\theta|Y] \\ &= \sigma^2 + \tau_n^2.\end{aligned}$$

Thus, the posterior predictive distribution is

$$p(\tilde{y}|y) = N(\tilde{y}|\mu_n, \sigma^2 + \tau_n^2).$$

Poisson distribution

The Poisson distribution is often used to model rare incidents, such as traffic accidents or rare diseases. For a vector $y = (y_1, ..., y_n)$ of iid observation, the likelihood is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} \propto \theta^{\sum y_i} e^{-n\theta}.$$

Poisson distribution

The Poisson distribution is often used to model rare incidents, such as traffic accidents or rare diseases. For a vector $y = (y_1, ..., y_n)$ of iid observation, the likelihood is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} \propto \theta^{\sum y_i} e^{-n\theta}.$$

Given that the prior distribution is Gamma(α, β), the posterior

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{\sum y_i} e^{-n\theta} \\ &\propto \theta^{\alpha+\sum y_i-1} e^{-(\beta+n)\theta} \end{aligned}$$

is Gamma$(\alpha + \sum y_i, \beta + n)$.

Poisson distribution (cont)

The negative binomial distribution. When the prior and posterior distributions can be written in closed form, the marginal likelihood $p(y)$ can be computed using the formula

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

Poisson distribution (cont)

The negative binomial distribution. When the prior and posterior distributions can be written in closed form, the marginal likelihood $p(y)$ can be computed using the formula

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

For example, if y is a single observation from Poi(θ), then

$$\begin{aligned} p(y) &= \frac{\frac{\theta^y}{y!}e^{-\theta} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}}{\frac{(\beta+1)^{\alpha+y}}{\Gamma(\alpha+y)}\theta^{\alpha+y-1}e^{-(\beta+1)\theta}} \\ &= \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^y, \end{aligned}$$

which is Neg-Bin(α, β), the negative binomial distribution.

Poisson distribution (cont)

On the other hand,

$$p(y) = \int p(y|\theta)p(\theta) = \int \text{Poi}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)d\theta,$$

implying that the negative binomial distribution is a *compound distribution* where the Poisson distribution is compounded using the Gamma distribution as a weight distribution.

Poisson distribution (cont)

On the other hand,

$$p(y) = \int p(y|\theta)p(\theta) = \int \text{Poi}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)d\theta,$$

implying that the negative binomial distribution is a *compound distribution* where the Poisson distribution is compounded using the Gamma distribution as a weight distribution.

In many applications, the data are distributed as

$$y_i \sim \text{Poi}(x_i\theta),$$

where the x_i are known values of an explanatory variable. In epidemiology, x_i is called *exposure* of the ith unit. With prior distribution $\text{Gamma}(\alpha,\beta)$, the posterior becomes $\text{Gamma}(\alpha + \sum y_i, \beta + \sum x_i)$.

Poisson distribution: Example

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

Table 2: Worldwide airline fatalities 1976-85. Death rate is passenger deaths per 100 million passenger miles. Source: Statistical Abstract of the United States.

Poisson distribution: Example (cont)

In Table 2, the death rate is $d_i = y_i/x_i$ where y_i is the number of passenger deaths and x_i the 'exposure' given in 100 million passenger miles. Thus $x_i = y_i/d_i$. Assuming the model $y_i \sim \text{Poi}(\theta x_i)$, the rate θ can be estimated using BUGS as follows:

```
model{
  theta ~ dgamma(alpha,beta)
  for(i in 1:n){
    y[i] ~ dpois(theta*x[i])
  }
}
```

```
air <- list(n=10,y=deaths,x=deaths/rate,alpha=0.01,beta=0.01)
...
  2.5%    25%    50%    75%  97.5%
0.1182 0.1201 0.1210 0.1220 0.1239
```

Exponential distribution

In a Poisson process having intensity θ, the number of events in a time interval of length τ follows the Poisson distribution with parameter $\tau\theta$. Further, the waiting time between two Poisson events follows the exponential distribution $\text{Exp}(\theta)$, and the waiting time until the nth event is $\text{Gamma}(n, \theta)$.

Exponential distribution

In a Poisson process having intensity θ, the number of events in a time interval of length τ follows the Poisson distribution with parameter $\tau\theta$. Further, the waiting time between two Poisson events follows the exponential distribution $\text{Exp}(\theta)$, and the waiting time until the nth event is $\text{Gamma}(n, \theta)$.

The exponential distribution can also be used to model life times of objects that do not wear out, since in this model the expected remaining life time is independent of the time the object has already survived.

Exponential distribution

In a Poisson process having intensity θ, the number of events in a time interval of length τ follows the Poisson distribution with parameter $\tau\theta$. Further, the waiting time between two Poisson events follows the exponential distribution $\text{Exp}(\theta)$, and the waiting time until the nth event is $\text{Gamma}(n, \theta)$.

The exponential distribution can also be used to model life times of objects that do not wear out, since in this model the expected remaining life time is independent of the time the object has already survived. If $Y \sim \text{Exp}(\theta)$, then

$$\begin{aligned}\Pr(Y \leq y | Y > y_0) &= \frac{\Pr(y_0 < Y \leq y)}{\Pr(Y > y_0} = \frac{\Pr(Y \leq y) - \Pr(Y \leq y_0)}{\Pr(Y > y_0)} \\ &= \frac{(1 - e^{-\theta y}) - (1 - e^{-\theta y_0})}{e^{-\theta y_0}} = 1 - e^{-\theta(y - y_0)},\end{aligned}$$

which is the exponential distribution function starting at y_0.

Exponential distribution (cont)

Bayesian analysis. Let $y = (y_1, ..., y_n)$ be a random sample form $\text{Exp}(\theta)$ and let $\text{Gamma}(\alpha, \beta)$ be the prior. Then the posterior is

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\theta e^{-\theta y_i} \\ &\propto \theta^{\alpha+n-1}e^{-\theta(\beta+\sum y_i)}, \end{aligned}$$

which is $\text{Gamma}(\alpha + n, \beta + \sum y_i)$ distribution.

Exponential distribution (cont)

Bayesian analysis. Let $y = (y_1, ..., y_n)$ be a random sample form $\text{Exp}(\theta)$ and let $\text{Gamma}(\alpha, \beta)$ be the prior. Then the posterior is

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\theta e^{-\theta y_i} \\ &\propto \theta^{\alpha+n-1}e^{-\theta(\beta+\sum y_i)}, \end{aligned}$$

which is $\text{Gamma}(\alpha + n, \beta + \sum y_i)$ distribution.

Censored observations. Assume that the observations $y_1, ...y_m$ are known to be larger than U, while the exact values of $y_{m+1}, ..., y_n$ are known. Then the values $y_1, ...y_m$ are called *right-censored*. On the other hand, if some observations are known to be less or equal to some threshold L, they are called *left-censored*.

Exponential distribution (cont)

In the exponential case of right-censoring, the likelihood is

$$p(y|\theta) = \prod_{i=1}^{m} \Pr(Y_i > U|\theta) \prod_{i=m+1}^{n} p(y_i|\theta)$$

$$= \prod_{i=1}^{m} e^{-\theta U} \prod_{i=m+1}^{n} \theta e^{-\theta y_i} = \theta^{n-m} e^{-\theta(mU+\sum y_i)}.$$

Exponential distribution (cont)

In the exponential case of right-censoring, the likelihood is

$$p(y|\theta) = \prod_{i=1}^{m} \Pr(Y_i > U|\theta) \prod_{i=m+1}^{n} p(y_i|\theta)$$

$$= \prod_{i=1}^{m} e^{-\theta U} \prod_{i=m+1}^{n} \theta e^{-\theta y_i} = \theta^{n-m} e^{-\theta(mU+\sum y_i)}.$$

Thus, with prior Gamma(α, β), the posterior is Gamma$(\alpha + n - m, \beta + mU + \sum y_i)$.

Exponential distribution (cont)

In the exponential case of right-censoring, the likelihood is

$$p(y|\theta) = \prod_{i=1}^{m} \Pr(Y_i > U|\theta) \prod_{i=m+1}^{n} p(y_i|\theta)$$

$$= \prod_{i=1}^{m} e^{-\theta U} \prod_{i=m+1}^{n} \theta e^{-\theta y_i} = \theta^{n-m} e^{-\theta(mU+\sum y_i)}.$$

Thus, with prior Gamma(α, β), the posterior is Gamma$(\alpha + n - m, \beta + mU + \sum y_i)$.

In the case of left-censoring, the likelihood is

$$p(y|\theta) = (1 - e^{-\theta L})^m \ \theta^{n-m} e^{-\theta \sum y_i},$$

so that the posterior distribution is nonstandard.

Exponential distribution: Example

Let us assume that the life time of an electronical component is exponentially distributed. After 2 years it is observed that 3 out of 10 components have broken and the life times of the remaining components are $2.7, 3.7, 4.0, 4.7, 5.9, 6.6, 12.1$.
The JAGS code (in this case different from OpenBUGS or WinBUGS) and the related R code:

```
model{
  theta ~ dgamma(alpha,beta)
  for(i in 1:n){
    x[i] ~ dinterval(y[i],L)
    y[i] ~ dexp(theta)
  }
}
comp <- list(n=10,L=2,y=c(NA,NA,NA,2.7,3.7,4.0,4.7,5.9,6.6,12
  x=c(0,0,0,1,1,1,1,1,1,1),alpha=0.01,beta=0.01)
```

Hypothesis testing

Bayesian hypothesis testing

The frequentist approach to hypothesis testing would compare a null hypothesis H_0 with an alternative H_1 through a test statistic T which typically obtains a larger value when H_1 is true than when H_0 is true. The null hypothesis is rejected with a level α if the observed value of the test statistic, t_{obs}, is larger than the critical value t_C where $\Pr(T > t_C|H_0) = \alpha$. The so-called p-value, $p = \Pr(T \geq t_{obs}|H_0)$, is a related concept.

Bayesian hypothesis testing

The frequentist approach to hypothesis testing would compare a null hypothesis H_0 with an alternative H_1 through a test statistic T which typically obtains a larger value when H_1 is true than when H_0 is true. The null hypothesis is rejected with a level α if the observed value of the test statistic, t_{obs}, is larger than the critical value t_C where $\Pr(T > t_C|H_0) = \alpha$. The so-called p-value, $p = \Pr(T \geq t_{obs}|H_0)$, is a related concept.

In frequentist statistics, we do not assign probabilities to hypotheses. In particular, the p-value cannot be interpreted as $p(H_0)$. On the contrary, in the Bayesian approach, we may assign the prior probabilities $p(H_0)$ and $p(H_1)$, and, using Bayes' theorem, compute the posterior probabilities

$$p(H_i|y) = \frac{p(H_i)p(y|H_i)}{p(H_0)p(y|H_0) + p(H_1)p(y|H_1)}, \quad i = 0, 1.$$

Bayesian hypothesis testing (cont)

In the frequentist approach it is not absolutely necessary to specify an alternative hypothesis. Further, if an alternative is specified, the p-value is independent of it. In the Bayesian approach, the both hypotheses must be fully specified.

Bayesian hypothesis testing (cont)

In the frequentist approach it is not absolutely necessary to specify an alternative hypothesis. Further, if an alternative is specified, the p-value is independent of it. In the Bayesian approach, the both hypotheses must be fully specified.

One usually computes the posterior odds

$$\frac{p(H_1|y}{p(H_0|y)} = \frac{p(y|H_1)}{p(y|H_0)} \times \frac{p(H_1)}{p(H_0)},$$

which depends on the data y only through the *Bayes factor* $B_{10} = p(y|H_1)/p(y|H_0)$.

Bayesian hypothesis testing (cont)

In the frequentist approach it is not absolutely necessary to specify an alternative hypothesis. Further, if an alternative is specified, the p-value is independent of it. In the Bayesian approach, the both hypotheses must be fully specified.

One usually computes the posterior odds

$$\frac{p(H_1|y}{p(H_0|y)} = \frac{p(y|H_1)}{p(y|H_0)} \times \frac{p(H_1)}{p(H_0)},$$

which depends on the data y only through the *Bayes factor* $B_{10} = p(y|H_1)/p(y|H_0)$.

In the case that a hypothesis is composite (not simple), the unknown parameters should be first integrated out:

$$p(y|H_i) = \int p(y|\theta_i, H_i)p(\theta_i|H_i)d\theta_i, \quad i = 0, 1.$$

Bayesian hypothesis testing (cont)

Table 3: Interpretation of Bayes factor B_{10} in favor of H_1 over H_0. From Robert E. Kass and Adrian E. Raftery (1995). "Bayes Factors". JASA 90 (430): 791.

B_{10}	$2 \log B_{10}$	Evidence against H_0
1-3	0-2	Hardly worth a mention
3-20	2-6	Positive
20-150	6-10	Strong
>150	>10	Very strong

Rough interpretations for B_1, and, equivalently for $2 \log B_{10}$, are provided in Table 3. The quantity $2 \log B_{10}$ corresponds to the likelihood ratio statistics in likelihood inference.

Bayesian hypothesis testing: Example

Table 4: The log Bayes factors $2 \log B_{\tau 0}$ for HUS data.

	1970	1971	1972	1973	1974	1975	1976
y	1	5	3	2	2	1	0
$\alpha = \beta = 1$	4.9	-0.5	0.6	3.9	7.5	13	24
$\alpha = \beta = 0.01$	-1.3	-5.9	-4.5	-1.0	3.0	9.7	20
$\alpha = \beta = 0.0001$	-10	-15	-14	-10	-6.1	0.6	11
	1977	1978	1979	1980	1981	1982	1983
y	0	2	1	1	7	11	4
$\alpha = \beta = 1$	35	41	51	63	55	38	42
$\alpha = \beta = 0.01$	32	39	51	64	57	40	47
$\alpha = \beta = 0.0001$	23	30	42	55	48	31	38
	1984	1985	1986	1987	1988	1989	
y	7	10	16	16	9	15	
$\alpha = \beta = 1$	40	31	11	-2.9	-5.3	0	
$\alpha = \beta = 0.01$	46	38	18	1.8	1.2	0	
$\alpha = \beta = 0.0001$	37	29	8.8	-7.1	-7.7	0	

Bayesian hypothesis testing: Example

Table 4 shows the numbers of cases of haemolytic uraemic syndrome (HUS) treated at a clinic in Birmingham from 1970 to 1989. There seems to be a rise in 1981. We assume that the annual counts $y_1, ..., y_n$ are independent and Poisson-distributed with means $\mathsf{E}(Y_j) = \lambda_1$ for $j = 1, ..., \tau$, and $\mathsf{E}(Y_j) = \lambda_2$ for $j = \tau + 1, ..., n$. The changepoint can take values $1, ..., n - 1$.

Bayesian hypothesis testing: Example

Table 4 shows the numbers of cases of haemolytic uraemic syndrome (HUS) treated at a clinic in Birmingham from 1970 to 1989. There seems to be a rise in 1981. We assume that the annual counts $y_1, ..., y_n$ are independent and Poisson-distributed with means $\mathsf{E}(Y_j) = \lambda_1$ for $j = 1, ..., \tau$, and $\mathsf{E}(Y_j) = \lambda_2$ for $j = \tau + 1, ..., n$. The changepoint can take values $1, ..., n - 1$.

Our baseline model H_0 is that there is no change, $\lambda_1 = \lambda_2 = \lambda$, and the alternative H_τ that there is a change after τ years. Under H_τ we assume that λ_1 and λ_2 have independent priors with parameters α and β. Then $p(y|H_\tau)$ equals

$$\int_0^\infty \prod_{j=1}^{\tau} \frac{\lambda_1^{y_j}}{y_j!} e^{-\lambda_1} \cdot \frac{\beta^\alpha \lambda_1^{\alpha-1} e^{-\beta\lambda_1}}{\Gamma(\alpha)} d\lambda_1 \int_0^\infty \prod_{j=\tau+1}^{n} \frac{\lambda_1^{y_j}}{y_j!} e^{-\lambda_2} \cdot \frac{\beta^\alpha \lambda_2^{\alpha-1} e^{-\beta\lambda_2}}{\Gamma(\alpha)} d\lambda_2,$$

Bayesian hypothesis testing: Example

which can be simplified as

$$\frac{\beta^{2\alpha}}{\Gamma(\alpha)^2 \prod_{j=1}^{n} y_j!} \frac{\Gamma(\alpha + s_\tau)\Gamma(\alpha + s_n - s_\tau)}{(\beta + \tau)^{\alpha + s_\tau}(\beta + n - \tau)^{\alpha + s_n - s_\tau}},$$

where $s_\tau = y_1 + ... + y_\tau$ and $s_n = y_1 + ... + y_n$.

Bayesian hypothesis testing: Example

which can be simplified as

$$\frac{\beta^{2\alpha}}{\Gamma(\alpha)^2 \prod_{j=1}^{n} y_j!} \frac{\Gamma(\alpha + s_\tau)\Gamma(\alpha + s_n - s_\tau)}{(\beta + \tau)^{\alpha+s_\tau}(\beta + n - \tau)^{\alpha+s_n-s_\tau}},$$

where $s_\tau = y_1 + ... + y_\tau$ and $s_n = y_1 + ... + y_n$.

Under H_0 we also assume that $\lambda \sim \text{Gamma}(\alpha, \beta)$. Then the Bayes factor for a changepoint in year τ is

$$B_{\tau 0} = \frac{\Gamma(\alpha + s_\tau)\Gamma(\alpha + s_n - s_\tau)\beta^\alpha(\beta + n)^{\alpha+s_n}}{\Gamma(\alpha)\Gamma(\alpha + s_n)(\beta + \tau)^{\alpha+s_\tau}(\beta + n - \tau)^{\alpha+s_n-s_\tau}}, \quad \tau = 1, ..., n$$

From Table 4 we see that there is a very strong evidence for change in 1976–1985 for all priors.

Computing marginal likelihoods

Bayes factors can be presented in closed form only in simple conjugate situations, but various simulation-based methods have been suggested. One simple example is the *harmonic mean* method, which is based on the result

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p(y|\theta^{(t)})} \xrightarrow{p} \frac{1}{p(y)} \quad \text{as } T \to \infty,$$

where $\theta^{(t)},\ t = 1, ..., T$ are independent simulations from $p(\theta|y)$. The result follows from the law of large numbers.

Computing marginal likelihoods

Bayes factors can be presented in closed form only in simple conjugate situations, but various simulation-based methods have been suggested. One simple example is the *harmonic mean* method, which is based on the result

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{p(y|\theta^{(t)})} \xrightarrow{p} \frac{1}{p(y)} \quad \text{as } T \to \infty,$$

where $\theta^{(t)},\ t = 1, ..., T$ are independent simulations from $p(\theta|y)$. The result follows from the law of large numbers.

This estimator is somewhat unstable, since occasional values of $\theta^{(t)}$ with small likelihood have a large effect on it. Therefore, several modifications of the method have been developed. More advanced methods, such as *path sampling*, are effective, but usually require problem-specific tuning.

Model choice: Example (cont)

Another approach is to consider the model choice as a discrete parameter. This is generally a more reliable method to obtain posterior model probabilities with BUGS (BUGS book, 2013).

Model choice: Example (cont)

Another approach is to consider the model choice as a discrete parameter. This is generally a more reliable method to obtain posterior model probabilities with BUGS (BUGS book, 2013).

In the following, we present a code to estimate the model probabilities in the HUS example. We give an equal prior probability, $1/n$, to each of the models H_τ, $\tau = 1, ..., n$. Here, H_n corresponds to H_0.

Model choice: Example (cont)

Another approach is to consider the model choice as a discrete parameter. This is generally a more reliable method to obtain posterior model probabilities with BUGS (BUGS book, 2013).

In the following, we present a code to estimate the model probabilities in the HUS example. We give an equal prior probability, $1/n$, to each of the models H_τ, $\tau = 1, ..., n$. Here, H_n corresponds to H_0.

Figure 3 shows the posterior model probabilities. The values 11 and 12 are the most probable change points:
$\Pr(\tau = 11|y) \approx 0.97$ and $\Pr(\tau = 12|y) \approx 0.03$.

Model choice: Example (cont)


```
model{
  for(i in 1:n){
    q[i] <- 1/n
  }
  tau ~ dcat(q[])
  for(i in 1:2){
    lambda[i] ~ dgamma(alpha,beta)
  }
  for(i in 1:n){
      mu[i] <- lambda[1]+
                step(i-tau-0.1)*(lambda[2]-lambda[1])
      y[i] ~ dpois(mu[i])
  }
}

HUS <- list(n=20, y= c(1,5,3,2,2,1,0,0,2,1,1,7,11,4,7,10,16,1
            alpha=0.01,beta=0.01)
```

Model choice: Example (cont)

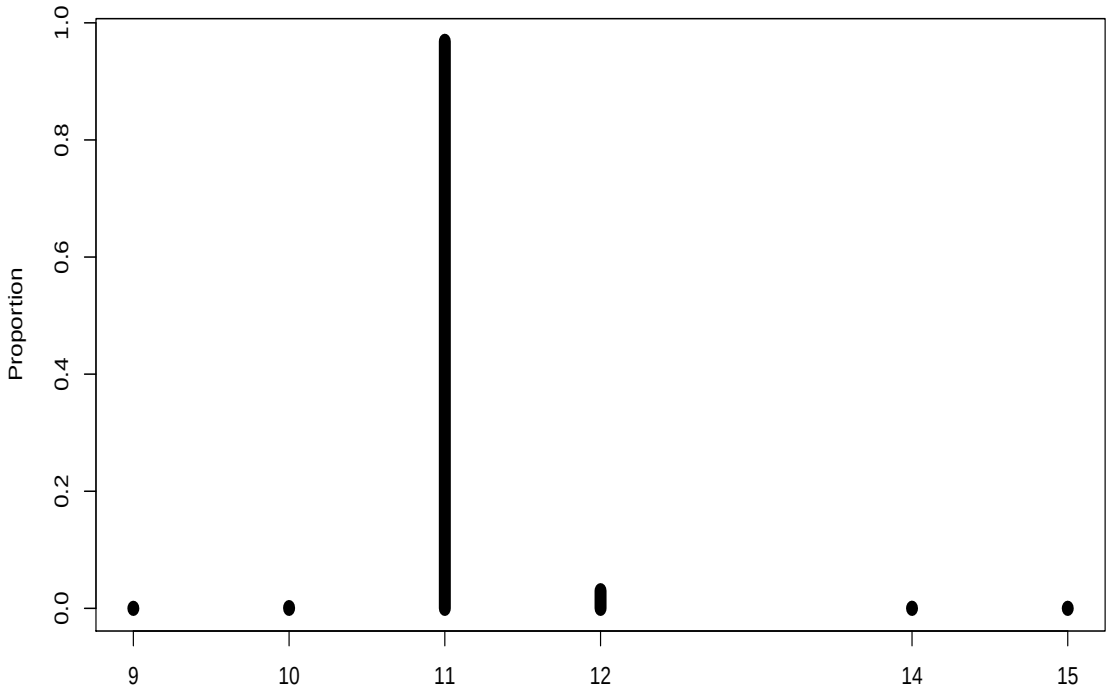


Figure 3: Posterior model probabilities in the HUS example.

Simple multiparameter models

Normal distribution with unknown mean and variance

Next we consider simple models having more than one parameter. Let us assume that $y = (y_1, ..., y_n)$ is a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. If the joint prior is $p(\mu, \sigma^2) \propto 1/\sigma^2$, or equivalently $p(\mu, \log(\sigma^2)) \propto 1$, the posterior is

Normal distribution with unknown mean and variance

Next we consider simple models having more than one parameter. Let us assume that $y = (y_1, ..., y_n)$ is a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. If the joint prior is $p(\mu, \sigma^2) \propto 1/\sigma^2$, or equivalently $p(\mu, \log(\sigma^2)) \propto 1$, the posterior is

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto \frac{1}{\sigma^2} \times \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\ &= \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right]\right) \\ &= \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right), \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance.

Normal distribution (cont)

The marginal posterior of σ^2 is obtained by integrating μ out:

$$p(\sigma^2|y) \propto \int_{-\infty}^{\infty} \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu.$$

Normal distribution (cont)

The marginal posterior of σ^2 is obtained by integrating μ out:

$$p(\sigma^2|y) \propto \int_{-\infty}^{\infty} \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu.$$

The integral of the factor $\exp\left(-\frac{1}{2\sigma^2}n(\bar{y}-\mu)^2\right)$ is a simple normal integral, so

Normal distribution (cont)

The marginal posterior of σ^2 is obtained by integrating μ out:

$$p(\sigma^2|y) \propto \int_{-\infty}^{\infty} \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu.$$

The integral of the factor $\exp\left(-\frac{1}{2\sigma^2}n(\bar{y}-\mu)^2\right)$ is a simple normal integral, so

$$\begin{aligned} p(\sigma^2|y) &\propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{2\pi\sigma^2/n} \\ &\propto \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right). \end{aligned}$$

Normal distribution (cont)

The marginal posterior of σ^2 is obtained by integrating μ out:

$$p(\sigma^2|y) \propto \int_{-\infty}^{\infty} \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu.$$

The integral of the factor $\exp\left(-\frac{1}{2\sigma^2}n(\bar{y}-\mu)^2\right)$ is a simple normal integral, so

$$\begin{aligned} p(\sigma^2|y) &\propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{2\pi\sigma^2/n} \\ &\propto \frac{1}{(\sigma^2)^{(n+1)/2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right). \end{aligned}$$

This is a scaled inverse-χ^2-density:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2).$$

Normal distribution (cont)

Thus, $\{(n-1)s^2/\sigma^2 \mid y\} \sim \chi^2_{n-1}$. This is analogous with the corresponding sampling theory result. However, in sampling theory, s^2 is considered random, while here σ^2 is random.

Normal distribution (cont)

Thus, $\{(n-1)s^2/\sigma^2 \mid y\} \sim \chi^2_{n-1}$. This is analogous with the corresponding sampling theory result. However, in sampling theory, s^2 is considered random, while here σ^2 is random.

By making the substitution

$$z = \frac{A}{\sigma^2}, \quad \text{where } A = (n-1)s^2 + n(\bar{y} - \mu)^2,$$

Normal distribution (cont)

Thus, $\{(n-1)s^2/\sigma^2 \mid y\} \sim \chi^2_{n-1}$. This is analogous with the corresponding sampling theory result. However, in sampling theory, s^2 is considered random, while here σ^2 is random.

By making the substitution

$$z = \frac{A}{\sigma^2}, \quad \text{where } A = (n-1)s^2 + n(\bar{y}-\mu)^2,$$

we obtain the marginal density of μ:

$$\begin{aligned} p(\mu|y) &\propto \int_0^\infty \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\sigma^2 \\ &\propto A^{-n/2} \int_0^\infty z^{n/2-1} \exp(-z) dz \\ &\propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}. \end{aligned}$$

Normal distribution: Speed of light (example)

This is the $t_{n-1}(\bar{y}, s^2/n)$ density. Thus, $\{(\mu - \bar{y})/(s/\sqrt{n}) \mid y\} \sim t_{n-1}$. This is again analogous to the sampling theory result. It can also be shown (exercise) that the density of a new observation $\tilde{y}$ is $t_{n-1}(\bar{y}, s^2(1 + 1/n))$. The posterior can be simulated using $p(\sigma^2|y)$ and $p(\mu|\sigma^2, y) = N(\mu|\bar{y}, \sigma^2/n)$.

Normal distribution: Speed of light (example)

This is the $t_{n-1}(\bar{y}, s^2/n)$ density. Thus, $\{(\mu - \bar{y})/(s/\sqrt{n}) \mid y\} \sim t_{n-1}$. This is again analogous to the sampling theory result. It can also be shown (exercise) that the density of a new observation $\tilde{y}$ is $t_{n-1}(\bar{y}, s^2(1 + 1/n))$. The posterior can be simulated using $p(\sigma^2|y)$ and $p(\mu|\sigma^2, y) = N(\mu|\bar{y}, \sigma^2/n)$.

Example. Estimating the speed of light. Simon Newcomb made an experiment in 1882 to measure the speed of light. He measured the time light travels 7442 meters. Figure 4 shows that there are two outliers, so the normal distribution as such is not a very good model. However, for the sake of illustration, we assume that the observations are independent and from $N(\mu, \sigma^2)$. With the noninformative prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, the 95% posterior interval is $(\bar{y} \pm t_{(n-1);0.025} s/\sqrt{n}) = (23.6, 28.9)$ where $n = 66$, $\bar{y} = 26.2$ and $s = 10.8$. Further, the prediction interval is $(\bar{y} \pm t_{(n-1);0.025} s\sqrt{1 + 1/n}) = (4.6, 47.8)$.

Example: Speed of light (cont)

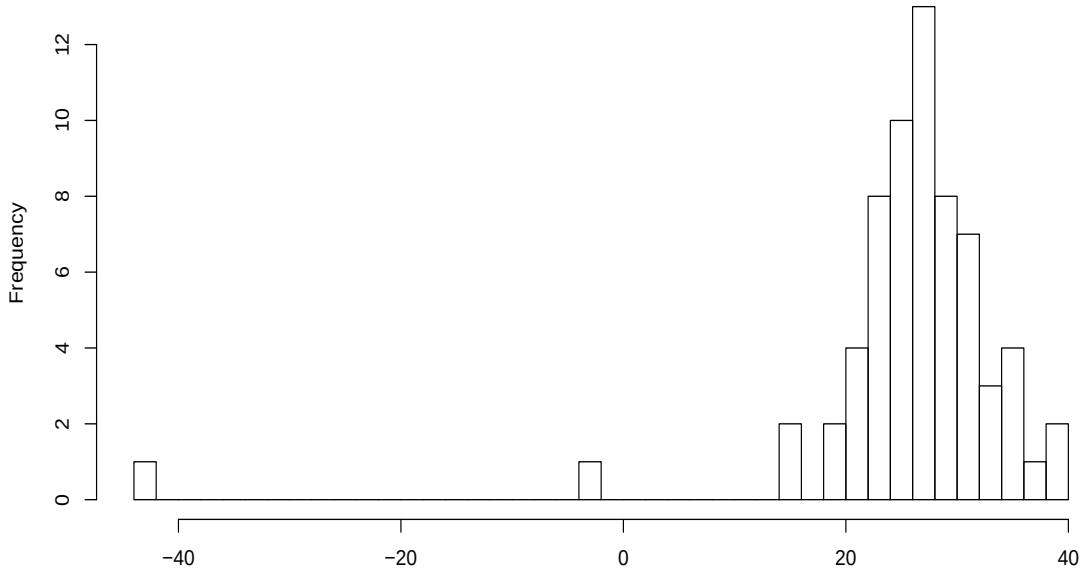


Figure 4: Newcomb's measurements for speed of light.

Multinomial distribution

If $y = (y_1, ...y_k)$ is multinomially distributed with parameters n and $\theta = (\theta_1, ..., \theta_k)$ (denoted as Multin$(n; \theta)$) then the likelihood is

$$p(\theta|y) \propto \theta_1^{y_1} \theta_2^{y_2} ... \theta_k^{y_k}$$

where $\theta_i \geq 0$ for all $i = 1, ..., k$ and $\sum_{i=1}^{k} \theta_i = 1$.

Multinomial distribution

If $y = (y_1, ...y_k)$ is multinomially distributed with parameters n and $\theta = (\theta_1, ..., \theta_k)$ (denoted as Multin$(n; \theta)$) then the likelihood is

$$p(\theta|y) \propto \theta_1^{y_1} \theta_2^{y_2} ... \theta_k^{y_k}$$

where $\theta_i \geq 0$ for all $i = 1, ..., k$ and $\sum_{i=1}^{k} \theta_i = 1$.

It is easy to see that the conjugate prior is the Dirichlet distribution (denoted as Dirichlet$(\alpha_1, ...\alpha_k)$):

$$p(\theta) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} ... \theta_k^{\alpha_k - 1},$$

where $\theta_i \geq 0$ and $\alpha_i > 0$ for all $i = 1, ..., k$, and $\sum_{i=1}^{k} \theta_i = 1$.

Multinomial distribution

If $y = (y_1, ...y_k)$ is multinomially distributed with parameters n and $\theta = (\theta_1, ..., \theta_k)$ (denoted as Multin$(n; \theta)$) then the likelihood is

$$p(\theta|y) \propto \theta_1^{y_1} \theta_2^{y_2} ... \theta_k^{y_k}$$

where $\theta_i \geq 0$ for all $i = 1, ..., k$ and $\sum_{i=1}^{k} \theta_i = 1$.

It is easy to see that the conjugate prior is the Dirichlet distribution (denoted as Dirichlet$(\alpha_1, ...\alpha_k)$):

$$p(\theta) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} ... \theta_k^{\alpha_k - 1},$$

where $\theta_i \geq 0$ and $\alpha_i > 0$ for all $i = 1, ..., k$, and $\sum_{i=1}^{k} \theta_i = 1$.

The posterior distribution is Dirichlet$(\alpha_1 + y_1, ...\alpha_k + y_k)$:

$$p(\theta) \propto \theta_1^{\alpha_1 + y_1 - 1} \theta_2^{\alpha_2 + y_2 - 1} ... \theta_k^{\alpha_k + y_k - 1}.$$

Multinomial distribution: Presidential elections (example)

In January 2006, Taloustutkimus (Economic Survey in Finland) interviewed 1582 adults about their preferences in the forthcoming presidential election. Out of those who expressed their opinion, 52% supported Halonen, 20% Niinistö, 18% Vanhanen, and 10% other candidates. The proportion of uncertain respondents was 29%.

Multinomial distribution: Presidential elections (example)

In January 2006, Taloustutkimus (Economic Survey in Finland) interviewed 1582 adults about their preferences in the forthcoming presidential election. Out of those who expressed their opinion, 52% supported Halonen, 20% Niinistö, 18% Vanhanen, and 10% other candidates. The proportion of uncertain respondents was 29%.

If we assume simple random sampling (which is not exactly true), the numbers of the supporters in the sample follow a multinomial distribution where $n \approx 0.71 \cdot 1582 \approx 1123$, and $\theta_1, \theta_2, \theta_3, \theta_4$ are the true proportions of the supporters of Halonen, Niinistö, Vanhanen, and other candidates, in the population of those expressing their opinion. With a uniform prior, the posterior is Dirichlet$(0.52 \cdot 1123 + 1, 0.20 \cdot 1123 + 1, 0.18 \cdot 1123 + 1, 0.1 \cdot 1123 + 1)$.

Example: Presidential elections (cont)

There were two interesting questions: 1) Will Halonen have more than 50% of the votes in the first round? 2) Will Niinistö win Vanhanen? By posterior simulation we find out that $\Pr(\theta_1 > 0.5|y) = 0.90$ and $\Pr(\theta_2 - \theta_3 > 0|y) = 0.86$. Further, the 95% posterior interval for Halonen's support is (49,55)%. Below the related JAGS code and the data given in R:

```
model{
  y ~ dmulti(theta,n)
  theta ~ ddirch(alpha)
  p1 <- step(theta[1]-0.5)
  p2 <- step(theta[2]-theta[3])
}

el <- list(n=1123,y=round(c(0.52,0.2,0.18,0.1)*1123),
      alpha=c(1,1,1,1))
```

Markov chains

Estimation of Markov Chains

Assume that we have observations $y_0, ..., y_T$ from a time homogenous Markov chain measured at time points $t = 0, 1, 2, ..., T$. Then the likelihood can be written as

$$\begin{aligned}
& \Pr(Y_0 = y_0, ..., Y_T = y_T) \\
= \; & \Pr(Y_0 = y_0 \prod_{t=1}^{T} \Pr(Y_t = y_t | Y_{t-1} = y_{t-1}) \\
= \; & \Pr(Y_0 = y_0) \prod_{t=1}^{T} p_{y_{t-1}, y_t} \\
= \; & \Pr(Y_0 = y_0) \prod_{r=1}^{S} \prod_{r=1}^{S} p_{rs}^{n_{rs}},
\end{aligned}$$

where n_{rs} denotes the number of transitions from r to s.

Estimation of Markov Chains (cont)

If we ignoring the information of the first observation, Y_0, we can write the log-likelihood as

$$l(p) = \sum_{r=1}^{S} \sum_{s=1}^{S} n_{rs} \log(p_{rs}), \tag{4}$$

and the $S \times S$ matrix of transition counts n_{rs} is a sufficient statistic. Conditioning on the row sums $n_{r.}$, the numbers of transitions starting from state r are multinomially distributed, $(n_{r1}, ..., n_{rS}) \sim \text{Multin}(n_{r.}; (p_{r1}, ..., p_{rS}))$ for all $r = 1, ..., S$.

Estimation of Markov Chains (cont)

If we ignoring the information of the first observation, Y_0, we can write the log-likelihood as

$$l(p) = \sum_{r=1}^{S} \sum_{s=1}^{S} n_{rs} \log(p_{rs}), \tag{4}$$

and the $S \times S$ matrix of transition counts n_{rs} is a sufficient statistic. Conditioning on the row sums $n_{r.}$, the numbers of transitions starting from state r are multinomially distributed, $(n_{r1}, ..., n_{rS}) \sim \text{Multin}(n_{r.}; (p_{r1}, ..., p_{rS}))$ for all $r = 1, ..., S$.

Further, the rows of this matrix are independent. From results concerning the multinomial distribution it follows that the ML estimate is $\hat{p}_{rs} = n_{rs}/n_{r.}$, for $s = 1, ..., S$ and $r = 1, ..., S$.

Estimation of Markov Chains (cont)

In a more simple model where the states Y_t are independent p_{rs} can be replaced with p_s in equation (4). The ML estimates are now $\hat{p}_s = n_{.s}/n_{..}$ where $n_{.s}$ is the sth column sum and $n_{..}$ the number of all transitions.

Estimation of Markov Chains (cont)

In a more simple model where the states Y_t are independent p_{rs} can be replaced with p_s in equation (4). The ML estimates are now $\hat{p}_s = n_{.s}/n_{..}$ where $n_{.s}$ is the sth column sum and $n_{..}$ the number of all transitions.

The likelihood ratio statistics for testing the independence hypothesis is given by

$$W = 2\sum_{r,s} n_{rs} \log\left(\frac{\hat{p}_{rs}}{\hat{p}_s}\right) = 2\sum_{r,s} n_{rs} \log\left(\frac{n_{rs}n_{..}}{n_{r.}n_{.s}}\right).$$

Under independence, there are $S-1$ free parameters, while in the general case, $S(S-1)$ parameters. Thus, under independence, the test statistic is approximately χ^2-distributed with $S(S-1)-(S-1) = (S-1)^2$ degrees of freedom. W approximately equals the Pearson statistic for independence.

Estimation of Markov Chains: Example

Table 5: Observed frequencies of one-step transitions in a DNA chain

Observed frequency

First base	A	C	G	T	Sum
A	185	74	86	171	516
C	101	41	6	115	263
G	69	45	34	78	226
T	161	103	100	202	566
Sum	516	263	226	566	1571

Let us test independence of bases in a DNA chain. Under independence, we obtain estimates $\hat{p}_A = 516/1571 = 0.328$, $\hat{p}_C = 263/1571 = 0.167$ etc. In the Markovian case, we obtain $\hat{p}_{AA} = 185/516 = 0.359$, $\hat{p}_{AC} = 74/516 = 0.143$ etc.

Estimation of Markov Chains: Example (cont)

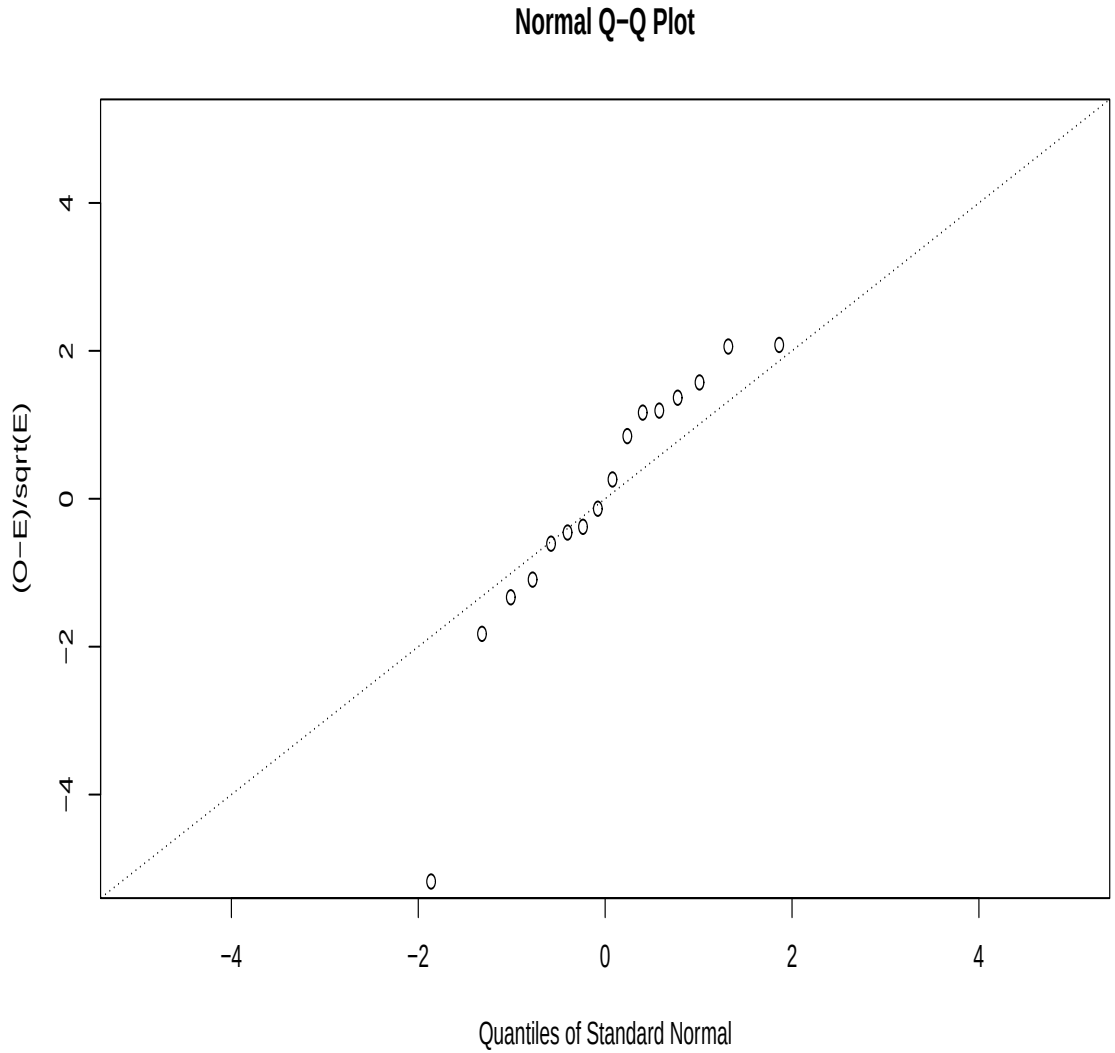
If the independence hypothesis was correct, the test statistics P and W would have approximate χ^2_9-distributions. Now their observed values are 64.45 and 50.3 which make this hypothesis highly implausible.

Estimation of Markov Chains: Example (cont)

If the independence hypothesis was correct, the test statistics P and W would have approximate χ^2_9-distributions. Now their observed values are 64.45 and 50.3 which make this hypothesis highly implausible.

The fit of the independence assumption can also be studied graphically. If this assumption was correct, the normalized deviations $Z_{rs} = (O_{rs} - E_{rs})/E_{rs}^{1/2}$, where $O_{rs} = n_{rs}$ denotes the observed and $E_{rs} = n_{r.}n_{.s}/n_{..}$ the expected frequency, would be approximately distributed as $N(0, 1)$. Figure 5 shows the normal probability plot. One observed frequency clearly deviates from the observed one (Z_{rs} is less than -5). This value belongs to the CG cell.

Estimation of Markov Chains: Example (cont)



MCMC methods

MCMC methods

In the following, we will introduce computationally intensive methods based on Markov chains which can be used in the simulation of multivariate distributions. These are called Markov Chain Monte Carlo (MCMC) methods, and they are especially useful in the computations of Bayesian statistics. The general idea is to generate a time-reversible Markov chain with a desired stationary distribution.

MCMC methods

In the following, we will introduce computationally intensive methods based on Markov chains which can be used in the simulation of multivariate distributions. These are called Markov Chain Monte Carlo (MCMC) methods, and they are especially useful in the computations of Bayesian statistics. The general idea is to generate a time-reversible Markov chain with a desired stationary distribution.

We will assume that target distribution is discrete, so that we can apply the theory of discrete state-space Markov chains. However, MCMC methods are often applied to continuous distributions, so that their proper treatment would require the theory of general state-space Markov chains. But since continuous distributions can be approximated by discrete ones with arbitrary accuracy, we can content ourselves with the theory presented by far.

Gibbs sampler

The Gibbs sampler can be used to simulate a multivarite distribution with probability function $p(\mathbf{x})$. The Gibbs sampler can be implemented if it is possible to generate random numbers from all of the full conditional distributions, denoted as $p_i(x_i|\mathbf{x}_{-i}),\ i = 1, ..., d$, where $\mathbf{x}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$.

Gibbs sampler

The Gibbs sampler can be used to simulate a multivarite distribution with probability function $p(\mathbf{x})$. The Gibbs sampler can be implemented if it is possible to generate random numbers from all of the full conditional distributions, denoted as $p_i(x_i|\mathbf{x}_{-i}),\ i = 1, ..., d$, where $\mathbf{x}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$.

The algorithm is implemented so that one first chooses the initial value vector $\mathbf{x}^0 = (x_1^0, ..., x_d^0)$. After generating the random vectors $\mathbf{x}^1, ..., \mathbf{x}^t$, the vector $\mathbf{x}^{t+1}$ is generated componentwise as follows:

- ✔ Generate x_1^{t+1} from $p_1(x_1|x_2^t, ..., x_d^t)$
- ✔ Generate x_2^{t+1} from $p_2(x_2|x_1^{t+1}, x_3^t, ..., x_d^t)$
- ✔ Generate x_3^{t+1} from $p_3(x_3|x_1^{t+1}, x_2^{t+1}, x_4^t, ..., x_d^t)$

  ...
- ✔ Generate x_d^{t+1} from $p_d(x_d|x_1^{t+1}, x_2^{t+1}, ..., x_{d-1}^{t+1})$

Gibbs sampler (cont)

The algorithm produces a Markov chain, since the distribution of $\mathbf{x}^{(t+1)}$ is independent of $\mathbf{x}^0, ..., \mathbf{x}^{(t-1)}$ given $\mathbf{x}^t$. It is time homogenous, since the transition probabilites are based on the distributions $p_j(x_j|\mathbf{x}_{-j})$ all the time. The chain is not necessarily irreducible, but it is so if the set $\{\mathbf{x} : p(\mathbf{x}) > 0\}$ is 'sufficiently' connected enabling the process to move to all points of the state space.

Gibbs sampler (cont)

The algorithm produces a Markov chain, since the distribution of $\mathbf{x}^{(t+1)}$ is independent of $\mathbf{x}^0, ..., \mathbf{x}^{(t-1)}$ given $\mathbf{x}^t$. It is time homogenous, since the transition probabilites are based on the distributions $p_j(x_j|\mathbf{x}_{-j})$ all the time. The chain is not necessarily irreducible, but it is so if the set $\{\mathbf{x} : p(\mathbf{x}) > 0\}$ is 'sufficiently' connected enabling the process to move to all points of the state space.

We show next that $p(\mathbf{x})$ fulfils the detailed balance condition

$$p(\mathbf{x}) \Pr(\mathbf{X}^{t+1} = \mathbf{x}^* | X^t = \mathbf{x}) = p(\mathbf{x}^*) \Pr(\mathbf{X}^{t+1} = \mathbf{x} | \mathbf{X}^t = \mathbf{x}^*),$$

where $\mathbf{x} = (x_1, ..., x_j, ..., x_d)$ and $\mathbf{x}^* = (x_1, ..., x_j^*, ..., x_d)$. For the moment we consider that one time step corresponds to changing only one component of $\mathbf{x}$.

Gibbs sampler (cont)

We obtain that

$$\begin{aligned} p(\mathbf{x})\Pr(\mathbf{X}^{t+1}=\mathbf{x}^*|\mathbf{X}^t=\mathbf{x}) &= p(\mathbf{x})p_j(x_j^*|\mathbf{x}_{-j}) = p(\mathbf{x})\frac{p(\mathbf{x}^*)}{p(\mathbf{x}_{-j})} \\ &= p(\mathbf{x}^*)\frac{p(\mathbf{x})}{p(\mathbf{x}_{-j})} = p(\mathbf{x}^*)p_j(x_j|\mathbf{x}_{-j}) \\ &= p(\mathbf{x}^*)\Pr(\mathbf{X}^{t+1}=\mathbf{x}|\mathbf{X}^t=\mathbf{x}^*); \end{aligned}$$

thus $p(\mathbf{x})$ is a stationary distribution.

Gibbs sampler (cont)

We obtain that

$$\begin{aligned} p(\mathbf{x})\Pr(\mathbf{X}^{t+1}=\mathbf{x}^*|\mathbf{X}^t=\mathbf{x}) &= p(\mathbf{x})p_j(x_j^*|\mathbf{x}_{-j}) = p(\mathbf{x})\frac{p(\mathbf{x}^*)}{p(\mathbf{x}_{-j})} \\ &= p(\mathbf{x}^*)\frac{p(\mathbf{x})}{p(\mathbf{x}_{-j})} = p(\mathbf{x}^*)p_j(x_j|\mathbf{x}_{-j}) \\ &= p(\mathbf{x}^*)\Pr(\mathbf{X}^{t+1}=\mathbf{x}|\mathbf{X}^t=\mathbf{x}^*); \end{aligned}$$

thus $p(\mathbf{x})$ is a stationary distribution.

Irreducibility implies the uniqueness of the stationary distribution. The chain is also positively recurrent, since transient and null recurrent chains do not posses a stationary distribution. Further, it is aperiodic, since the new value can be the same as the old. It follows from these properties that the chain is ergodic.

Metropolis algorithm

The Metropolis algorithm is different from Gibbs sampling in that it does not require ability to generate random variates from conditional distributions. It is sufficient to know the probability function (or density) of the target density upto a constant of proportionality.

Metropolis algorithm

The Metropolis algorithm is different from Gibbs sampling in that it does not require ability to generate random variates from conditional distributions. It is sufficient to know the probability function (or density) of the target density upto a constant of proportionality.

Assume that we want to simulate a distribution with probability function $p(x)$ where x may be scalar or vector. We need to define a jumping distribution (or proposal distribution) $J(y|x)$ from which a proposal y may be generated when the current value is x. In the Metropolis algorithm it is assumed that $J(y|x) = J(x|y)$.

Metropolis algorithm (cont)

An initial value x_0 is first generated. After generating $x_0, ..., x_t$, the new value x_{t+1} is obtained as follows: 1) A new proposal y is generated from $J(y|x)$. The new value y is accepted with probability

$$\min\left(1, \frac{p(y)}{p(x_t)}\right).$$

2) If the new value is accepted, we set $x_{t+1} = y$, otherwise the old value is kept, so that $x_{t+1} = x_t$.

Metropolis algorithm (cont)

An initial value x_0 is first generated. After generating $x_0, ..., x_t$, the new value x_{t+1} is obtained as follows: 1) A new proposal y is generated from $J(y|x)$. The new value y is accepted with probability

$$\min\left(1, \frac{p(y)}{p(x_t)}\right).$$

2) If the new value is accepted, we set $x_{t+1} = y$, otherwise the old value is kept, so that $x_{t+1} = x_t$.

The Metropolis algorithm produces a Markov chain, since the distribution of the new value x_{t+1} only depends on the current value x_t. The chain is also time-homogenous, since the transition probabilities are based on the jumping distribution $J(y|x)$, which is not changed during the simulation. Further, it is irreducible if $J(y|x)$ is so chosen that the chain may reach all points of the state space.

Metropolis algorithm (cont)

Next we show that $p(x)$ fulfils the detailed balance condition. Let x and x^* be two points in the state space such that $p(x^*) \leq p(x)$. Then

$$\begin{aligned} p(x)\Pr(X_{t+1} = x^* | X_t = x) &= p(x) J(x^*|x) \frac{p(x^*)}{p(x)} \\ &= p(x^*) J(x|x^*) \\ &= p(x^*) \Pr(X_{t+1} = x | X_t = x^*). \end{aligned}$$

Metropolis algorithm (cont)

Next we show that $p(x)$ fulfils the detailed balance condition. Let x and x^* be two points in the state space such that $p(x^*) \leq p(x)$. Then

$$\begin{aligned} p(x)\Pr(X_{t+1} = x^* | X_t = x) &= p(x) J(x^*|x) \frac{p(x^*)}{p(x)} \\ &= p(x^*) J(x|x^*) \\ &= p(x^*) \Pr(X_{t+1} = x | X_t = x^*). \end{aligned}$$

Thus, $p(x)$ is the stationary distribution and the chain is positively recurrent. Further, since it is also aperiodic, it is ergodic.

Metropolis algorithm (cont)

It is said that a Markov chain *mixes* slowly if it moves slowly around the support of $p(x)$. Then there is strong autocorrelation between the consequtive observations, and the mean converges slowly to the theoretical mean of the stationary distribution.

Metropolis algorithm (cont)

It is said that a Markov chain *mixes* slowly if it moves slowly around the support of $p(x)$. Then there is strong autocorrelation between the consequtive observations, and the mean converges slowly to the theoretical mean of the stationary distribution.

There are two possible reasons for this problem. First, if the deviation of the jumping distribution is too small for some component, the chain moves slowly with respect to that component. On the other hand, if the deviation is too large, new proposals are rarely accepted and the chain remains long in the same position.

Metropolis algorithm (cont)

It is said that a Markov chain *mixes* slowly if it moves slowly around the support of $p(x)$. Then there is strong autocorrelation between the consequtive observations, and the mean converges slowly to the theoretical mean of the stationary distribution.

There are two possible reasons for this problem. First, if the deviation of the jumping distribution is too small for some component, the chain moves slowly with respect to that component. On the other hand, if the deviation is too large, new proposals are rarely accepted and the chain remains long in the same position.

It is possible to optimize the jumping distribution. If the jumping distribution is a d-dimensional normal distribution, then its optimal covariance matrix is $c^2\Sigma$ where $c \approx 2.4/\sqrt{d}$ and Σ is the covariance matrix of the target distribution.

Metropolis algorithm: Example

Let us consider a two-parameter Weibull distribution with the density

$$f(x;\beta,\delta) = \frac{\delta}{\beta^{\delta}} x^{\delta-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{\delta}\right\}, \quad x,\beta,\delta > 0.$$

Metropolis algorithm: Example

Let us consider a two-parameter Weibull distribution with the density

$$f(x;\beta,\delta) = \frac{\delta}{\beta^{\delta}} x^{\delta-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{\delta}\right\}, \quad x,\beta,\delta > 0.$$

With a random sample $y_1, ..., y_n$ the likelihood is

$$p(y|\theta) = \frac{\delta^n}{\beta^{n\delta}} \left(\prod_i y_i\right)^{\delta-1} \exp\left\{-\sum_i \left(\frac{y_i}{\beta}\right)^{\delta}\right\}.$$

Metropolis algorithm: Example

Let us consider a two-parameter Weibull distribution with the density

$$f(x;\beta,\delta) = \frac{\delta}{\beta^{\delta}} x^{\delta-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{\delta}\right\}, \quad x,\beta,\delta > 0.$$

With a random sample $y_1, ..., y_n$ the likelihood is

$$p(y|\theta) = \frac{\delta^n}{\beta^{n\delta}} \left(\prod_i y_i\right)^{\delta-1} \exp\left\{-\sum_i \left(\frac{y_i}{\beta}\right)^{\delta}\right\}.$$

By choosing $p(\beta,\delta) \propto 1/(\beta\delta)$ as the prior, the posterior becomes

$$p(\beta,\delta|y) \propto \frac{\delta^{n-1}}{\beta^{n\delta+1}} \left(\prod_i y_i\right)^{\delta-1} \exp\left\{-\sum_i \left(\frac{y_i}{\beta}\right)^{\delta}\right\}.$$

Metropolis algorithm: Example (cont)

It would be possible to derive the full conditional posterior distributions and simulate the posterior distribution using Gibbs sampling. We could generate random numbers from the conditional distributions using adaptive rejection sampling. However, it is here simpler to apply the Metropolis algorithm.

Metropolis algorithm: Example (cont)

It would be possible to derive the full conditional posterior distributions and simulate the posterior distribution using Gibbs sampling. We could generate random numbers from the conditional distributions using adaptive rejection sampling. However, it is here simpler to apply the Metropolis algorithm.

To illustrate the estimation, we generate an artificial data set of 100 observations from the Weibull(0.3,10) distribution. Figure 6 shows a simulated Markov chain with 10000 iterations, starting from the initial values $\delta = \beta = 1$. As a jumping distribution we use the bivariate normal distribution, the mean vector being the 'old' vector and the covariance matrix $\mathsf{diag}(0.01, 10)$.

Metropolis algorithm: Example (cont)

It would be possible to derive the full conditional posterior distributions and simulate the posterior distribution using Gibbs sampling. We could generate random numbers from the conditional distributions using adaptive rejection sampling. However, it is here simpler to apply the Metropolis algorithm.

To illustrate the estimation, we generate an artificial data set of 100 observations from the Weibull(0.3,10) distribution. Figure 6 shows a simulated Markov chain with 10000 iterations, starting from the initial values $\delta = \beta = 1$. As a jumping distribution we use the bivariate normal distribution, the mean vector being the 'old' vector and the covariance matrix $\mathsf{diag}(0.01, 10)$.

The figure shows that the chain converges to its stationary distribution rapidly but the chain for β seems to mix poorly.

Metropolis algorithm: Example (cont)

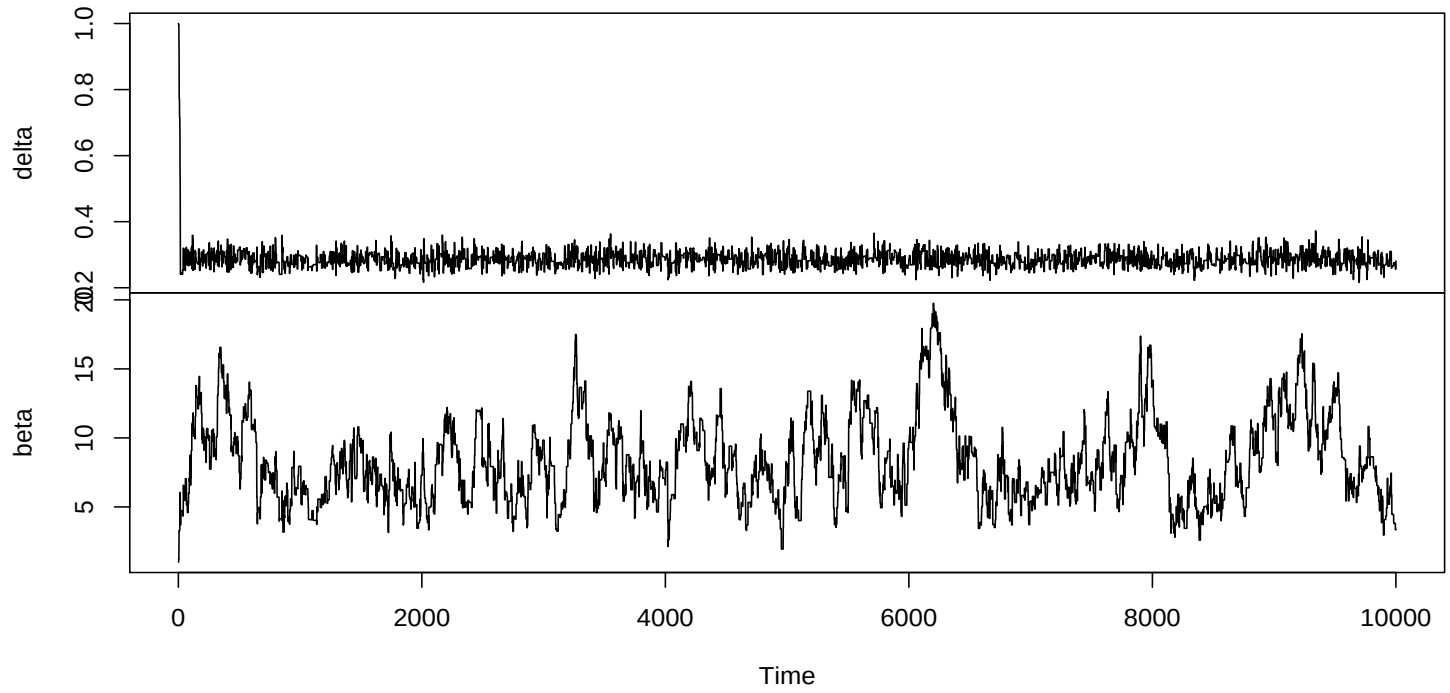


Figure 6: Estimating the parameteres of the Weibull distribution: 10000 iterations of the Metropolis algorithm

Metropolis algorithm: Example (cont)

Next we simulate 10000 new observations using the optimal covariance matrix $2.4^2\Sigma/2$ where Σ is the covariance matrix of the target distribution, estimated using the most recent simulations of the original chain. As an initial value we use the last simulated vector of the first chain. On the basis of Figure 7 the mixing is more rapid now. Figure 8 shows the graphs of the 2.5%, 50% and 97% cumulative quantiles.

Metropolis algorithm: Example (cont)

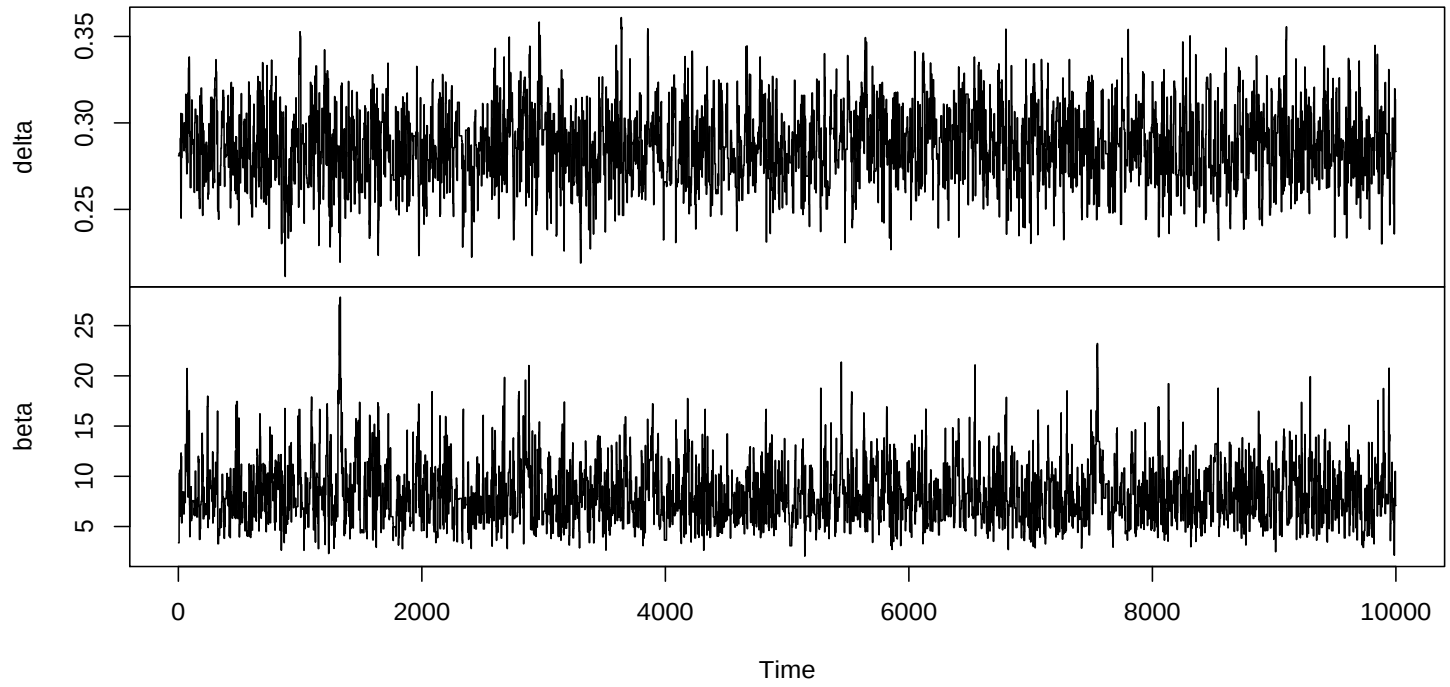


Figure 7: Estimating the parameteres of the Weibull distribution: 10000 further iterations of the Metropolis algorithm

Metropolis algorithm: Example (cont)

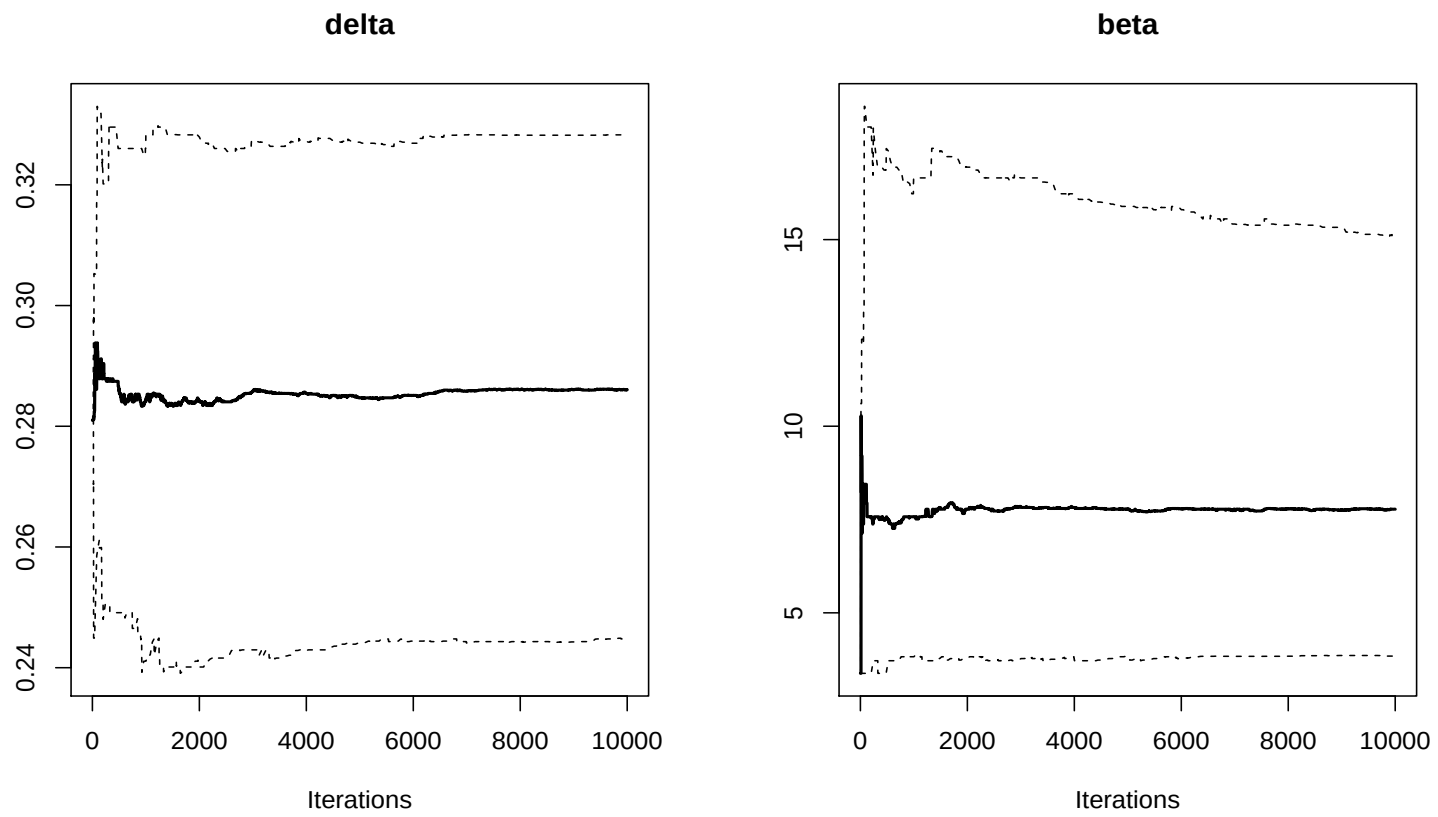


Figure 8: 2.5%, 50% and 97% cumulative quantiles of 10000 posterior simulations

Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is similar to the Metropolis algorithm except that it is not assumed that the jumping distribution $J(y|x)$ is symmetric with respect to the 'old' value x. The acceptance probablity of a proposal is now

$$\min\left(1, \frac{p(y)/J(y|x_t)}{p(x_t)/J(x_t|y)}\right).$$

It can be shown (exercise) that the algorithm produces a Markov chain with stationary distribution $p(x)$.

Detecting convergence

Markov Chain simulation should be continued until reaching the stationary distribution, and after this until reliable estimates for the summary statistics of the stationary distribution have been obtained. The iterations before the convergence are usually disregarded as a *burn-in* phase.

Detecting convergence

Markov Chain simulation should be continued until reaching the stationary distribution, and after this until reliable estimates for the summary statistics of the stationary distribution have been obtained. The iterations before the convergence are usually disregarded as a *burn-in* phase.

In practice, convergence to stationary distribution can be detected by studying various time series plots, such as trace plots, and plots of cumulative summary statistics and autoregression functions.

Detecting convergence

Markov Chain simulation should be continued until reaching the stationary distribution, and after this until reliable estimates for the summary statistics of the stationary distribution have been obtained. The iterations before the convergence are usually disregarded as a *burn-in* phase.

In practice, convergence to stationary distribution can be detected by studying various time series plots, such as trace plots, and plots of cumulative summary statistics and autoregression functions.

However, it is usually more reliable to also use convergence diagnostics. Geweke's diagnostic is based on comparing the means of the beginning and last parts of the chain. In the following, we will introduce Gelman and Rubin's diagnostic, which is based on comparing several simulated chains.

Gelman and Rubin's diagnostic

Suppose we have simulated m chains of n iterations (after removing the burn-in phase). We denote the simulations by ψ_{ij} $(i = 1, ..., n; j = 1, ..., n)$, and compute B and W, the between- and within-sequence variances:

Gelman and Rubin's diagnostic

Suppose we have simulated m chains of n iterations (after removing the burn-in phase). We denote the simulations by ψ_{ij} $(i = 1, ..., n; j = 1, ..., n)$, and compute B and W, the between- and within-sequence variances:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})^2,$$

where $\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$, $\quad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$, and

Gelman and Rubin's diagnostic

Suppose we have simulated m chains of n iterations (after removing the burn-in phase). We denote the simulations by ψ_{ij} $(i = 1, ..., n; j = 1, ..., n)$, and compute B and W, the between- and within-sequence variances:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})^2,$$

where $\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$, $\quad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$, and

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{.j})^2$$

Gelman and Rubin's diagnostic

Suppose we have simulated m chains of n iterations (after removing the burn-in phase). We denote the simulations by ψ_{ij} $(i = 1, ..., n; j = 1, ..., n)$, and compute B and W, the between- and within-sequence variances:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})^2,$$

where $\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$, $\quad \bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$, and

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \text{ where } s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{.j})^2$$

We can estimate the posterior variance $\mathsf{Var}(\psi|y)$ by the weighted average $\widehat{\mathsf{Var}}^{+}(\psi|y) = \frac{n-1}{n} W + \frac{1}{n} B$.

Gelman and Rubin's diagnostic (cont)

The quantity $\widehat{\mathsf{Var}}^{+}(\psi|y)$ *overestimates* the posterior variance if the starting values are overdispersed, but is unbiased under stationarity. On the other hand, W *underestimates* posterior variance for any finite n because the individual sequences have not had time to range over all of the target distribution.

Gelman and Rubin's diagnostic (cont)

The quantity $\widehat{\mathsf{Var}}^{+}(\psi|y)$ *overestimates* the posterior variance if the starting values are overdispersed, but is unbiased under stationarity. On the other hand, W *underestimates* posterior variance for any finite n because the individual sequences have not had time to range over all of the target distribution.

We may monitor convergence using the *potential scale factor*

$$\hat{R} = \sqrt{\frac{\widehat{\mathsf{Var}}^{+}(\psi|y)}{W}}$$

which tells by which factor the posterior deviation estimate can be decreased if simulation is continued. Simulation should be continued until $\hat{R}$ is close to 1 for each parameter ψ. In most practical cases, values below 1.1 would be acceptable.

Gelman and Rubin's diagnostic: Example

To illustrate the use of the diagnostic, we continue our example on the Weibull distribution. We generate 5 chains of length 1000 using random initial values. After removing the first 500 simulations from each chain, we obtain the following diagnostics. Also a multivariate version of the diagnostic is computed. Here, gelman.diag is a function in R package coda and SIMS is an mcmc object containing the chains.

```
 1> gelman.diag(SIMS)
Potential scale reduction factors:

      Point est. Upper C.I.
delta       1.00       1.01
beta        1.01       1.02

Multivariate psrf
1.01
```

Gelman and Rubin's diagnostic: Example (cont)

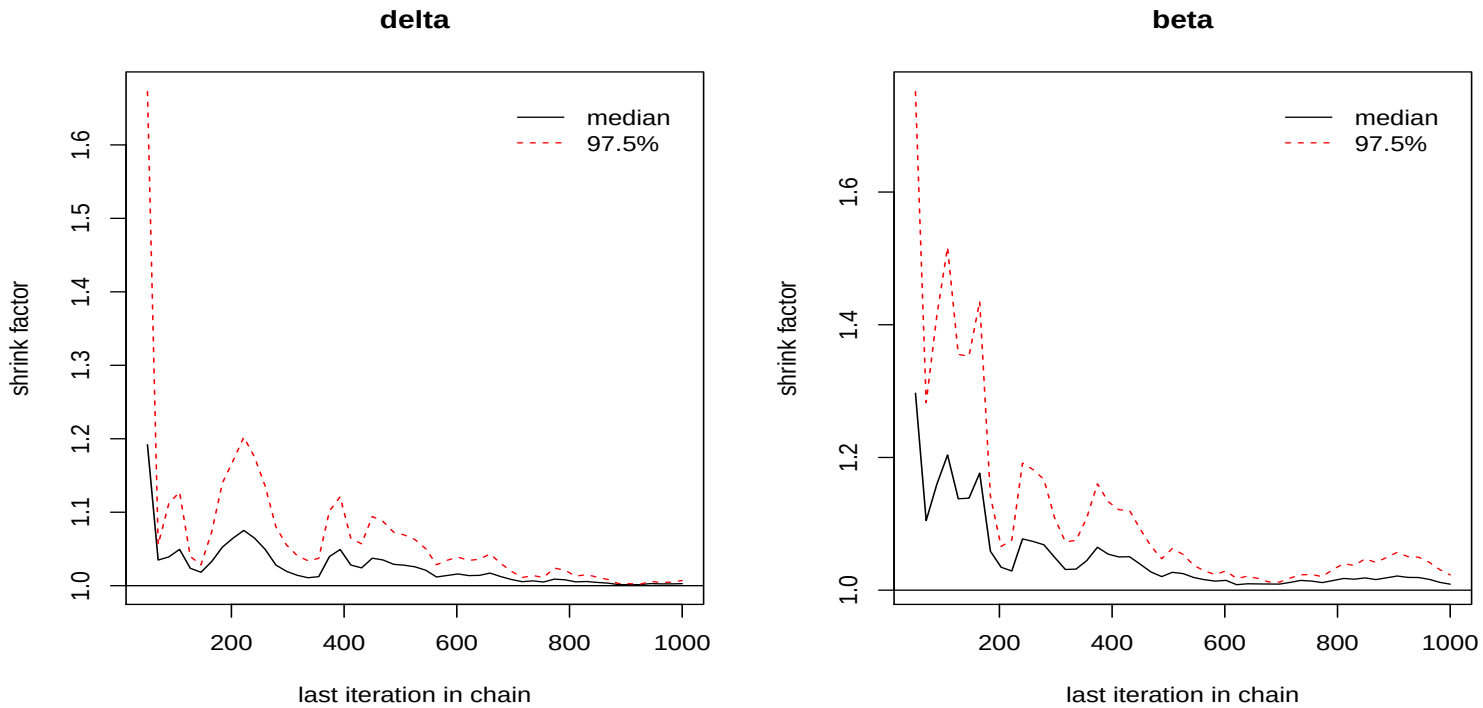


Figure 9: The Gelman-Rubin shrink factor might be close to 1 by chance. Therefore, a graph (gelman.plot) showing its convergence is useful. Here, the curves show the diagnostic and its 97.5% quantile for the observation intervals 25:50, 30:60, ..., 500:1000.

Model checking and comparison

Model checking and comparison

The conclusions of a Bayesian analysis are conditional on the chosen probability model. Therefore, it is essential to check that the model is a reasonable approximation to reality. Model checking can be done with respect to outliers, sampling distribution, prior distribution, link function, covariates and so on.

Model checking and comparison

The conclusions of a Bayesian analysis are conditional on the chosen probability model. Therefore, it is essential to check that the model is a reasonable approximation to reality. Model checking can be done with respect to outliers, sampling distribution, prior distribution, link function, covariates and so on.

We can distinguish three aspects of modelling:

- ✔ *Criticism:* exploratory checking of a single model
- ✔ *Extension:* embedding a model in a larger model
- ✔ *Comparison:* comparing candidate models in terms of their fit and predictive power

Residuals

A widely used and useful technique for model checking is plotting residuals. They help, for example, detect outliers, autocorrelation and problems in distributional assumptions. They measure the deviation between observations and estimated expected values.

Residuals

A widely used and useful technique for model checking is plotting residuals. They help, for example, detect outliers, autocorrelation and problems in distributional assumptions. They measure the deviation between observations and estimated expected values.

A *Pearson residual* is defined as

$$r_i(\theta) = \frac{y_i - \mathsf{E}(y_i|\theta)}{\sqrt{\mathsf{Var}(y_i|\theta)}}.$$

In classical analysis, θ is replaced by its fitted value, while in Bayesian analysis the residuals have a posterior distribution.

Residuals

A widely used and useful technique for model checking is plotting residuals. They help, for example, detect outliers, autocorrelation and problems in distributional assumptions. They measure the deviation between observations and estimated expected values.

A *Pearson residual* is defined as

$$r_i(\theta) = \frac{y_i - \mathsf{E}(y_i|\theta)}{\sqrt{\mathsf{Var}(y_i|\theta)}}.$$

In classical analysis, θ is replaced by its fitted value, while in Bayesian analysis the residuals have a posterior distribution.

Example. We consider the child heart surgery data in Table 1. Figure 10 shows the box plot of Pearson residuals assuming that $y_i \sim \text{Bin}(\theta, m_j)$. Hospital H appears to be an outlier.

Residuals (cont)

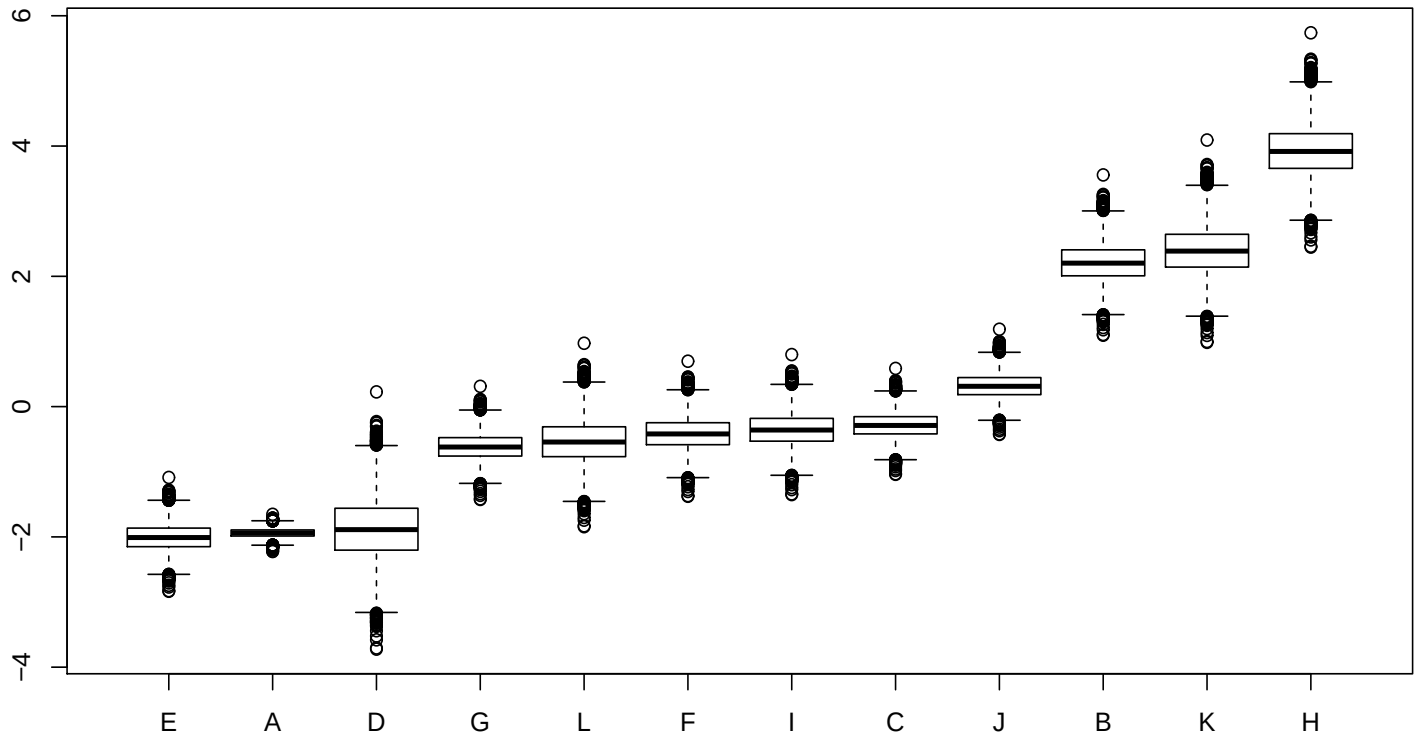


Figure 10: Box plot of Pearson residuals for heart surgery data

Residuals: code for making the box plot


```
model{
  theta ~ dbeta(1,1)
  for(j in 1:J){
    y[j] ~ dbin(theta,m[j])
    res[j] <- (y[j]-m[j]*theta)/sqrt(m[j]*theta*(1-theta))
  }}

hospital <- list(J=J,m=m,y=y)
hospital.jag <- jags.model("Hospital2.txt",hospital)
hospital.coda <- coda.samples(hospital.jag,c("theta","res"),1

med <- apply(hospital.coda[[1]][,-13],2,median)
ind <- order(med)
Res <- as.list(1:J)
for(j in 1:J) Res[[j]] <-
  c(hospital.coda[[1]][,paste("res[",ind[j],"]",sep="")])
boxplot(Res,names=names(y)[ind])
```

Predictive checks and Bayesian p-values

Residuals are examples of statistics which measure the discrepancy between the data and the assumed model. These statistics are usually easy to calculate, but we beed a method to determine if the observed discrepancy is significant. Here, we may use so-called Bayesian p-values obtained by simulating the posterior predictive distribution of the test statistic.

Predictive checks and Bayesian p-values

Residuals are examples of statistics which measure the discrepancy between the data and the assumed model. These statistics are usually easy to calculate, but we beed a method to determine if the observed discrepancy is significant. Here, we may use so-called Bayesian p-values obtained by simulating the posterior predictive distribution of the test statistic.

Ideally, models should be checked by comparing the predictions of a model to new data. Suppose that the data y is divided into two parts: y_f for fitting the model, and y_c for model criticism. Then the comparisons are based on the predictive distribution,

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

simulated by drawing θ from $p(\theta|y_f)$ and y_c^{pred} from $p(y_c^{pred}|\theta)$.

Predictive checks and Bayesian p-values (cont)

A function $T(y_c)$ is called a test statistic (Gelman et al., 2004) if it would have an extreme value if the data y_c conflict with the assumed model. By choosing $T(y_c) = y_{ci}$ one can check for individual outliers.

Predictive checks and Bayesian p-values (cont)

A function $T(y_c)$ is called a test statistic (Gelman et al., 2004) if it would have an extreme value if the data y_c conflict with the assumed model. By choosing $T(y_c) = y_{ci}$ one can check for individual outliers.

One can check whether $T(y_c)$ is extreme graphically or by computing the Bayesian p-value

$$p = \Pr(T(y_c^{pred}) \leq T(y_c)|y_f).$$

This can be obtained by drawing simulations y_c^{pred} from the posterior predictive distribution, and by calculating the proportion of cases where $T(y_c^{pred}) \leq T(y_c)$.

Predictive checks and Bayesian p-values (cont)

A function $T(y_c)$ is called a test statistic (Gelman et al., 2004) if it would have an extreme value if the data y_c conflict with the assumed model. By choosing $T(y_c) = y_{ci}$ one can check for individual outliers.

One can check whether $T(y_c)$ is extreme graphically or by computing the Bayesian p-value

$$p = \Pr(T(y_c^{pred}) \leq T(y_c)|y_f).$$

This can be obtained by drawing simulations y_c^{pred} from the posterior predictive distribution, and by calculating the proportion of cases where $T(y_c^{pred}) \leq T(y_c)$.

In practice, the same data set is often used for fitting and checking ($y_c = y_f = y$). In this case the diagnostics are likely to be conservative.

Predictive checks and Bayesian p-values: Example 1

In the previous example of cardic surgery death rates the value of the hospital H appeared to be an outlier. We may compute its predictive p-value (using the *mid p-value* $\Pr(y_i^{pred} > y_i|y_{-i}) + \frac{1}{2}\Pr(y_i^{pred} = y_i|y_{-i})$ for discrete data):

```
model{
  theta ~ dbeta(1,1)
  for(j in 1:7){y[j] ~ dbin(theta,m[j])}
  for(j in 9:J){y[j] ~ dbin(theta,m[j])}
  #predicted number of deaths in the 8th hospital
  y8.pred ~ dbin(theta,m[8])
  P <- step(y8.pred-y[8]-0.001)+0.5*equals(y8.pred,y[8])
}
```

```
           Mean      SD  Naive SE Time-series SE
P       0.00035 0.01803 0.0001803      0.0001803
y8.pred 14.68810 3.86418 0.0386418      0.0386418
```

Predictive checks and Bayesian p-values: Example 2

We continue our study of the Newcomb data, and use the statistics $T_1 = \min(y)$ and $T_2 = (y_{(1)} - y_{(n/2)})/(y_{(n/4)} - y_{(n/2)})$, where $y_{(j)}$ is the jth lowest value of y.

```
for(i in 1:n){
   y[i] ~ dnorm(mu,tau)
   yrep[i] ~ dnorm(mu,tau)
 }
 n.50 <- round(n/2)
 n.25 <- round(n/4)
 yrep.sort <- sort(yrep[])
 T1.rep <- yrep.sort[1]
 yrep.50 <- yrep.sort[n.50]
 yrep.25 <- yrep.sort[n.25]
 T2.rep <- (T1.rep-yrep.50)/(yrep.25-yrep.50)
 P1 <- step(T1.rep-T1.obs)
 P2 <- step(T2.rep-T2.obs)
```

Predictive checks and Bayesian p-values: Example 2

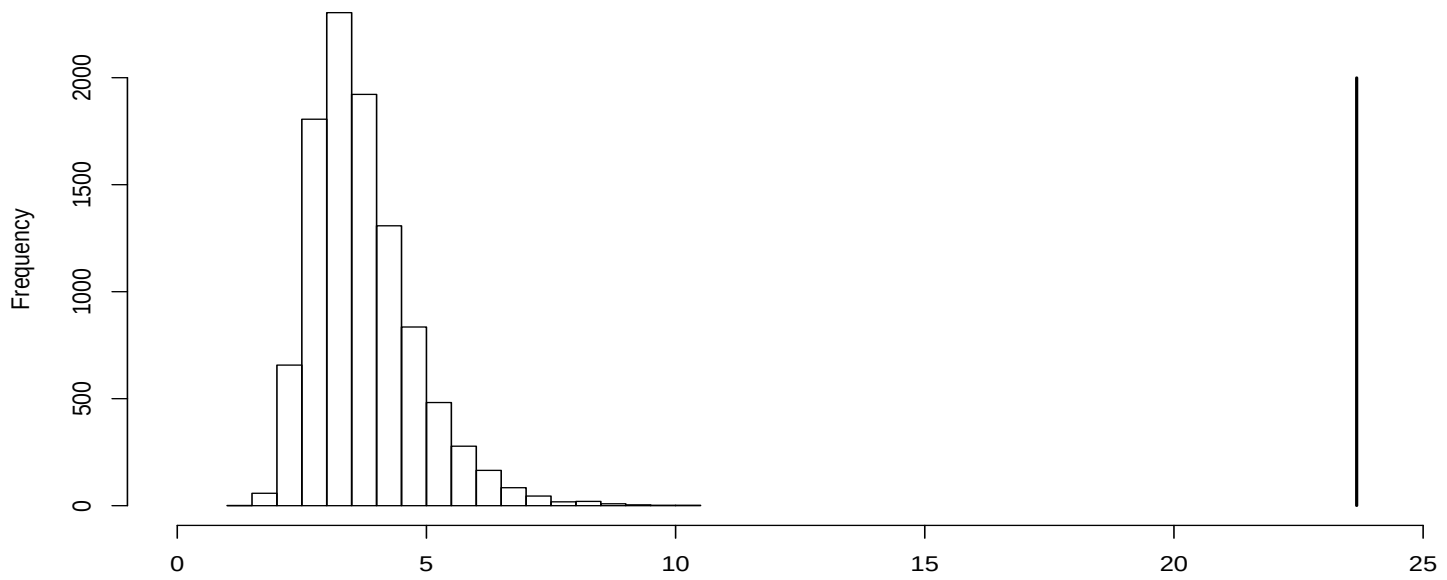


Figure 11: The figure shows the posterior predictive distribution of T_2. We see that T_2^{obs} indicated by a vertical line would be implausibly large if the model was correct.

Model comparison using deviances

Model fit can be summarized with the deviance, defined as

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

where $p(y|\theta)$ is the likelihood function.

Model comparison using deviances

Model fit can be summarized with the deviance, defined as

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

where $p(y|\theta)$ is the likelihood function.To obtain a summary that depends on y only, θ can be replaced with a point estimate $\hat{\theta}$, such as posterior mean. We obtain

$$D(\hat{\theta}) = -2 \log p(y|\hat{\theta}).$$

Model comparison using deviances

Model fit can be summarized with the deviance, defined as

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

where $p(y|\theta)$ is the likelihood function.To obtain a summary that depends on y only, θ can be replaced with a point estimate $\hat{\theta}$, such as posterior mean. We obtain

$$D(\hat{\theta}) = -2 \log p(y|\hat{\theta}).$$

This may give an over-optimistic picture of the model fit.

Model comparison using deviances

Model fit can be summarized with the deviance, defined as

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

where $p(y|\theta)$ is the likelihood function.To obtain a summary that depends on y only, θ can be replaced with a point estimate $\hat{\theta}$, such as posterior mean. We obtain

$$D(\hat{\theta}) = -2 \log p(y|\hat{\theta}).$$

This may give an over-optimistic picture of the model fit. A natural Bayesian alternative is the posterior mean deviance

$$\bar{D} = \mathsf{E}(D(\theta)|y).$$

Model comparison using deviances (cont)

It is easy to estimate $\bar{D}$ using posterior simulations θ^l:

$$\hat{\bar{D}} = \frac{1}{L} \sum_{l=1}^{L} D(\theta^l).$$

Model comparison using deviances (cont)

It is easy to estimate $\bar{D}$ using posterior simulations θ^l:

$$\hat{\bar{D}} = \frac{1}{L}\sum_{l=1}^{L} D(\theta^l).$$

The difference between the posterior mean deviance and the deviance at $\hat{\theta}$ represents the effect of model fitting and is called the *effective number of parameters*:

$$p_D = \bar{D} - D(\hat{\theta}).$$

Model comparison using deviances (cont)

It is easy to estimate $\bar{D}$ using posterior simulations θ^l:

$$\hat{\bar{D}} = \frac{1}{L}\sum_{l=1}^{L} D(\theta^l).$$

The difference between the posterior mean deviance and the deviance at $\hat{\theta}$ represents the effect of model fitting and is called the *effective number of parameters*:

$$p_D = \bar{D} - D(\hat{\theta}).$$

In nonhierarchical models, if the number of observations is large or the prior information is weak, p_D is usually approximately equal to the actual number of parameters.

Deviance information criterion, DIC

When the goal is to choose an optimal model for prediction, the expected predictive deviance,

$$\mathsf{E}[-2\log(p(y^{rep},\hat{\theta}(y)))],$$

has been suggested as a criterion of model fit. Here the expectation is taken over the unknown true distribution of y^{rep}.

Deviance information criterion, DIC

When the goal is to choose an optimal model for prediction, the expected predictive deviance,

$$\mathsf{E}[-2\log(p(y^{rep}, \hat{\theta}(y)))],$$

has been suggested as a criterion of model fit. Here the expectation is taken over the unknown true distribution of y^{rep}.

This can be approximated by the *deviance information criterion* (DIC):

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

Deviance information criterion, DIC

When the goal is to choose an optimal model for prediction, the expected predictive deviance,

$$\mathsf{E}[-2\log(p(y^{rep},\hat{\theta}(y)))],$$

has been suggested as a criterion of model fit. Here the expectation is taken over the unknown true distribution of y^{rep}.

This can be approximated by the *deviance information criterion* (DIC):

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

This can usually be easily computed using posterior simulation. When the prior information is weak or the sample size large, $p \approx p_D$, implying that DIC $\approx$ AIC.

Deviance information criterion (example)

We regress the incidence of pine processionary caterpillars on 8 potential exploratory variables. (The data set is caterpillar in R package bayess.) The response variable y is the log transform of the average number of nests per tree.

The explanatory variables:
x_1 altitude (in meters)
x_2 slope (in degrees)
x_3 number of pine trees in the area
x_4 height (in meters) of the tree sampled at the center of the area
x_5 orientation of the area (from 1 if southbound to 2 otherwise)
x_6 height (in meters) of the dominant tree
x_7 number of vegetation strata
x_8 mix settlement index (from 1 if not mixed to 2 if mixed)

DIC example (cont)


```
#JAGS code
for(i in 1:n){
  y[i] ~ dnorm(mu[i],tau)
  mu[i] <- b0 +b[1]*X[i,1]+b[2]*X[i,2]+b[3]*X[i,3]+
           b[4]*X[i,4]+b[5]*X[i,5]+b[6]*X[i,6]+
           b[7]*X[i,7]+b[8]*X[i,8]
}
b0 ~ dnorm(0,0.001)
tau ~ dgamma(0.001,0.001)
for(j in 1:8){
   b[j] ~ dnorm(0,0.001)

#R code
cp.jag <- jags.model("caterpillar.txt",data,n.chains=2)
cp.coda <- coda.samples(cp.jag,c("b0","b","tau"),10000)
summary(cp.coda)
dic.samples(cp.jag,n.iter=100000,type="pD")
```

DIC example (cont)

According to the results, only β_1, β_2 and β_7 are 'significant' in the sense that 0 is not included in their 95% posterior intervals.

	2.5%	25%	50%	75%	97.5%
b[1]	-0.6285	-0.4439	-0.35067	-0.25715	-0.06983
b[2]	-0.4792	-0.3324	-0.25874	-0.18520	-0.03537
b[3]	-0.1268	0.2279	0.39554	0.55959	0.89309
b[4]	-0.4393	-0.1638	-0.02962	0.10728	0.36899
b[5]	-0.3364	-0.1891	-0.11807	-0.04442	0.10419
b[6]	-0.6400	-0.2028	0.03315	0.26520	0.72730
b[7]	-1.2519	-0.8400	-0.63957	-0.44101	-0.04077
b[8]	-0.2799	-0.1311	-0.05702	0.01804	0.17237
b0	0.6114	0.7450	0.81175	0.87794	1.01362
tau	1.6534	2.5258	3.10635	3.76617	5.25581

DIC example (cont)

Now, the model selection criteria are estimated as follows: $\hat{\bar{D}} = 56.7$, $\hat{p}_D = 10.95$ and $\widehat{\text{DIC}} = 67.65$. When the unsignificant variables are removed both $\hat{\bar{D}}$ and $\widehat{\text{DIC}}$ become smaller, indicating a better model.

```
#Original model
Mean deviance:  56.7
penalty 10.95
Penalized deviance: 67.65

#Restricted model
Mean deviance:  55.52
penalty 5.369
Penalized deviance: 60.89
```

Hierarchical and regression models

Categorical data

Generalized linear model

In linear models it is assumed that the response variable is normally distributed and its expected value is a linear combination of the explanatory variables. Generalized linear models extend the idea of linear modelling to cases where either of these assumptions may not be appropriate.

Generalized linear model

In linear models it is assumed that the response variable is normally distributed and its expected value is a linear combination of the explanatory variables. Generalized linear models extend the idea of linear modelling to cases where either of these assumptions may not be appropriate.

A generalized linear model is specified in three stages:

- ✔ The linear predictor $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$
- ✔ The *link function* $g(.)$ which relates the linear predictor to the mean of the response variable: $g(\mu_i) = \eta_i$, where $\mu_i = \mathsf{E}(y_i)$
- ✔ The distribution of y_i given its mean μ_i. In general, this distribution may also depend on a *dispersion parameter* ϕ.

Binomial regression

Binomial regression is perhaps the most popular application of the generalized linear model. Suppose that $y_i \sim \text{Bin}(n_i, \mu_i)$ where n_i is known. Then one usually specifies a model for μ_i, the mean of y_i/n_i. Choosing the logistic transformation $g(\mu_i) = \log(\mu_i/(1-\mu_i))$ leads to logistic regression. The likelihood in this case is

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{n_i - y_i}.$$

Binomial regression

Binomial regression is perhaps the most popular application of the generalized linear model. Suppose that $y_i \sim \text{Bin}(n_i, \mu_i)$ where n_i is known. Then one usually specifies a model for μ_i, the mean of y_i/n_i. Choosing the logistic transformation $g(\mu_i) = \log(\mu_i/(1-\mu_i))$ leads to logistic regression. The likelihood in this case is

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{n_i - y_i}.$$

Another popular choice for a link function is the probit link $g(\mu) = \Phi^{-1}(\mu)$ where $\Phi(.)$ is the distribution function of a standard normal variable. The likelihood becomes

$$p(y|\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} (\Phi(\eta_i))^{y_i} (1-\Phi(\eta_i))^{n_i - y_i}.$$

Binomial regression: Example

Table 6: Bioassay data from Racine et al. (1986).

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths , y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

As an example we consider 4 batches of 5 animals, each of which is given a different dose of a drug. We are interested in determining the toxity of the drug. Table 6 reports the numbers of deaths for the different dose levels.

Binomial regression: Example (cont)

We assume that the numbers of deaths are binomially distributed,

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i),$$

and that there is a simple linear linear relationship between the logit of the mortality θ_i and the dose level x_i:

$$\text{logit}(\theta_i) = \alpha + \beta x_i,$$

where $\text{logit}(\theta_i) = \log(\theta_i/(1-\theta_i))$.

Binomial regression: Example (cont)

We assume that the numbers of deaths are binomially distributed,

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i),$$

and that there is a simple linear linear relationship between the logit of the mortality θ_i and the dose level x_i:

$$\text{logit}(\theta_i) = \alpha + \beta x_i,$$

where $\text{logit}(\theta_i) = \log(\theta_i/(1-\theta_i))$. Now the posterior of (α, β) is

$$p(\alpha, \beta|y) \propto \; p(\alpha, \beta) \prod_{i=1}^{k} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i},$$

where $p(\alpha, \beta)$ is the prior, $\theta_i = e^{\eta_i}/(1+e^{\eta_i})$ and $\eta_i = \alpha + \beta x_i$.

Binomial regression: Example (cont)

There are several ways to specify the prior information. However, we assume that $\alpha \sim N(0, 1000)$ and $\beta \sim \text{TN}(0, 1000; 0, \infty)$ (truncated normal distribution). We truncate the prior of β from below at 0, since we believe that the dose is harmful so that $\beta \geq 0$.

Binomial regression: Example (cont)

There are several ways to specify the prior information. However, we assume that $\alpha \sim N(0, 1000)$ and $\beta \sim \text{TN}(0, 1000; 0, \infty)$ (truncated normal distribution). We truncate the prior of β from below at 0, since we believe that the dose is harmful so that $\beta \geq 0$.

We also wish to determine the LD50, the dose level at which the probability of death is 50%. Thus, we determine x so that

$$\text{logit}(0.5) = \alpha + \beta x.$$

Solving this gives that the LD50 is $x = -\beta/\alpha$.

Binomial regression: Example (cont)

There are several ways to specify the prior information. However, we assume that $\alpha \sim N(0, 1000)$ and $\beta \sim \text{TN}(0, 1000; 0, \infty)$ (truncated normal distribution). We truncate the prior of β from below at 0, since we believe that the dose is harmful so that $\beta \geq 0$.

We also wish to determine the LD50, the dose level at which the probability of death is 50%. Thus, we determine x so that

$$\text{logit}(0.5) = \alpha + \beta x.$$

Solving this gives that the LD50 is $x = -\beta/\alpha$.

Figure 12 shows the results of the analysis.

Binomial regression: Example (cont)

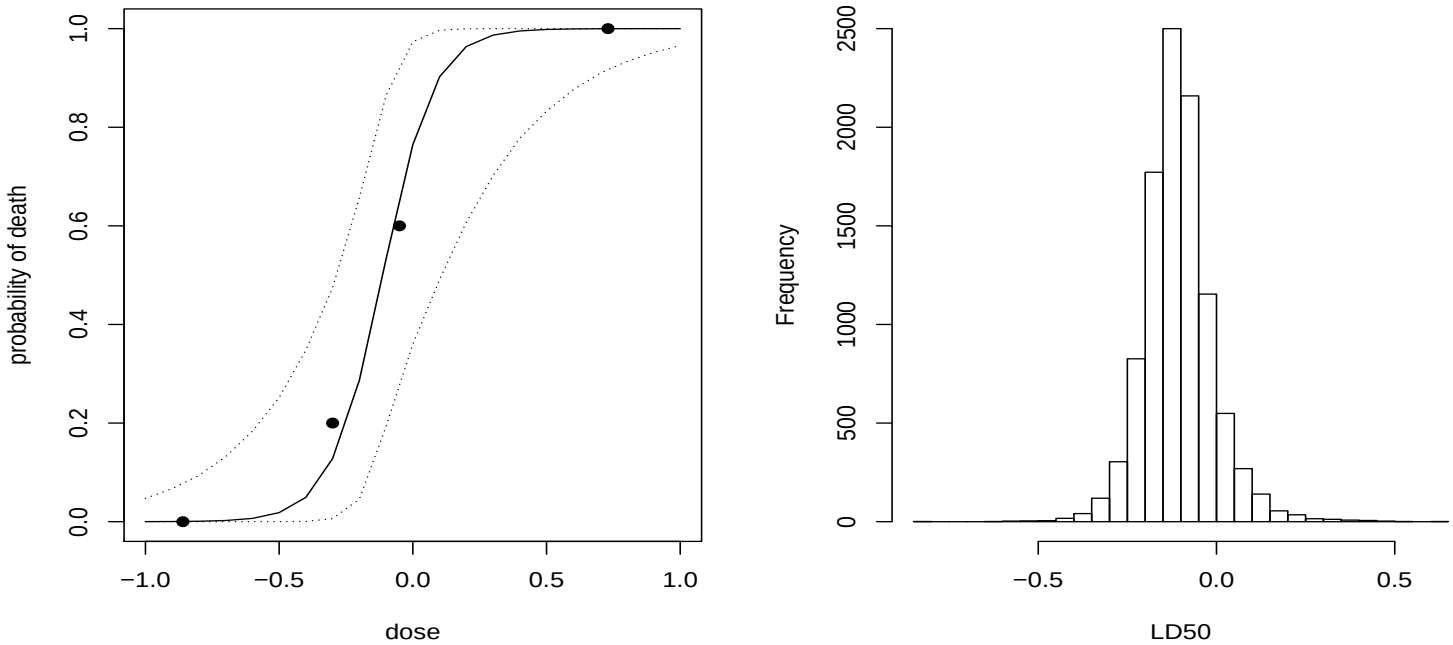


Figure 12: Results of the bioassay experiment. Left: Probability of death as a function of dose with 95% posterior interval and 'observed values'. Right: The posterior distribution of LD50.

Binomial regression: Example (cont)


```
alpha ~ dnorm(0,0.001)
beta ~ dnorm(0,0.001)T(0,) #Truncated distribution
for(i in 1:k){
  logit(theta[i]) <- alpha+beta*x[i]
  y[i] ~ dbinom(theta[i],n[i])
}
LD50 <- -alpha/beta
for(i in 1:K){
  logit(theta.pred[i]) <- alpha+beta*xpred[i]}

#R code:
bioassay.coda <- coda.samples(bioassay.jag,c("alpha","beta",
                                  "LD50","theta.pred"),10000)
a <- summary(bioassay.coda)
med <- a$quantiles[-(1:3),3]
plot(xpred,med,type="l",ylim=c(0,1),xlab="dose",ylab="probabi
points(x,y/n,pch=19)
```

Literature

- ✔ Congdon, Peter, *Bayesian Statistical Modelling*, Wiley, 2001
- ✔ Davison, A. C. *Statistical Models*, Cambridge University Press, 2003
- ✔ Gelman et al, *Bayesian Data Analysis*, Chapman & Hall/CRC, 2nd edition, 2004
- ✔ Kruschke, John K, Doing Bayesian Data Analysis: A Tutorial with R and BUGS, Elsevier, 2011
- ✔ Lee, Peter M., *Bayesian Statistics: An Introduction*, Wiley, 4th edition, 2012.
- ✔ Lunn et al, The BUGS Book: A Practical Introduction to Bayesian Analysis, CRC Press, 2013
- ✔ Rohatgi, Vijay K. *Statistical Inference*, Dover Publications, 2003
- ✔ Ross, Sheldon M, *Introduction to Probability Models*, Academic Press, 6th (or newer) edition, 1997