Bayesian Inference I

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

MSc in Biostatistics,
Department of Mathematics and School of Medicine

Outline of the course

This course provides theory and practice of the Bayesian approach to statistical inference. Applications are performed with the statistical package R.

Topics:

- Bayesian Updating through Bayes' Theorem
- Prior Distributions
- Multi-parameter Problems
- Summarizing Posterior Information
- Prediction
- The Gibbs Sampler

Multi-parameter problems

Most statistical models contain more than one parameter. The method of analysing multi-parameter problems in Bayesian statistics is much more straightforward than in classical statistics. Indeed, there is absolutely no new theory required.

We now have a vector $\theta = (\theta_1, \ldots, \theta_d)$ of parameters. We specify a multivariate prior $f(\theta)$, and combine it with a likelihood $f(x|\theta)$ via Bayes' theorem to obtain

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)d\theta}.$$

Of course, the posterior will now also be a multivariate distribution and inference about any subset of parameters within θ is obtained by straightforward probability calculations on this joint distribution.

Conditional Posterior Distributions

The *conditional posterior distribution* of a component of θ, θ_i say, given the values of the remaining components θ_{-i} is given by

$$f_i(\theta_i \mid x, \theta_{-i}) \propto f(\theta \mid x),$$

where the values of θ_{-i} are held fixed.

Conditional Posterior Distributions

The *conditional posterior distribution* of a component of θ, θ_i say, given the values of the remaining components θ_{-i} is given by

$$f_i(\theta_i \mid x, \theta_{-i}) \propto f(\theta \mid x),$$

where the values of θ_{-i} are held fixed.

That is the conditional posterior distribution of θ_i is given by the joint posterior distribution of θ, $f(\theta \mid \mathbf{x})$, regarded as a function of θ_i alone with the other components θ_{-i} of θ fixed, normalised to be a density function as appropriate.

Marginal Posterior Distributions

Exact Bayesian inference about the scalar parameter θ_i can only be made from the posterior distribution integrated over θ_{-i},

$$f(\theta_i \mid \mathbf{x}) = \int f(\theta \mid \mathbf{x}) d\theta_{-i}.$$

This resulting *marginal posterior* of a given parameter of interest θ_i, after eliminating the nuisance parameters θ_{-i} by integration, can be used for drawing inferences about that parameter.

Marginal Posterior Distributions

Exact Bayesian inference about the scalar parameter θ_i can only be made from the posterior distribution integrated over θ_{-i},

$$f(\theta_i \mid \mathbf{x}) = \int f(\theta \mid \mathbf{x}) d\theta_{-i}.$$

This resulting *marginal posterior* of a given parameter of interest θ_i, after eliminating the nuisance parameters θ_{-i} by integration, can be used for drawing inferences about that parameter.

If marginalization is not possible, another approach which can be used to eliminate the nuisance parameters is to compute the posterior distribution of the parameter of interest conditioning on the maximum likelihood estimates of the other components of the parameter vector. This technique, which is not fully Bayesian, is called the *empirical Bayes* method, to be distinguished from fully Bayesian inferential methods.

Practical Issues

1. **Prior specification.** Priors are now multivariate distributions. This means that the prior specification needs to reflect prior belief not just about each parameter individually, but also about dependence between different parameters.
2. **Computation.** With multivariate problems the integrals are very difficult to evaluate. This makes the use of conjugate prior families even more valuable, and creates the need for numerical techniques to obtain inferences when conjugate families are either unavailable or inappropriate.
3. **Interpretation.** The entire posterior inference is contained in the posterior distribution, which will have as many dimensions as the variable θ. The structure of the posterior distribution may be highly complex.

Multivariate Prior Distributions

Consider a statistical problem parameterised by $\theta = (\theta_1, \theta_2)$.

- The simplest choice of prior is to assume prior independence between θ_1 and θ_2: $f(\theta) = f(\theta_1)f(\theta_2)$

Multivariate Prior Distributions

Consider a statistical problem parameterised by $\theta = (\theta_1, \theta_2)$.

- The simplest choice of prior is to assume prior independence between θ_1 and θ_2: $f(\theta) = f(\theta_1)f(\theta_2)$
- Another choice can be a bivariate distribution factorised as a product of a conditional times a marginal density:

$$f(\theta) = f(\theta_1 \mid \theta_2)f(\theta_2)$$

Multivariate Prior Distributions

Consider a statistical problem parameterised by $\theta = (\theta_1, \theta_2)$.

- The simplest choice of prior is to assume prior independence between θ_1 and θ_2: $f(\theta) = f(\theta_1)f(\theta_2)$
- Another choice can be a bivariate distribution factorised as a product of a conditional times a marginal density:

$$f(\theta) = f(\theta_1 \mid \theta_2)f(\theta_2)$$

- The most general case is to assume a bivariate prior distribution allowing for dependence (correlation) between θ_1 and θ_2, for example a bivariate normal density.

Note: Generalisation to the case of a multivariate parameter θ.

A Discrete Example

Suppose a machine is either satisfactory ($x = 1$) or unsatisfactory ($x = 2$). The probability of the machine being satisfactory depends on the room temperature ($\theta_1 = 0$: cool, $\theta_1 = 1$: hot) and humidity ($\theta_2 = 0$: dry, $\theta_2 = 1$: humid). The probabilities of $x = 1$ are given in the following table.

$\Pr(x = 1 \mid \theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.6	0.8
$\theta_2 = 1$	0.7	0.6

The joint prior distribution of (θ_1, θ_2) is

$\Pr(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0.3	0.2
$\theta_2 = 1$	0.2	0.3

The Posterior Distribution

The joint posterior distribution can be calculated as follows.

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x = 1 \mid \theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	0.18	0.16
$= \Pr(x = 1, \theta_1, \theta_2)$	$\theta_2 = 1$	0.14	0.18

The Posterior Distribution

The joint posterior distribution can be calculated as follows.

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x=1\|\theta_1,\theta_2)\times\Pr(\theta_1,\theta_2)$	$\theta_2 = 0$	0.18	0.16
$=\Pr(x=1,\theta_1,\theta_2)$	$\theta_2 = 1$	0.14	0.18
$\Pr(x=1)$		0.66	

The Posterior Distribution

The joint posterior distribution can be calculated as follows.

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x = 1\|\theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	0.18	0.16
$= \Pr(x = 1, \theta_1, \theta_2)$	$\theta_2 = 1$	0.14	0.18
$\Pr(x = 1)$		0.66	
$\Pr(\theta_1, \theta_2\|x = 1)$	$\theta_2 = 0$	18/66	16/66
	$\theta_2 = 1$	14/66	18/66

The Posterior Distribution

The joint posterior distribution can be calculated as follows.

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x = 1 \mid \theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	0.18	0.16
$= \Pr(x = 1, \theta_1, \theta_2)$	$\theta_2 = 1$	0.14	0.18
$\Pr(x = 1)$		0.66	
$\Pr(\theta_1, \theta_2 \mid x = 1)$	$\theta_2 = 0$	18/66	16/66
	$\theta_2 = 1$	14/66	18/66

By summing across margins we obtain the marginal posterior distributions:

$$\Pr(\theta_1 = 0) = 32/66, \quad \Pr(\theta_1 = 1) = 34/66$$

and

$$\Pr(\theta_2 = 0) = 34/66, \quad \Pr(\theta_2 = 1) = 32/66.$$

A Continuous Example

Suppose $Y_1 \sim Poisson(\alpha\beta)$ and $Y_2 \sim Poisson(1-\alpha)\beta)$ with Y_1 and Y_2 independent given α and β.

Suppose our prior information for α and β can be expressed as: $\alpha \sim Beta(p, q)$ and $\beta \sim Gamma(p+q, 1)$ with α and β independent, for specified hyperparameters p and q.

Then we have the following likelihood:

$$f(y_1, y_2|\alpha, \beta) = \frac{\exp(-\alpha\beta)(\alpha\beta)^{y_1}}{y_1!} \times \frac{\exp(-(1-\alpha)\beta)[(1-\alpha)\beta]^{y_2}}{y_2!}$$

and the prior

$$f(\alpha, \beta) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\alpha^{p-1}(1-\alpha)^{q-1} \times \frac{e^{-\beta}\beta^{p+q-1}}{\Gamma(p+q)}.$$

The Joint Posterior

By Bayes' theorem:

$$\begin{aligned} f(\alpha, \beta | y_1, y_2) &\propto e^{-\beta} \beta^{y_1+y_2} \alpha^{y_1} (1-\alpha)^{y_2} \alpha^{p-1} (1-\alpha)^{q-1} e^{-\beta} \beta^{p+q-1} \\ &= \beta^{y_1+y_2+p+q-1} e^{-2\beta} \alpha^{y_1+p-1} (1-\alpha)^{y_2+q-1} \end{aligned}$$

over the region $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq \infty$. This is the (joint) posterior distribution for α and β and contains all the information from the prior and data.

The Joint Posterior

By Bayes' theorem:

$$\begin{aligned} f(\alpha, \beta | y_1, y_2) &\propto e^{-\beta} \beta^{y_1+y_2} \alpha^{y_1} (1-\alpha)^{y_2} \alpha^{p-1} (1-\alpha)^{q-1} e^{-\beta} \beta^{p+q-1} \\ &= \beta^{y_1+y_2+p+q-1} e^{-2\beta} \alpha^{y_1+p-1} (1-\alpha)^{y_2+q-1} \end{aligned}$$

over the region $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq \infty$. This is the (joint) posterior distribution for α and β and contains all the information from the prior and data.

In this particular case, the posterior factorises into functions of α and β. Therefore, we can write:

$$f(\alpha, \beta \mid y_1, y_2) \propto g(\alpha) h(\beta), \quad \text{where}$$

$$g(\alpha) = \alpha^{y_1+p-1}(1-\alpha)^{y_2+q-1} \quad \text{and} \quad h(\beta) = \beta^{y_1+y_2+p+q-1} e^{-2\beta}.$$

The Marginals

It follows, therefore, that the marginal posterior distributions are given by

$$f(\alpha|y_1, y_2) = \int_0^\infty f(\alpha, \beta|y_1, y_2) d\beta \ \propto \ g(\alpha) \int_0^\infty h(\beta) d\beta \propto \ g(\alpha),$$

and

$$f(\beta|y_1, y_2) = \int_0^1 f(\alpha, \beta|y_1, y_2) d\alpha \ \propto \ h(\beta) \int_0^1 g(\alpha) d\alpha \ \propto \ h(\beta).$$

The Marginals

It follows, therefore, that the marginal posterior distributions are given by

$$f(\alpha|y_1, y_2) = \int_0^\infty f(\alpha, \beta|y_1, y_2) d\beta \;\propto\; g(\alpha) \int_0^\infty h(\beta) d\beta \propto\; g(\alpha),$$

and

$$f(\beta|y_1, y_2) = \int_0^1 f(\alpha, \beta|y_1, y_2) d\alpha \;\propto\; h(\beta) \int_0^1 g(\alpha) d\alpha \;\propto\; h(\beta).$$

That is, $\alpha|y_1, y_2 \sim Beta(y_1 + p, y_2 + q)$ and $\beta|y_1, y_2 \sim Gamma(y_1 + y_2 + p + q, 2)$.

The Marginals

It follows, therefore, that the marginal posterior distributions are given by

$$f(\alpha|y_1, y_2) = \int_0^\infty f(\alpha, \beta|y_1, y_2)d\beta \ \propto \ g(\alpha)\int_0^\infty h(\beta)d\beta \propto \ g(\alpha),$$

and

$$f(\beta|y_1, y_2) = \int_0^1 f(\alpha, \beta|y_1, y_2)d\alpha \ \propto \ h(\beta)\int_0^1 g(\alpha)d\alpha \ \propto \ h(\beta).$$

That is, $\alpha|y_1, y_2 \sim Beta(y_1 + p, y_2 + q)$ and $\beta|y_1, y_2 \sim Gamma(y_1 + y_2 + p + q, 2)$.

Note: The posterior belongs to the same family with the prior, therefore the prior we chose was conjugate to this likelihood model.

Outline of the course

This course provides theory and practice of the Bayesian approach to statistical inference. Applications are performed with the statistical package R.

Topics:

- Bayesian Updating through Bayes' Theorem
- Prior Distributions
- Multi-parameter Problems
- Summarizing Posterior Information
- Prediction
- The Gibbs Sampler

Summarizing posterior information

We've stressed that the posterior distribution is a complete summary of the inference about a parameter θ. In essence, the posterior distribution is the inference. However, for some applications it is desirable to summarize this information.

- Point Estimation. Point summaries of the posterior distribution obtained within a decision theoretic framework. Common choices: posterior mean, median or mode.

Summarizing posterior information

We've stressed that the posterior distribution is a complete summary of the inference about a parameter θ. In essence, the posterior distribution is the inference. However, for some applications it is desirable to summarize this information.

- Point Estimation. Point summaries of the posterior distribution obtained within a decision theoretic framework. Common choices: posterior mean, median or mode.
- Credibility Regions. Analogue of a classical confidence interval. Point estimates give no measure of accuracy, so it is preferable to give a region within which it is 'likely' that the parameter lies. Bayesian credibility regions are probabilistic.

Summarizing posterior information

We've stressed that the posterior distribution is a complete summary of the inference about a parameter θ. In essence, the posterior distribution is the inference. However, for some applications it is desirable to summarize this information.

- Point Estimation. Point summaries of the posterior distribution obtained within a decision theoretic framework. Common choices: posterior mean, median or mode.
- Credibility Regions. Analogue of a classical confidence interval. Point estimates give no measure of accuracy, so it is preferable to give a region within which it is 'likely' that the parameter lies. Bayesian credibility regions are probabilistic.
- Hypothesis Testing. Comparisons of two (or more) alternative hypotheses, e.g $H_0: \theta \in \Omega_0$, $H_1: \theta \in \Omega_1$. Probabilistic statements about and symmetric treatment of the hypotheses.

Decision Theory

Many problems in the real world are those of making decisions in the face of uncertainty: '*which political party will be best to vote for?*'; '*should I accept one job offer or wait in the hope that I get offered a better job?*'.

Decision Theory

Many problems in the real world are those of making decisions in the face of uncertainty: '*which political party will be best to vote for?*'; '*should I accept one job offer or wait in the hope that I get offered a better job?*'.

All of statistical inference can also be thought of as decision making: *having observed a particular set of data, what value should we decide to estimate a parameter by?*

Decision Theory

Many problems in the real world are those of making decisions in the face of uncertainty: '*which political party will be best to vote for?*'; '*should I accept one job offer or wait in the hope that I get offered a better job?*'.

All of statistical inference can also be thought of as decision making: *having observed a particular set of data, what value should we decide to estimate a parameter by?*

The elements needed to construct a decision problem are:

1. A *parameter space* Θ which contains the possible *states of nature*;
2. A set A of *actions* which are available to the decision maker;
3. A *loss function* L, where $L(\theta, a)$ is the loss incurred by adopting action a when the true state of nature is θ.

An Example

A public health officer is seeking a rational policy of vaccination against a relatively mild ailment which causes absence from work.

Surveys suggest that 60% of the population are already immune.

It is estimated that the money–equivalent of man–hours lost from failing to vaccinate a vulnerable individual is 20, that the unnecessary cost of vaccinating an immune person is 8, and that there is no cost incurred in vaccinating a vulnerable person or failing to vaccinate an immune person.

So, for this example we have:

1. The parameter space $\Theta = \{\theta_1, \theta_2\}$, where θ_1 and θ_2 correspond to the individual being immune and vulnerable respectively;
2. The set of actions $A = \{a_1, a_2\}$ where a_1 and a_2 correspond to vaccinating and not vaccinating respectively;
3. The loss function is

$L(\theta, a)$	θ_1	θ_2
a_1	8	0
a_2	0	20

So, for this example we have:

1. The parameter space $\Theta = \{\theta_1, \theta_2\}$, where θ_1 and θ_2 correspond to the individual being immune and vulnerable respectively;
2. The set of actions $A = \{a_1, a_2\}$ where a_1 and a_2 correspond to vaccinating and not vaccinating respectively;
3. The loss function is

$L(\theta, a)$	θ_1	θ_2
a_1	8	0
a_2	0	20

The **decision strategy** is then to evaluate the expected loss for each action and choose the action which has the minimum expected loss.

Minimising the Prior Expected Loss

The expected loss is calculated based on the prior distribution of θ.

Minimising the Prior Expected Loss

The expected loss is calculated based on the prior distribution of θ.

Surveys suggest that 60% of the population are already immune. Therefore, $f(\theta_1) = 0.6$ and $f(\theta_2) = 0.4$.

Minimising the Prior Expected Loss

The expected loss is calculated based on the prior distribution of θ.

Surveys suggest that 60% of the population are already immune. Therefore, $f(\theta_1) = 0.6$ and $f(\theta_2) = 0.4$.

$f(\theta)$	0.6	0.4	
$L(\theta, a)$	θ_1	θ_2	$E[L(\theta, a)]$
a_1	8	0	$0.6 \times 8 + 0.4 \times 0 = 4.8$
a_2	0	20	$0.6 \times 0 + 0.4 \times 20 = 8.0$

Minimising the Prior Expected Loss

The expected loss is calculated based on the prior distribution of θ.

Surveys suggest that 60% of the population are already immune. Therefore, $f(\theta_1) = 0.6$ and $f(\theta_2) = 0.4$.

$f(\theta)$	0.6	0.4	
$L(\theta, a)$	θ_1	θ_2	$E[L(\theta, a)]$
a_1	8	0	$0.6 \times 8 + 0.4 \times 0 = 4.8$
a_2	0	20	$0.6 \times 0 + 0.4 \times 20 = 8.0$

The conclusion is that it is preferable (according to minimisation of cost) to vaccinate everyone. The cost (or loss) is 4.8 per individual.

Example: Continuation

Suppose now that we had further information or data x available to us which reflected the value of θ, i.e. we have observed x from $f(x|\theta)$. Then we can replace $f(\theta)$ by the posterior $f(\theta|x)$ in the calculation of the expected loss. The best action will then depend on the particular outcome x.

Example: Continuation

Suppose now that we had further information or data x available to us which reflected the value of θ, i.e. we have observed x from $f(x|\theta)$. Then we can replace $f(\theta)$ by the posterior $f(\theta|x)$ in the calculation of the expected loss. The best action will then depend on the particular outcome x.

A simple skin test has been developed which, though not completely reliable, tends to indicate the immune status of the individual. The probabilities of reaction are given below.

			Immune θ_1	Vulnerable θ_2
Reaction	Negligible	x_1	0.35	0.09
	Mild	x_2	0.30	0.17
	Moderate	x_3	0.21	0.25
	Strong	x_4	0.14	0.49

Posterior Expected Loss

Our general procedure is to use Bayes' theorem to compute the posterior distribution $f(\theta|x)$. Then, for any particular action a, the *posterior expected loss* is

$$\rho(a, x) = E\left[L(\theta, a)|x\right] = \int L(\theta, a)f(\theta|x)d\theta.$$

Having observed a particular value of x, we choose the action a which results in the lowest value of ρ. Writing $a = d(x)$, we call $d(x)$ the *Bayes decision rule*.

Posterior Expected Loss

Our general procedure is to use Bayes' theorem to compute the posterior distribution $f(\theta|x)$. Then, for any particular action a, the *posterior expected loss* is

$$\rho(a, x) = E\left[L(\theta, a)|x\right] = \int L(\theta, a) f(\theta|x) d\theta.$$

Having observed a particular value of x, we choose the action a which results in the lowest value of ρ. Writing $a = d(x)$, we call $d(x)$ the *Bayes decision rule*.

For our example, we consider all the possible outcomes x, calculating for each of these the corresponding posterior $f(\theta|x)$. For each of these we next work out the posterior expected loss for each action. Finally we select the best action, that with the minimum posterior expected loss, for that outcome.

		θ_1	θ_2	
	$f(x_1\|\theta)$	0.35	0.09	
Likelihoods	$f(x_2\|\theta)$	0.30	0.17	
	$f(x_3\|\theta)$	0.21	0.25	
	$f(x_4\|\theta)$	0.14	0.49	

		θ_1	θ_2	
Likelihoods	$f(x_1\|\theta)$	0.35	0.09	
	$f(x_2\|\theta)$	0.30	0.17	
	$f(x_3\|\theta)$	0.21	0.25	
	$f(x_4\|\theta)$	0.14	0.49	
Prior	$f(\theta)$	0.6	0.4	

		θ_1	θ_2		
Likelihoods	$f(x_1\|\theta)$	0.35	0.09		
	$f(x_2\|\theta)$	0.30	0.17		
	$f(x_3\|\theta)$	0.21	0.25		
	$f(x_4\|\theta)$	0.14	0.49		
Prior	$f(\theta)$	0.6	0.4		
Joints	$f(x_1, \theta)$	0.210	0.036	0.246	$f(x_1)$
	$f(x_2, \theta)$	0.180	0.068	0.248	$f(x_2)$
	$f(x_3, \theta)$	0.126	0.100	0.226	$f(x_3)$
	$f(x_4, \theta)$	0.084	0.196	0.280	$f(x_4)$

		θ_1	θ_2		
Likelihoods	$f(x_1\|\theta)$	0.35	0.09		
	$f(x_2\|\theta)$	0.30	0.17		
	$f(x_3\|\theta)$	0.21	0.25		
	$f(x_4\|\theta)$	0.14	0.49		
Prior	$f(\theta)$	0.6	0.4		
Joints	$f(x_1, \theta)$	0.210	0.036	0.246	$f(x_1)$
	$f(x_2, \theta)$	0.180	0.068	0.248	$f(x_2)$
	$f(x_3, \theta)$	0.126	0.100	0.226	$f(x_3)$
	$f(x_4, \theta)$	0.084	0.196	0.280	$f(x_4)$
				a_1	a_2
Posteriors	$f(\theta\|x_1)$	0.854	0.146	6.829	2.927
	$f(\theta\|x_2)$	0.726	0.274	5.806	5.484
	$f(\theta\|x_3)$	0.558	0.442	4.460	8.847
	$f(\theta\|x_4)$	0.300	0.700	2.400	14.000

Expected Losses

Bayes Decision Rule

The decisions for each value of x, together with their associated minimum posterior expected loss, are summarised below.

x	$d(x)$	$\rho(d(x), x)$
x_1	a_2	2.927
x_2	a_2	5.484
x_3	a_1	4.460
x_4	a_1	2.400

Bayes Decision Rule

The decisions for each value of x, together with their associated minimum posterior expected loss, are summarised below.

x	$d(x)$	$\rho(d(x), x)$
x_1	a_2	2.927
x_2	a_2	5.484
x_3	a_1	4.460
x_4	a_1	2.400

Conclusion: if either a negligible or mild reaction is observed, the Bayes decision is not to vaccinate, whereas if a moderate or strong reaction is observed, the decision is to vaccinate.

Bayes Risk

We can go one stage further and calculate the *risk* associated with this policy, by averaging across the uncertainty in the observations x. That is, we define the Bayes risk by:

$$BR(d) = \int \rho(d(x), x) f(x) dx$$

Bayes Risk

We can go one stage further and calculate the *risk* associated with this policy, by averaging across the uncertainty in the observations x. That is, we define the Bayes risk by:

$$BR(d) = \int \rho(d(x), x) f(x) dx$$

For our example this becomes the sum

$$BR(d) = \sum \rho(d(x), x) f(x) = 2.93 \times 0.25 + 5.48 \times 0.25 + 4.46 \times 0.23 + 2.40 \times 0.28$$

Bayes Risk

We can go one stage further and calculate the *risk* associated with this policy, by averaging across the uncertainty in the observations x. That is, we define the Bayes risk by:

$$BR(d) = \int \rho(d(x), x) f(x) dx$$

For our example this becomes the sum

$$BR(d) = \sum \rho(d(x), x) f(x) = 2.93 \times 0.25 + 5.48 \times 0.25 + 4.46 \times 0.23 + 2.40 \times 0.28$$

That is $BR(d) = 3.76$, which is smaller than the least cost per individual, of 4.8, obtained by using the prior information alone, without the knowledge of x. Therefore, measuring x is worth while.

Point Estimation

Under the Bayesian approach, the posterior distribution *is* the inference. However, for some applications it is desirable (or necessary) to summarize this information in some way. In particular, we may wish to give a single 'best' estimate of the unknown parameter.

Point Estimation

Under the Bayesian approach, the posterior distribution *is* the inference. However, for some applications it is desirable (or necessary) to summarize this information in some way. In particular, we may wish to give a single 'best' estimate of the unknown parameter.

So, in the Bayesian framework, *how do we reduce the information in a posterior distribution to give a single 'best' estimate?*

Point Estimation

Under the Bayesian approach, the posterior distribution *is* the inference. However, for some applications it is desirable (or necessary) to summarize this information in some way. In particular, we may wish to give a single 'best' estimate of the unknown parameter.

So, in the Bayesian framework, *how do we reduce the information in a posterior distribution to give a single 'best' estimate?*
In fact, the answer depends on what we mean by 'best', and this in turn is specified by turning the problem into a decision problem.

Point Estimation

Under the Bayesian approach, the posterior distribution *is* the inference. However, for some applications it is desirable (or necessary) to summarize this information in some way. In particular, we may wish to give a single 'best' estimate of the unknown parameter.

So, in the Bayesian framework, *how do we reduce the information in a posterior distribution to give a single 'best' estimate?*
In fact, the answer depends on what we mean by 'best', and this in turn is specified by turning the problem into a decision problem.

We specify a loss function $L(\theta, a)$ which measures our perceived penalty in estimating θ by a. There are a range of natural loss functions we could use, and the particular choice for any specified problem will depend on the context.

Loss Functions

The most commonly used loss functions are:

1. *Squared Error (or Quadratic) loss*: $L(\theta, a) = (\theta - a)^2$;
2. *Absolute Error loss*: $L(\theta, a) = |\theta - a|$;
3. *0—1 loss*:

$$L(\theta, a) = \begin{cases} 0 & \text{if } |\theta - a| \leq \epsilon \\ 1 & \text{if } |\theta - a| > \epsilon \end{cases}$$

In each of these cases, by *minimizing the posterior expected loss*, we obtain simple forms for the Bayes decision rule, which is taken to be the **point estimate** of θ for that particular choice of loss function.

Squared Error Loss

In this case we can simplify $\rho(a, x) = E\left[(\theta - a)^2 | x\right]$ by letting $\mu = E(\theta|x)$ and expanding:

$$E\left[(\theta - a)^2 | x\right] = E\left\{[(\theta - \mu) + (\mu - a)]^2 \, | x\right\}$$

Squared Error Loss

In this case we can simplify $\rho(a,x) = E\left[(\theta - a)^2|x\right]$ by letting $\mu = E(\theta|x)$ and expanding:

$$\begin{aligned} E\left[(\theta - a)^2|x\right] &= E\left\{\left[(\theta - \mu) + (\mu - a)\right]^2|x\right\} \\ &= E\left[(\theta - \mu)^2|x\right] + (\mu - a)^2 + 2\,E\left[(\theta - \mu)|x\right](\mu - a) \end{aligned}$$

Squared Error Loss

In this case we can simplify $\rho(a,x) = E\left[(\theta - a)^2|x\right]$ by letting $\mu = E(\theta|x)$ and expanding:

$$\begin{aligned} E\left[(\theta - a)^2|x\right] &= E\left\{[(\theta-\mu)+(\mu-a)]^2\,|x\right\} \\ &= E\left[(\theta-\mu)^2|x\right] + (\mu-a)^2 + 2\,E\left[(\theta-\mu)|x\right](\mu-a) \\ &= Var\left[\theta|x\right] + (\mu-a)^2 \end{aligned}$$

Squared Error Loss

In this case we can simplify $\rho(a, x) = E\left[(\theta - a)^2|x\right]$ by letting $\mu = E(\theta|x)$ and expanding:

$$\begin{aligned} E\left[(\theta - a)^2|x\right] &= E\left\{\left[(\theta - \mu) + (\mu - a)\right]^2|x\right\} \\ &= E\left[(\theta - \mu)^2|x\right] + (\mu - a)^2 + 2\,E\left[(\theta - \mu)|x\right](\mu - a) \\ &= Var\left[\theta|x\right] + (\mu - a)^2 \end{aligned}$$

On the right, the first term no longer depends on a, and the second term attains its minimum of zero by taking $a = \mu$. In summary, the posterior expected squared error loss has its minimum value of $Var\left[\theta|x\right]$, the posterior variance of θ, when $a = E(\theta|x)$, the **posterior expectation** of θ.

Absolute Error Loss

We show that in this case the minimum posterior expected loss is obtained by taking $a = m$, the **median** of the posterior distribution $f(\theta|x)$. We assume that this is unique, and is defined by

$$\Pr(\theta < m|x) = \Pr(\theta > m|x) = 1/2.$$

Absolute Error Loss

We show that in this case the minimum posterior expected loss is obtained by taking $a = m$, the **median** of the posterior distribution $f(\theta|x)$. We assume that this is unique, and is defined by

$$\Pr(\theta < m|x) = \Pr(\theta > m|x) = 1/2.$$

To prove the result note first that the function

$$s(\theta) = \begin{cases} -1, & \text{for } \theta < m \\ +1, & \text{for } \theta > m \end{cases}$$

has the property

$$\begin{aligned} E\left[s(\theta) \mid x\right] &= -\int_{-\infty}^{m} f(\theta \mid x)d\theta + \int_{m}^{\infty} f(\theta \mid x)d\theta \\ &= -\Pr(\theta < m \mid x) + \Pr(\theta > m \mid x) = 0. \end{aligned}$$

Absolute Error Loss

Now consider $L(\theta, a) - L(\theta, m) = |\theta - a| - |\theta - m|$ for some $a < m$.

Absolute Error Loss

Now consider $L(\theta, a) - L(\theta, m) = |\theta - a| - |\theta - m|$ for some $a < m$.

If $\theta < a$:
$L(\theta, a) - L(\theta, m) = -\theta + a + \theta - m = a - m = (m - a)s(\theta)$

Absolute Error Loss

Now consider $L(\theta, a) - L(\theta, m) = |\theta - a| - |\theta - m|$ for some $a < m$.

If $\theta < a$:
$L(\theta, a) - L(\theta, m) = -\theta + a + \theta - m = a - m = (m - a)s(\theta)$

If $\theta > m$:
$L(\theta, a) - L(\theta, m) = -a + \theta - \theta + m = -a + m = (m - a)s(\theta)$

Absolute Error Loss

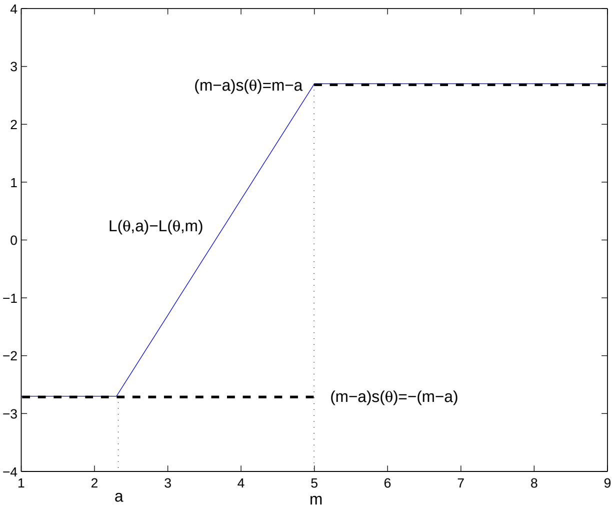
Now consider $L(\theta, a) - L(\theta, m) = |\theta - a| - |\theta - m|$ for some $a < m$.

If $\theta < a$:
$L(\theta, a) - L(\theta, m) = -\theta + a + \theta - m = a - m = (m - a)s(\theta)$

If $\theta > m$:
$L(\theta, a) - L(\theta, m) = -a + \theta - \theta + m = -a + m = (m - a)s(\theta)$

If $a < \theta < m$:
$L(\theta, a) - L(\theta, m) = -a + \theta + \theta - m = 2\theta - a - m > (m - a)s(\theta)$

Plot of $L(\theta, a) - L(\theta, m)$ and $(m - a)s(\theta)$



Absolute Error Loss

It can be seen that $L(\theta, a) - L(\theta, m)$ is greater than $(m - a)s(\theta)$ so

$$E\left[L(\theta, a) - L(\theta, m)|x\right] > (m - a)E\left[s(\theta)|x\right] = 0.$$

Absolute Error Loss

It can be seen that $L(\theta, a) - L(\theta, m)$ is greater than $(m - a)s(\theta)$ so

$$E\left[L(\theta, a) - L(\theta, m)|x\right] > (m - a)E\left[s(\theta)|x\right] = 0.$$

So

$$E\left[L(\theta, a)|x\right] > E\left[L(\theta, m)|x\right].$$

Absolute Error Loss

It can be seen that $L(\theta, a) - L(\theta, m)$ is greater than $(m - a)s(\theta)$ so

$$E\left[L(\theta, a) - L(\theta, m)|x\right] > (m - a)E\left[s(\theta)|x\right] = 0.$$

So

$$E\left[L(\theta, a)|x\right] > E\left[L(\theta, m)|x\right].$$

This also holds by a similar argument when $a > m$, so $E\left[L(\theta, a)|x\right]$ is a minimum when $a = m$, the **posterior median**.

0–1 Loss

Clearly in this case

$$\rho(a, x) = \Pr\{|\theta - a| > \epsilon | x) = 1 - \Pr\{|\theta - a| \leq \epsilon | x).$$

Consequently, if we define a *modal interval of length* 2ϵ as the interval $[\theta - \epsilon, \theta + \epsilon]$ which has highest probability, then the Bayes estimate is the **midpoint** of the interval with highest probability.

By choosing ϵ arbitrarily small, this procedure will lead to the **posterior mode** as the Bayesian estimate.

0–1 Loss

Clearly in this case

$$\rho(a,x) = \Pr\{|\theta - a| > \epsilon|x) = 1 - \Pr\{|\theta - a| \leq \epsilon|x).$$

Consequently, if we define a *modal interval of length* 2ϵ as the interval $[\theta - \epsilon, \theta + \epsilon]$ which has highest probability, then the Bayes estimate is the **midpoint** of the interval with highest probability.

By choosing ϵ arbitrarily small, this procedure will lead to the **posterior mode** as the Bayesian estimate.

Conclusion: in the Bayesian framework a point estimate is a single summary statistic of the posterior distribution. By defining the quality of an estimator through a loss function, the decision theory methodology leads to optimal choices of point estimates.

Example

If the posterior density for θ is

$$f(\theta|x) = 1 \text{ for } 0 \leq \theta \leq 1,$$

calculate the best estimator of $\phi = \theta^2$ with respect to quadratic loss.

Example

If the posterior density for θ is

$$f(\theta|x) = 1 \text{ for } 0 \leq \theta \leq 1,$$

calculate the best estimator of $\phi = \theta^2$ with respect to quadratic loss.

The best estimator of ϕ with respect to quadratic loss is

$$E(\phi \mid x) = E(\theta^2 \mid x) = \int_0^1 \theta^2 d\theta = \left[\frac{\theta^3}{3}\right]_0^1 = \frac{1}{3}.$$

Credibility Regions

In classical statistics parameters are not regarded as random, so it is not possible to give an interval with the interpretation that there is a certain probability that the parameter lies in the interval. Instead, confidence intervals have the interpretation that if the sampling were repeated, there is a specified probability that the interval so obtained would contain the parameter (it is the interval which is random and not the parameter).

There is no such difficulty in the Bayesian approach because parameters are treated as random. Thus, a region $C_\alpha(x)$ is a $100(1-\alpha)\%$ *credible region* for θ if

$$\int_{C_\alpha(x)} f(\theta|x)d\theta = 1-\alpha.$$

That is, there is a probability of $1-\alpha$, based on the posterior distribution, that θ lies in $C_\alpha(x)$.

Highest Posterior Density Credibility Regions

One difficulty with credibility regions (in common with confidence intervals) is that they are not uniquely defined. Any region with probability $1-\alpha$ will do. Since we want the region to contain the 'most probable' values of the parameter, it is usual to impose an additional constraint:

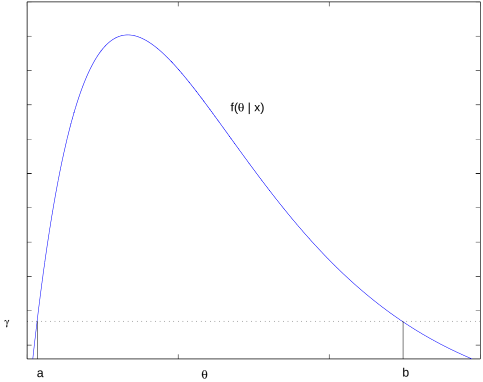
$$C_\alpha(x) = \{\theta : \ f(\theta|x) \geq \gamma\}$$

where γ is chosen to ensure that

$$\int_{C_\alpha(x)} f(\theta|x)d\theta = 1-\alpha.$$

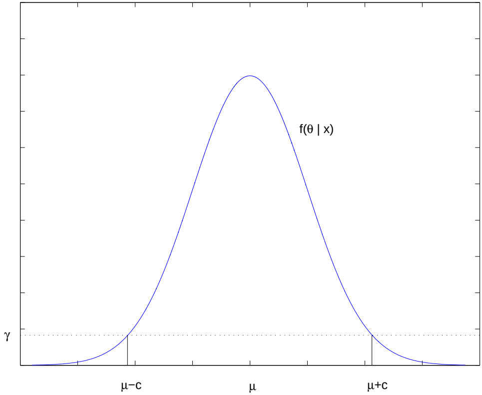
Unimodal Posterior Distribution

HPD region for a unimodal posterior distribution. The region is an interval of the form (a, b).



Symmetric Unimodal Posterior Distribution

HPD region for a unimodal and symmetric posterior distribution. The region is an interval of the form $(\mu - c, \mu + c)$.



Example. Normal Mean

Let $X_1, \ldots, X_n$ be independent variables from $N(\theta, \sigma^2)$ (σ^2 known) with a prior for θ of the form $\theta \sim N(b, d^2)$.

With this construction we obtained the posterior:

$$\theta | x \sim N(\mu, s^2)$$

where $\mu = \frac{\frac{b}{d^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$ and $s^2 = \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$.

Example. Normal Mean

Let $X_1, \ldots, X_n$ be independent variables from $N(\theta, \sigma^2)$ (σ^2 known) with a prior for θ of the form $\theta \sim N(b, d^2)$.

With this construction we obtained the posterior:

$$\theta | x \sim N(\mu, s^2)$$

where $\mu = \frac{\frac{b}{d^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$ and $s^2 = \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$.

Since the normal distribution is uni–modal and symmetric, the HPD regions are symmetric intervals of the form $(\mu - c, \mu + c)$. It follows that the $100(1-\alpha)\%$ HPD interval for θ is:

$$\mu \pm z_{\alpha/2} s,$$

where $z_{\alpha/2}$ is the appropriate percentile of the $N(0,1)$ distribution.

Example. Normal Mean

Let $X_1, \ldots, X_n$ be independent variables from $N(\theta, \sigma^2)$ (σ^2 known) with a prior for θ of the form $\theta \sim N(b, d^2)$.

With this construction we obtained the posterior:

$$\theta | x \sim N(\mu, s^2)$$

where $\mu = \frac{\frac{b}{d^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$ and $s^2 = \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}$.

Since the normal distribution is uni–modal and symmetric, the HPD regions are symmetric intervals of the form $(\mu - c, \mu + c)$. It follows that the $100(1-\alpha)\%$ HPD interval for θ is:

$$\mu \pm z_{\alpha/2} s,$$

where $z_{\alpha/2}$ is the appropriate percentile of the $N(0, 1)$ distribution.

As $n \to \infty$ this interval becomes $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$.

Example

Suppose $x \sim Binomial(n, \theta)$ with the prior

$$\theta \sim Beta(p, q).$$

This gives the posterior distribution

$$\theta|x \sim Beta(p + x, q + n - x)$$

Example

Suppose $x \sim Binomial(n, \theta)$ with the prior

$$\theta \sim Beta(p, q).$$

This gives the posterior distribution

$$\theta|x \sim Beta(p + x, q + n - x)$$

Thus, the $100(1 - \alpha)\%$ HPD interval $[a, b]$ satisfies:

$$\frac{1}{B(p + x, q + n - x)} \int_a^b \theta^{p+x-1}(1 - \theta)^{q+n-x-1} d\theta = 1 - \alpha,$$

and

$$a^{p+x-1}(1 - a)^{q+n-x-1} = b^{p+x-1}(1 - b)^{q+n-x-1} = \gamma.$$

Generally, this has to be solved numerically.

Hypothesis Tests

Hypothesis tests are decisions of the form of choosing between two different hypotheses:

$$H_0: \ \theta \in \Omega_0,$$
$$H_1: \ \theta \in \Omega_1.$$

In the simplest case where Ω_1 and Ω_2 consist of single points, the test is of the form

$$H_0: \ \theta = \theta_0,$$
$$H_1: \ \theta = \theta_1.$$

Hypothesis Tests

The classical approach to this problem is usually to base the test on the *likelihood ratio*:

$$\lambda = \frac{f(x|\theta_1)}{f(x|\theta_0)}.$$

Large values of λ indicate that the observed data x is more likely to have occurred if θ_1 is the true value of θ than if θ_0 is.

Hypothesis Tests

The classical approach to this problem is usually to base the test on the *likelihood ratio*:

$$\lambda = \frac{f(x|\theta_1)}{f(x|\theta_0)}.$$

Large values of λ indicate that the observed data x is more likely to have occurred if θ_1 is the true value of θ than if θ_0 is.

In the Bayesian view of things, we should also bring to bear the prior information we have about θ. Therefore, we may compute the posterior probabilities of θ_1 and θ_0:

$$\begin{aligned} f(\theta_1|x) &= \frac{f(\theta_1)f(x|\theta_1)}{f(\theta_0)f(x|\theta_0) + f(\theta_1)f(x|\theta_1)} \\ f(\theta_0|x) &= 1 - f(\theta_1|x). \end{aligned}$$

Hypothesis Tests

In the general case of testing the hypotheses:

$$H_0: \ \theta \in \Omega_0,$$
$$H_1: \ \theta \in \Omega_1,$$

we can still calculate the posterior probabilities of the two hypotheses, after specifying prior probabilities, $f(\theta \in \Omega_0)$ and $f(\theta \in \Omega_1)$, on the hypotheses. Then we have

$$f(\theta \in \Omega_1|x) = \frac{f(\theta \in \Omega_1)f(x|\theta \in \Omega_1)}{f(\theta \in \Omega_0)f(x|\theta \in \Omega_0) + f(\theta \in \Omega_1)f(x|\theta \in \Omega_1)},$$

where

$$f(x|\theta \in \Omega) = \int_\Omega f(\theta)f(x|\theta)d\theta.$$

Obviously, it is straightforward to generalise the above testing approach to the case of testing more than two hypotheses.

Baysian Model Comparison

Bayesian model comparison is a generalisation of Bayesian hypothesis testing.

Baysian Model Comparison

Bayesian model comparison is a generalisation of Bayesian hypothesis testing.

Consider being interested in comparing k competing models for a given set of observed data: M_1, M_2,..., M_k.

Baysian Model Comparison

Bayesian model comparison is a generalisation of Bayesian hypothesis testing.

Consider being interested in comparing k competing models for a given set of observed data: M_1, M_2,..., M_k.

We assume prior model probabilities: $\Pr(M_j)$, $j = 1, \ldots, k$, $\sum_{j=1}^{k} \Pr(M_j) = 1$.

Baysian Model Comparison

Bayesian model comparison is a generalisation of Bayesian hypothesis testing.

Consider being interested in comparing k competing models for a given set of observed data: M_1, M_2,..., M_k.

We assume prior model probabilities: $\Pr(M_j)$, $j = 1, \ldots, k$, $\sum_{j=1}^{k} \Pr(M_j) = 1$.

We compute the posterior model probabilities as

$$\Pr(M_j \mid x) \propto \Pr(M_j) f(x \mid M_j)$$

Baysian Model Comparison

Bayesian model comparison is a generalisation of Bayesian hypothesis testing.

Consider being interested in comparing k competing models for a given set of observed data: M_1, M_2,..., M_k.

We assume prior model probabilities: $\Pr(M_j)$, $j = 1, \ldots, k$, $\sum_{j=1}^{k} \Pr(M_j) = 1$.

We compute the posterior model probabilities as

$$\Pr(M_j \mid x) \propto \Pr(M_j) f(x \mid M_j)$$

Before dealing with the problem of model comparison, let us define the marginal likelihood of a given model.

The Marginal Likelihood

The *marginal likelihood* or *evidence* $f(x)$ of a given model $f(x \mid \theta)$ is the **marginal distribution** of the data under that model. It is obtained by integrating the product of the likelihood times a prior distribution $f(\theta)$ on the model parameters θ over θ:

$$f(x) = \int f(x \mid \theta) f(\theta) d\theta.$$

That is $f(x)$ is the normalising constant of the posterior:

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}.$$

Equivalently, the marginal likelihood is defined as the expectation of the likelihood with respect to the prior distribution $f(\theta)$.

The Marginal Likelihood

The *marginal likelihood* or *evidence* $f(x)$ of a given model $f(x \mid \theta)$ is the **marginal distribution** of the data under that model. It is obtained by integrating the product of the likelihood times a prior distribution $f(\theta)$ on the model parameters θ over θ:

$$f(x) = \int f(x \mid \theta) f(\theta) d\theta.$$

That is $f(x)$ is the normalising constant of the posterior:

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}.$$

Equivalently, the marginal likelihood is defined as the expectation of the likelihood with respect to the prior distribution $f(\theta)$.

Note: for given data, x, $f(x)$ is the probability (or density) of observing x under the assumed model.

Bayesian Treatment

Consider a number of competing models $M_1, \dots, M_k$, parameterised respectively by $\theta_1, \dots, \theta_k$, for an observed data set. In the presence of uncertainty about the correct model, Bayesian inference involves:

1. Evaluation of the posterior probability $\Pr(M_j \mid x)$ of each model M_j, $j = 1, \dots, k$.
2. Evaluation of the posterior distribution $f(\theta_j \mid x, M_j)$ of the parameters θ_j of model M_j, $j = 1, \dots, k$.

Bayesian Treatment

Consider a number of competing models $M_1, \ldots, M_k$, parameterised respectively by $\theta_1, \ldots, \theta_k$, for an observed data set. In the presence of uncertainty about the correct model, Bayesian inference involves:

1. Evaluation of the posterior probability $\Pr(M_j \mid x)$ of each model M_j, $j = 1, \ldots, k$.
2. Evaluation of the posterior distribution $f(\theta_j \mid x, M_j)$ of the parameters θ_j of model M_j, $j = 1, \ldots, k$.

In fact, the unknown quantities in the process of statistical inference are **both the model and the parameters**. Under the Bayesian approach, all unknown quantities are treated as **random variables** and inferred through their posterior distributions.

Bayesian Inference

After specifying prior model probabilities, $\Pr(M_j)$, for all competing models and carefully choosing proper prior distributions for the model specific parameters, $f(\theta_j \mid M_j)$, $j = 1, \ldots, k$, posterior inferences are obtained as follows.

1. The posterior probability of model M_j is calculated using Bayes therom as

$$\Pr(M_j \mid x) = \frac{\Pr(M_j) f(x \mid M_j)}{\sum_{i=1}^{k} \Pr(M_i) f(x \mid M_i)}, \quad j = 1, \ldots, k,$$

 where $f(x \mid M_j)$ is the marginal likelihood of model M_j.

2. The posterior distribution of the parameters θ_j of model M_j is given by Bayes theorm as

$$f(\theta_j \mid x, M_j) = \frac{f(\theta_j \mid M_j) f(x \mid \theta_j, M_j)}{f(x \mid M_j)}, \quad j = 1, \ldots, k.$$

Proof

Consider the problem of joint inference for the model and the parameters. Let M be a *discrete r.v.* denoting the model and taking the values $M_1, \dots, M_k$. Let θ denote *generically* the parameter(s).

Proof

Consider the problem of joint inference for the model and the parameters. Let M be a *discrete r.v.* denoting the model and taking the values $M_1, \ldots, M_k$. Let θ denote *generically* the parameter(s).

Joint Prior: $f(\theta, M) = f(M)f(\theta \mid M)$

Proof

Consider the problem of joint inference for the model and the parameters. Let M be a *discrete r.v.* denoting the model and taking the values $M_1, \dots, M_k$. Let θ denote *generically* the parameter(s).

Joint Prior: $f(\theta, M) = f(M)f(\theta \mid M)$
Joint Posterior: $f(\theta, M \mid x) \propto f(M)f(\theta \mid M)f(x \mid \theta, M)$

Proof

Consider the problem of joint inference for the model and the parameters. Let M be a *discrete r.v.* denoting the model and taking the values $M_1, \ldots, M_k$. Let θ denote *generically* the parameter(s).

Joint Prior: $f(\theta, M) = f(M)f(\theta \mid M)$
Joint Posterior: $f(\theta, M \mid x) \propto f(M)f(\theta \mid M)f(x \mid \theta, M)$
Marginal Posterior of M:

$$
\begin{aligned}
f(M \mid x) &\propto \int f(M)f(\theta \mid M)f(x \mid \theta, M)d\theta \\
&= f(M)\int f(\theta \mid M)f(x \mid \theta, M)d\theta = f(M)f(x \mid M)
\end{aligned}
$$

Proof

Consider the problem of joint inference for the model and the parameters. Let M be a *discrete r.v.* denoting the model and taking the values $M_1, \ldots, M_k$. Let θ denote *generically* the parameter(s).

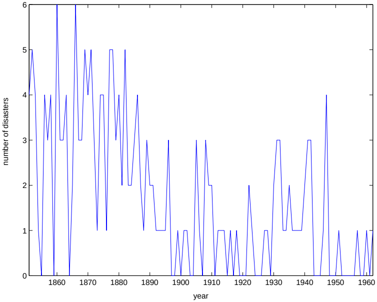
Joint Prior: $f(\theta, M) = f(M)f(\theta \mid M)$
Joint Posterior: $f(\theta, M \mid x) \propto f(M)f(\theta \mid M)f(x \mid \theta, M)$
Marginal Posterior of M:

$$
\begin{aligned}
f(M \mid x) &\propto \int f(M)f(\theta \mid M)f(x \mid \theta, M)d\theta \\
&= f(M) \int f(\theta \mid M)f(x \mid \theta, M)d\theta = f(M)f(x \mid M)
\end{aligned}
$$

Conditional Posterior of θ:
$f(\theta \mid x, M) \propto f(\theta, M \mid x) \propto f(\theta \mid M)f(x \mid \theta, M)$

Example. A Poisson Changepoint Problem

Consider data consisting of a series relating to the number of British coal mining disasters per year, over the period 1851 - 1962.



From this plot it does seem to be the case that there has been a reduction in the rate of disasters over the period.

Competing Models

For the coal-mining disasters data we consider two competing models:

M1 each x_i is an independent draw from a Poisson random variable with mean θ;

M2 for $i \leq t$, x_i is an independent draw from a Poisson random variable with mean θ_1, and for $i > t$, x_i is an independent draw from a Poisson random variable with mean θ_2.

In the first model there is just one unknown parameter, θ. In the second model, there are three unknown parameters: θ_1, θ_2 and t.

Prior Specification

Model M_1: $X_i \sim Poisson(\theta)$, $i = 1, \dots, n$
Model M_2: $X_i \sim Poisson(\theta_1)$, $i = 1, \dots, t$
$X_i \sim Poisson(\theta_2)$, $i = t+1, \dots, n$

We assume $\theta \sim Exp(1/2)$, i.e. $f(\theta) = \frac{1}{2} e^{-\theta/2}$

Furthermore, $\theta_1 \sim Exp(1/2)$, and $\theta_2 \sim Exp(1/2)$,
i.e. $f(\theta_1) = \frac{1}{2} e^{-\theta_1/2}$, $f(\theta_2) = \frac{1}{2} e^{-\theta_2/2}$,

$t \sim DU(1, \dots, n-1)$, i.e. $f(t) = \frac{1}{n-1}$, $t = 1, \dots, n-1$,
and $P(M_1) = P(M_2) = \frac{1}{2}$.

Model M_1

Likelihood: $f(x \mid \theta) = \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i}$

Conjugate Prior: $f(\theta) = \frac{1}{2} e^{-\theta/2}$

Posterior: $f(\theta \mid x) \propto f(\theta) f(x \mid \theta)$

$$= \frac{1}{2} e^{-\theta/2} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} = \frac{1}{2} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i}$$

$$\equiv Gamma(\textstyle\sum_{i=1}^{n} x_i + 1, n + \frac{1}{2}).$$

Model M_1

Likelihood: $f(x \mid \theta) = \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}$

Conjugate Prior: $f(\theta) = \frac{1}{2}e^{-\theta/2}$

Posterior: $f(\theta \mid x) \propto f(\theta)f(x \mid \theta)$

$$= \frac{1}{2}e^{-\theta/2}\left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i} = \frac{1}{2}\left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(n+\frac{1}{2})\theta}\theta^{\sum_{i=1}^{n} x_i}$$

$$\equiv Gamma(\textstyle\sum_{i=1}^{n} x_i + 1, n + \frac{1}{2}).$$

Evidence: $f(x \mid M_1) = \int f(\theta)f(x \mid \theta)d\theta$

Model M_1

Likelihood: $f(x \mid \theta) = \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i}$

Conjugate Prior: $f(\theta) = \frac{1}{2} e^{-\theta/2}$

Posterior: $f(\theta \mid x) \propto f(\theta) f(x \mid \theta)$

$$= \frac{1}{2} e^{-\theta/2} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} = \frac{1}{2} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i}$$

$$\equiv Gamma(\textstyle\sum_{i=1}^{n} x_i + 1, n + \frac{1}{2}).$$

Evidence: $f(x \mid M_1) = \int f(\theta) f(x \mid \theta) d\theta$

$$= \int_0^\infty \frac{1}{2} [\textstyle\prod_{i=1}^{n} \frac{1}{x_i!}] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i} d\theta$$

$$= \frac{1}{2} [\textstyle\prod_{i=1}^{n} \frac{1}{x_i!}] \int_0^\infty e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i} d\theta$$

Model M_1

Likelihood: $f(x \mid \theta) = \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i}$

Conjugate Prior: $f(\theta) = \frac{1}{2} e^{-\theta/2}$

Posterior: $f(\theta \mid x) \propto f(\theta) f(x \mid \theta)$

$$= \frac{1}{2} e^{-\theta/2} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} = \frac{1}{2} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i}$$

$$\equiv Gamma(\textstyle\sum_{i=1}^{n} x_i + 1, n + \frac{1}{2}).$$

Evidence: $f(x \mid M_1) = \int f(\theta) f(x \mid \theta) d\theta$

$$= \int_0^\infty \frac{1}{2} [\textstyle\prod_{i=1}^{n} \frac{1}{x_i!}] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i} d\theta$$

$$= \frac{1}{2} [\textstyle\prod_{i=1}^{n} \frac{1}{x_i!}] \int_0^\infty e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^{n} x_i} d\theta$$

$$= \frac{1}{2} [\textstyle\prod_{i=1}^{n} \frac{1}{x_i!}] \frac{\Gamma(\sum_{i=1}^{n} x_i + 1)}{(n+\frac{1}{2})^{\sum_{i=1}^{n} x_i + 1}}$$

Model M_2

Likelihood:

$$f(x \mid \theta_1, \theta_2, t) = \left[\prod_{i=1}^{t} \frac{1}{x_i!}\right] e^{-t\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} \left[\prod_{i=t+1}^{n} \frac{1}{x_i!}\right] e^{-(n-t)\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i}$$

Priors: $f(\theta_1) = \frac{1}{2} e^{-\theta_1/2}, \quad f(\theta_2) = \frac{1}{2} e^{-\theta_2/2}, \quad f(t) = \frac{1}{n-1}$

Posterior: $f(\theta_1, \theta_2, t \mid x) \propto f(\theta_1) f(\theta_2) f(t) f(x \mid \theta_1, \theta_2, t)$

$$= \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i}.$$

Model M_2

Likelihood:

$$f(x \mid \theta_1, \theta_2, t) = \left[\prod_{i=1}^{t} \frac{1}{x_i!}\right] e^{-t\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} \left[\prod_{i=t+1}^{n} \frac{1}{x_i!}\right] e^{-(n-t)\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i}$$

Priors: $f(\theta_1) = \frac{1}{2} e^{-\theta_1/2}, \quad f(\theta_2) = \frac{1}{2} e^{-\theta_2/2}, \quad f(t) = \frac{1}{n-1}$

Posterior: $f(\theta_1, \theta_2, t \mid x) \propto f(\theta_1) f(\theta_2) f(t) f(x \mid \theta_1, \theta_2, t)$

$$= \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i}.$$

Conditional Posteriors of θ_1 and θ_2 given t:

$$f(\theta_1 \mid x, t) \propto f(\theta_1, \theta_2, t \mid x) \propto e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} \equiv Gamma\left(\sum_{i=1}^{t} x_i + 1, t + \tfrac{1}{2}\right)$$

$$f(\theta_2 \mid x, t) \propto f(\theta_1, \theta_2, t \mid x) \propto e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i} \equiv Gamma\left(\sum_{i=t+1}^{n} x_i + 1, n - t + \tfrac{1}{2}\right)$$

Model M_2

Marginal Posterior of t:

$$
\begin{aligned}
f(t \mid x) &= \int_{\theta_1}\int_{\theta_2} f(\theta_1,\theta_2,t \mid x)d\theta_2 d\theta_1 \\
&\propto \int_{\theta_1}\int_{\theta_2} e^{-(t+\frac{1}{2})\theta_1}\theta_1^{\sum_{i=1}^t x_i} e^{-(n-t+\frac{1}{2})\theta_2}\theta_2^{\sum_{i=t+1}^n x_i} d\theta_2 d\theta_1 \\
&= \left[\int_0^\infty e^{-(t+\frac{1}{2})\theta_1}\theta_1^{\sum_{i=1}^t x_i} d\theta_1\right]\left[\int_0^\infty e^{-(n-t+\frac{1}{2})\theta_2}\theta_2^{\sum_{i=t+1}^n x_i} d\theta_2\right] \\
&= \frac{\Gamma(\sum_{i=1}^t x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^t x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^n x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^n x_i + 1}}
\end{aligned}
$$

Bayesian Model Comparison

Evidence:

$f(x \mid M_2) = \sum_{t=1}^{n-1} \int_{\theta_1} \int_{\theta_2} f(\theta_1) f(\theta_2) f(t) f(x \mid \theta_1, \theta_2, t) d\theta_2 d\theta_1 =$

Bayesian Model Comparison

Evidence:

$$f(x \mid M_2) = \sum_{t=1}^{n-1} \int_{\theta_1} \int_{\theta_2} f(\theta_1) f(\theta_2) f(t) f(x \mid \theta_1, \theta_2, t) d\theta_2 d\theta_1 =$$

$$\sum_{t=1}^{n-1} \int \int \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i} d\theta_2 d\theta_1$$

$$= \sum_{t=1}^{n-1} \left\{ \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] \frac{\Gamma(\sum_{i=1}^{t} x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^{t} x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^{n} x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^{n} x_i + 1}} \right\}$$

$$= \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] \sum_{t=1}^{n-1} \left\{ \frac{\Gamma(\sum_{i=1}^{t} x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^{t} x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^{n} x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^{n} x_i + 1}} \right\}$$

Bayesian Model Comparison

Evidence:

$$f(x \mid M_2) = \sum_{t=1}^{n-1} \int_{\theta_1} \int_{\theta_2} f(\theta_1) f(\theta_2) f(t) f(x \mid \theta_1, \theta_2, t) d\theta_2 d\theta_1 =$$

$$\sum_{t=1}^{n-1} \int \int \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^{t} x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^{n} x_i} d\theta_2 d\theta_1$$

$$= \sum_{t=1}^{n-1} \left\{ \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] \frac{\Gamma(\sum_{i=1}^{t} x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^{t} x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^{n} x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^{n} x_i + 1}} \right\}$$

$$= \frac{1}{4(n-1)} \left[\prod_{i=1}^{n} \frac{1}{x_i!}\right] \sum_{t=1}^{n-1} \left\{ \frac{\Gamma(\sum_{i=1}^{t} x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^{t} x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^{n} x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^{n} x_i + 1}} \right\}$$

Posterior model probabilities:

$$\Pr(M1 \mid x) = \frac{\Pr(M_1) f(x|M_1)}{\Pr(M_1) f(x|M_1) + \Pr(M_2) f(x|M_2)} = \frac{f(x|M_1)}{f(x|M_1) + f(x|M_2)}$$

$$\Pr(M2 \mid x) = 1 - \Pr(M1 \mid x).$$