

# Bayesian Inference

Loukia Meligkotsidou,  
National and Kapodistrian University of Athens

MSc in Statistics and Operational Research,  
Department of Mathematics

## 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
- ▶ Asymptotics
- ▶ Markov chain Monte Carlo Methods

# 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  $\theta = (\theta_1, \dots, \theta_d)$  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  $\theta$  is obtained by straightforward probability calculations on this joint distribution.

# Conditional Posterior Distributions

The *conditional posterior distribution* of a component of  $\theta$ ,  $\theta_i$  say, given the values of the remaining components  $\theta_{-i}$  is given by

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

where the values of  $\theta_{-i}$  are held fixed.

# Conditional Posterior Distributions

The *conditional posterior distribution* of a component of  $\theta$ ,  $\theta_i$  say, given the values of the remaining components  $\theta_{-i}$  is given by

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

where the values of  $\theta_{-i}$  are held fixed.

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

# Marginal Posterior Distributions

**Exact Bayesian inference** about the scalar parameter  $\theta_i$  can only be made from the posterior distribution integrated over  $\theta_{-i}$ ,

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

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

# Marginal Posterior Distributions

**Exact Bayesian inference** about the scalar parameter  $\theta_i$  can only be made from the posterior distribution integrated over  $\theta_{-i}$ ,

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

This resulting *marginal posterior* of a given parameter of interest  $\theta_i$ , after eliminating the nuisance parameters  $\theta_{-i}$  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  $\theta$ . 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  $\theta_1$  and  $\theta_2$ :  $f(\theta) = f(\theta_1)f(\theta_2)$

# 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  $\theta_1$  and  $\theta_2$ :  $f(\theta) = f(\theta_1)f(\theta_2)$
- ▶ Another choice can be a bivariate distribution factorised as a product of a conditional times a marginal density:

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

# 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  $\theta_1$  and  $\theta_2$ :  $f(\theta) = f(\theta_1)f(\theta_2)$
- ▶ Another choice can be a bivariate distribution factorised as a product of a conditional times a marginal density:

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

- ▶ The most general case is to assume a **bivariate** prior distribution allowing for dependence (correlation) between  $\theta_1$  and  $\theta_2$ , for example a bivariate normal density.

**Note:** Generalisation to the case of a multivariate parameter  $\theta$ .

## 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.

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

The joint prior distribution of  $(\theta_1, \theta_2)$  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.

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x = 1 \theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	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.

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

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(x = 1   \theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	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

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  $Y_1 \sim \text{Poisson}(\alpha\beta)$  and  $Y_2 \sim \text{Poisson}((1 - \alpha)\beta)$  with  $Y_1$  and  $Y_2$  independent given  $\alpha$  and  $\beta$ .

Suppose our prior information for  $\alpha$  and  $\beta$  can be expressed as:  $\alpha \sim \text{Beta}(p, q)$  and  $\beta \sim \text{Gamma}(p + q, 1)$  with  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$  and contains all the information from the prior and data.

In this particular case, the posterior factorises into functions of  $\alpha$  and  $\beta$ . Therefore, we can write:

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

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

# 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,  $\alpha|y_1, y_2 \sim \text{Beta}(y_1 + p, y_2 + q)$  and  $\beta|y_1, y_2 \sim \text{Gamma}(y_1 + y_2 + p + q, 2)$ .

# 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,  $\alpha|y_1, y_2 \sim \text{Beta}(y_1 + p, y_2 + q)$  and  $\beta|y_1, y_2 \sim \text{Gamma}(y_1 + y_2 + p + q, 2)$ .

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

## Example. The Normal Likelihood

Suppose  $X_1, \dots, X_n$  are independent variables distributed as  $N(\mu, \phi^{-1})$ . Obtain the conjugate prior for  $\theta = (\mu, \phi)$

$$\begin{aligned} f(x_i|\theta) &= \frac{\sqrt{\phi}}{\sqrt{2\pi}} \exp\left\{-\frac{\phi}{2}(x_i - \mu)^2\right\} \\ &= \frac{\sqrt{\phi}}{\sqrt{2\pi}} \exp\left\{-\frac{\phi}{2}x_i^2 - \frac{\phi}{2}\mu^2 + \phi x_i \mu\right\} \end{aligned}$$

## Example. The Normal Likelihood

Suppose  $X_1, \dots, X_n$  are independent variables distributed as  $N(\mu, \phi^{-1})$ . Obtain the conjugate prior for  $\theta = (\mu, \phi)$

$$\begin{aligned} f(x_i|\theta) &= \frac{\sqrt{\phi}}{\sqrt{2\pi}} \exp\left\{-\frac{\phi}{2}(x_i - \mu)^2\right\} \\ &= \frac{\sqrt{\phi}}{\sqrt{2\pi}} \exp\left\{-\frac{\phi}{2}x_i^2 - \frac{\phi}{2}\mu^2 + \phi x_i \mu\right\} \\ &= \frac{\sqrt{\phi}}{\sqrt{2\pi}} \exp\left\{-\frac{\phi}{2}\mu^2\right\} \exp\left\{-\frac{\phi}{2}x_i^2 + \phi x_i \mu\right\} \\ &= h(x_i)g(\theta) \exp\{t_1(x_i)c_1(\theta) + t_2(x_i)c_2(\theta)\}, \end{aligned}$$

where  $h(x_i) = \frac{1}{\sqrt{2\pi}}$ ,  $g(\theta) = \phi^{1/2} \exp\{-\frac{\phi}{2}\mu^2\}$ ,  $t_1(x_i) = x_i^2$ ,  
 $c_1(\theta) = -\frac{\phi}{2}$ ,  $t_2(x_i) = x_i$ ,  $c_2(\theta) = \phi\mu$ .



# The Joint Prior

The likelihood:

$$f(x | \theta) \propto [g(\theta)]^n \exp\left\{\sum_{i=1}^n t_1(x_i) c_1(\theta) + \sum_{i=1}^n t_1(x_i) t_2(x_i) c_2(\theta)\right\}$$

The conjugate prior:

$$f(\theta) \propto [g(\theta)]^d \exp\{b_1 c_1(\theta) + b_2 c_2(\theta)\}$$

# The Joint Prior

The likelihood:

$$f(x | \theta) \propto [g(\theta)]^n \exp\left\{\sum_{i=1}^n t_1(x_i) c_1(\theta) + \sum_{i=1}^n t_1(x_i) t_2(x_i) c_2(\theta)\right\}$$

The conjugate prior:

$$\begin{aligned} f(\theta) &\propto [g(\theta)]^d \exp\{b_1 c_1(\theta) + b_2 c_2(\theta)\} \\ &= \left[\phi^{1/2} \exp\left\{-\frac{\phi}{2} \mu^2\right\}\right]^d \exp\left\{-b_1 \frac{\phi}{2} + b_2 \phi \mu\right\} \end{aligned}$$

# The Joint Prior

The likelihood:

$$f(x | \theta) \propto [g(\theta)]^n \exp\left\{\sum_{i=1}^n t_1(x_i) c_1(\theta) + \sum_{i=1}^n t_1(x_i) t_2(x_i) c_2(\theta)\right\}$$

The conjugate prior:

$$\begin{aligned} f(\theta) &\propto [g(\theta)]^d \exp\{b_1 c_1(\theta) + b_2 c_2(\theta)\} \\ &= \left[\phi^{1/2} \exp\left\{-\frac{\phi}{2} \mu^2\right\}\right]^d \exp\left\{-b_1 \frac{\phi}{2} + b_2 \phi \mu\right\} \\ &= \phi^{d/2} \exp\left\{-\frac{d\phi}{2} \mu^2 - b_1 \frac{\phi}{2} + b_2 \phi \mu\right\} \end{aligned}$$

# The Joint Prior

The likelihood:

$$f(x | \theta) \propto [g(\theta)]^n \exp\left\{\sum_{i=1}^n t_1(x_i) c_1(\theta) + \sum_{i=1}^n t_1(x_i) t_2(x_i) c_2(\theta)\right\}$$

The conjugate prior:

$$\begin{aligned} f(\theta) &\propto [g(\theta)]^d \exp\{b_1 c_1(\theta) + b_2 c_2(\theta)\} \\ &= \left[\phi^{1/2} \exp\left\{-\frac{\phi}{2} \mu^2\right\}\right]^d \exp\left\{-b_1 \frac{\phi}{2} + b_2 \phi \mu\right\} \\ &= \phi^{d/2} \exp\left\{-\frac{d\phi}{2} \mu^2 - b_1 \frac{\phi}{2} + b_2 \phi \mu\right\} \\ &= \phi^{1/2} \exp\left\{-\frac{d\phi}{2} \left[\mu^2 - 2\mu \frac{b_2}{d}\right]\right\} \times \phi^{\frac{d-1}{2}} \exp\left\{-\phi \frac{b_1}{2}\right\} \end{aligned}$$

# The Joint Conjugate Prior

$$f(\theta) \propto \phi^{1/2} \exp\left\{-\frac{d\phi}{2}\left[\mu^2 - 2\mu\frac{b_2}{d}\right]\right\} \times \phi^{\frac{d-1}{2}} \exp\left\{-\phi\frac{b_1}{2}\right\}$$

# The Joint Conjugate Prior

$$\begin{aligned} f(\theta) &\propto \phi^{1/2} \exp\left\{-\frac{d\phi}{2}\left[\mu^2 - 2\mu\frac{b_2}{d}\right]\right\} \times \phi^{\frac{d-1}{2}} \exp\left\{-\phi\frac{b_1}{2}\right\} \\ &= \phi^{1/2} \exp\left\{-\frac{d\phi}{2}\left[\mu^2 - 2\mu\alpha\right]\right\} \times \phi^{p-1} \exp\{-\phi q\} \end{aligned}$$

# The Joint Conjugate Prior

$$\begin{aligned}f(\theta) &\propto \phi^{1/2} \exp\left\{-\frac{d\phi}{2}\left[\mu^2 - 2\mu\frac{b_2}{d}\right]\right\} \times \phi^{\frac{d-1}{2}} \exp\left\{-\phi\frac{b_1}{2}\right\} \\ &= \phi^{1/2} \exp\left\{-\frac{d\phi}{2}\left[\mu^2 - 2\mu\alpha\right]\right\} \times \phi^{p-1} \exp\{-\phi q\}\end{aligned}$$

That is the joint conjugate prior factorizes as

$$f(\theta) = f(\mu, \phi) = f(\phi)f(\mu | \phi),$$

where  $f(\phi) \equiv \text{Gamma}(p, q)$  and  $f(\mu | \phi) \equiv N(\alpha, (d\phi)^{-1})$ .

## 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  $\theta$ . 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  $\theta$ . 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  $\theta$ . 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**  $\Theta$  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  $\theta$ .

## 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  $\theta_1$  and  $\theta_2$  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

$L(\theta, a)$	$\theta_1$	$\theta_2$
$a_1$	8	0
$a_2$	0	20



So, for this example we have:

1. The parameter space  $\Theta = \{\theta_1, \theta_2\}$ , where  $\theta_1$  and  $\theta_2$  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

$L(\theta, a)$	$\theta_1$	$\theta_2$
$a_1$	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  $\theta$ .

## Minimising the Prior Expected Loss

The expected loss is calculated based on the prior distribution of  $\theta$ .

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  $\theta$ .

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	
$L(\theta, a)$	$\theta_1$	$\theta_2$	$E[L(\theta, a)]$
$a_1$	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  $\theta$ .

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	
$L(\theta, a)$	$\theta_1$	$\theta_2$	$E[L(\theta, a)]$
$a_1$	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  $\theta$ , 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  $\theta$ , 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	Vulnerable
			$\theta_1$	$\theta_2$
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

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

Having observed a particular value of  $x$ , we choose the action  $a$  which results in the lowest value of  $\rho$ . 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

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

Having observed a particular value of  $x$ , we choose the action  $a$  which results in the lowest value of  $\rho$ . 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.

		$\theta_1$	$\theta_2$
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

		$\theta_1$	$\theta_2$	
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	

		$\theta_1$	$\theta_2$		
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	$f(x_4)$

		$\theta_1$	$\theta_2$			
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	Expected Losses
	$f(\theta x_3)$	0.558	0.442	4.460	8.847	
	$f(\theta x_4)$	0.300	0.700	2.400	14.000	

# 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.

## Example

A baseball team needs to decide whether or not to adopt a promotional campaign with a cost of 0.5 million dollars.

Denote by  $\theta$  the proportion of matches that the team will win in the next period. The (prior) belief of the team is that  $\theta$  has a uniform distribution in  $[0, 1]$ .

They expect that, without the campaign, their benefit from the tickets will be  $1 + 5\theta$  million dollars, while, with the campaign, their benefit will be  $3 + 3\theta$  million dollars. Should they adopt the campaign?

So, for this example:

1. The parameter space is continuous  $\Theta = [0, 1]$ , with  $\theta$  denoting the probability that the team wins a baseball match.
2. The set of actions is  $A = \{a_1, a_2\}$  where  $a_1$  and  $a_2$  correspond to adopting and not adopting the campaign, respectively.
3. The loss function is

$$L(\theta, a_1) = 0.5 - 3 - 3\theta = -2.5 - 3\theta$$

$$L(\theta, a_2) = -1 - 5\theta$$

So, for this example:

1. The parameter space is continuous  $\Theta = [0, 1]$ , with  $\theta$  denoting the probability that the team wins a baseball match.
2. The set of actions is  $A = \{a_1, a_2\}$  where  $a_1$  and  $a_2$  correspond to adopting and not adopting the campaign, respectively.
3. The loss function is

$$L(\theta, a_1) = 0.5 - 3 - 3\theta = -2.5 - 3\theta$$

$$L(\theta, a_2) = -1 - 5\theta$$

$$E[L(\theta, a_1)] = \int L(\theta, a_1)f(\theta)d\theta = \int_0^1(-2.5 - 3\theta)d\theta = -4$$

$$E[L(\theta, a_2)] = \int L(\theta, a_2)f(\theta)d\theta = \int_0^1(-1 - 5\theta)d\theta = -3.5$$

So, for this example:

1. The parameter space is continuous  $\Theta = [0, 1]$ , with  $\theta$  denoting the probability that the team wins a baseball match.
2. The set of actions is  $A = \{a_1, a_2\}$  where  $a_1$  and  $a_2$  correspond to adopting and not adopting the campaign, respectively.
3. The loss function is

$$L(\theta, a_1) = 0.5 - 3 - 3\theta = -2.5 - 3\theta$$

$$L(\theta, a_2) = -1 - 5\theta$$

$$E[L(\theta, a_1)] = \int L(\theta, a_1)f(\theta)d\theta = \int_0^1(-2.5 - 3\theta)d\theta = -4$$

$$E[L(\theta, a_2)] = \int L(\theta, a_2)f(\theta)d\theta = \int_0^1(-1 - 5\theta)d\theta = -3.5$$

$E[L(\theta, a_1)] < E[L(\theta, a_2)]$ , therefore the expected loss is minimized (the expected profit is maximized) if the team adopts the campaign.

## Example. Continuation

In the playoffs, the team wins 4 out of 4 matches. If it is equally probable the team to win a match at the playoffs and at the championship, which is now the optimal decision with respect to the campaign?

## Example. Continuation

In the playoffs, the team wins 4 out of 4 matches. If it is equally probable the team to win a match at the playoffs and at the championship, which is now the optimal decision with respect to the campaign?

Let  $X$  denote the number of matches that the team wins at the playoffs and  $\theta$  denote the probability of win. The likelihood of observing  $x = n = 4$  is:

$$f(\theta | x) = \theta^4(1 - \theta)^{4-4}$$



## Example. Continuation

In the playoffs, the team wins 4 out of 4 matches. If it is equally probable the team to win a match at the playoffs and at the championship, which is now the optimal decision with respect to the campaign?

Let  $X$  denote the number of matches that the team wins at the playoffs and  $\theta$  denote the probability of win. The likelihood of observing  $x = n = 4$  is:

$$f(\theta | x) = \theta^4(1 - \theta)^{4-4}$$

Posterior:  $f(\theta | x) \propto f(\theta)f(x | \theta) \propto \theta^4$ .

$f(x) = \int_0^1 \theta^4 d\theta = \frac{1}{5}$ , therefore  $f(\theta | x) = 5\theta^4$ ,  $0 < \theta < 1$ .

## Posterior Expected Losses

$$E[L(\theta, a_1) | x] = \int L(\theta, a_1)f(\theta | x)d\theta = \int_0^1 (-2.5-3\theta)5\theta^4 d\theta = -5$$

$$E[L(\theta, a_2) | x] = \int L(\theta, a_2)f(\theta | x)d\theta = \int_0^1 (-1-5\theta)5\theta^4 d\theta = -3.5$$

# Posterior Expected Losses

$$E[L(\theta, a_1) | x] = \int L(\theta, a_1)f(\theta | x)d\theta = \int_0^1 (-2.5-3\theta)5\theta^4 d\theta = -5$$

$$E[L(\theta, a_2) | x] = \int L(\theta, a_2)f(\theta | x)d\theta = \int_0^1 (-1-5\theta)5\theta^4 d\theta = -3.5$$

$E[L(\theta, a_1) | x] > E[L(\theta, a_2) | x]$ , therefore, after observing the data the expected loss is minimized (the expected profit is maximized) if the team does not adopt the campaign.

# 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  $\theta$  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  $\theta$  for that particular choice of loss function.



# Squared Error Loss

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

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

# Squared Error Loss

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

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

# Squared Error Loss

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

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

## Squared Error Loss

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

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

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  $\text{Var}[\theta | x]$ , the posterior variance of  $\theta$ , when  $a = E(\theta | x)$ , the **posterior expectation** of  $\theta$ .

# 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

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

# 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  $\theta < a$ :

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

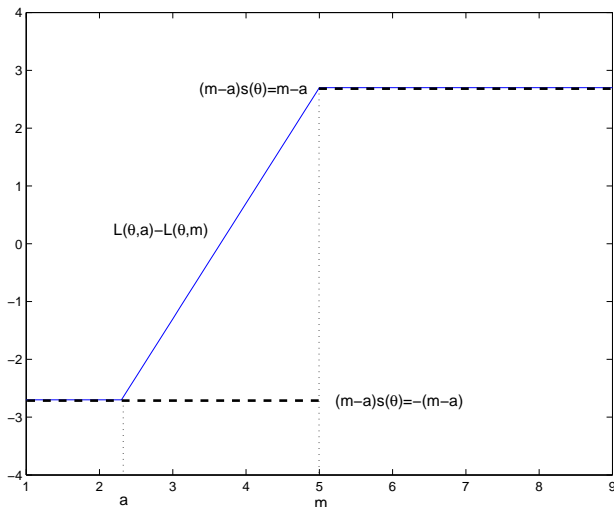
If  $\theta > m$ :

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

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

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

# Absolute Error Loss

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

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

So

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

# Absolute Error Loss

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

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

So

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

This also holds by a similar argument when  $a > m$ , so  $E [L(\theta, a)|x]$  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  $2\epsilon$*  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  $\epsilon$  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*  $2\epsilon$  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  $\epsilon$  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  $\theta$  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  $\theta$  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  $\phi$  with respect to quadratic loss is

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

# 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  $\theta$  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  $\theta$  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:

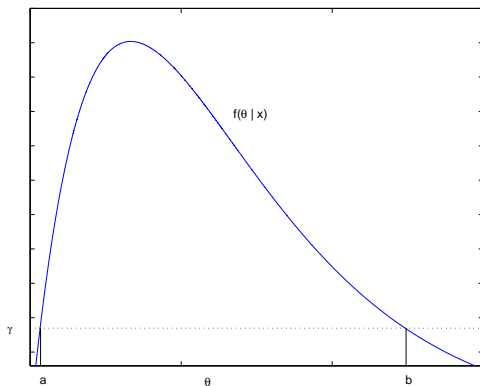
$$C_\alpha(x) = \{\theta : f(\theta|x) \geq \gamma\}$$

where  $\gamma$  is chosen to ensure that

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

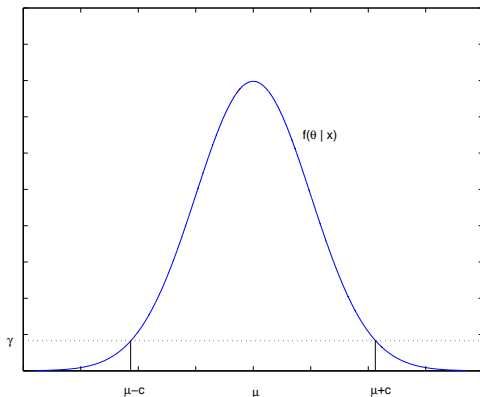
# Unimodal Posterior Distribution

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

With this construction we obtained the posterior:

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

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

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

With this construction we obtained the posterior:

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

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

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  $\theta$  is:

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

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

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

## Example

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

$$\theta \sim \text{Beta}(p, q).$$

This gives the posterior distribution

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

## Example

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

$$\theta \sim \text{Beta}(p, q).$$

This gives the posterior distribution

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

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  $\Omega_1$  and  $\Omega_2$  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  $\lambda$  indicate that the observed data  $x$  is more likely to have occurred if  $\theta_1$  is the true value of  $\theta$  than if  $\theta_0$  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  $\lambda$  indicate that the observed data  $x$  is more likely to have occurred if  $\theta_1$  is the true value of  $\theta$  than if  $\theta_0$  is.

In the Bayesian view of things, we should also bring to bear the prior information we have about  $\theta$ . Therefore, we may compute the posterior probabilities of  $\theta_1$  and  $\theta_0$ :

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

# 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.

# Bayesian Model Comparison

Bayesian model comparison is a generalisation of Bayesian hypothesis testing.



# Bayesian 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:  $M_1, M_2, \dots, M_k$ .

# Bayesian 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:  $M_1, M_2, \dots, M_k$ .

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

# Bayesian 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:  $M_1, M_2, \dots, M_k$ .

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

We compute the posterior model probabilities as

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

# Bayesian 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:  $M_1, M_2, \dots, M_k$ .

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

We compute the posterior model probabilities as

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

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  $f(x | \theta)$  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  $\theta$  over  $\theta$ :

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

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

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

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  $f(x | \theta)$  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  $\theta$  over  $\theta$ :

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

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

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

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

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

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,  $f(\theta_j | M_j)$ ,  $j = 1, \dots, k$ , posterior inferences are obtained as follows.

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

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

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

2. The posterior distribution of the parameters  $\theta_j$  of model  $M_j$  is given by Bayes theorem as

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

# 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  $\theta$  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  $M_1, \dots, M_k$ . Let  $\theta$  denote *generically* the parameter(s).

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

# 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  $M_1, \dots, M_k$ . Let  $\theta$  denote *generically* the parameter(s).

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

Joint Posterior:  $f(\theta, M | x) \propto f(M)f(\theta | M)f(x | \theta, M)$

## 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  $M_1, \dots, M_k$ . Let  $\theta$  denote *generically* the parameter(s).

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

Joint Posterior:  $f(\theta, M | x) \propto f(M)f(\theta | M)f(x | \theta, M)$

Marginal Posterior of  $M$ :

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

## 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  $M_1, \dots, M_k$ . Let  $\theta$  denote *generically* the parameter(s).

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

Joint Posterior:  $f(\theta, M | x) \propto f(M)f(\theta | M)f(x | \theta, M)$

Marginal Posterior of  $M$ :

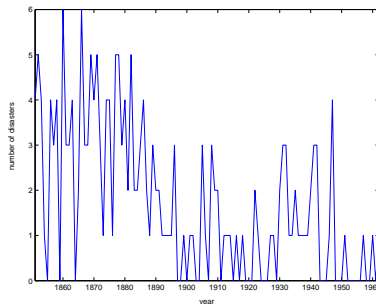
$$\begin{aligned} f(M | x) &\propto \int f(M)f(\theta | M)f(x | \theta, M)d\theta \\ &= f(M) \int f(\theta | M)f(x | \theta, M)d\theta = f(M)f(x | M) \end{aligned}$$

Conditional Posterior of  $\theta$ :

$$f(\theta | x, M) \propto f(\theta, M | x) \propto f(\theta | M)f(x | \theta, M)$$

## 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.

# Bayesian Model Comparison

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

- M1 each  $x_j$  is an independent draw from a Poisson random variable with mean  $\theta$ ;
- M2 for  $i \leq t$ ,  $x_i$  is an independent draw from a Poisson random variable with mean  $\theta_1$ , and for  $i > t$ ,  $x_i$  is an independent draw from a Poisson random variable with mean  $\theta_2$ .

In the first model there is just one unknown parameter,  $\theta$ . In the second model, there are three unknown parameters:  $\theta_1$ ,  $\theta_2$  and  $t$ .



# Prior Specification

Model  $M_1$ :  $X_i \sim \text{Poisson}(\theta)$ ,  $i = 1, \dots, n$

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

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

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

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

Model  $M_1$ 

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

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

$$\text{Posterior: } f(\theta | x) \propto f(\theta) f(x | \theta)$$

$$\begin{aligned} &= \frac{1}{2} e^{-\theta/2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-n\theta} \theta^{\sum_{i=1}^n x_i} = \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} \\ &\equiv \text{Gamma}(\sum_{i=1}^n x_i + 1, n + \frac{1}{2}). \end{aligned}$$

Model  $M_1$ 

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

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

$$\text{Posterior: } f(\theta | x) \propto f(\theta) f(x | \theta)$$

$$\begin{aligned} &= \frac{1}{2} e^{-\theta/2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-n\theta} \theta^{\sum_{i=1}^n x_i} = \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} \\ &\equiv \text{Gamma}(\sum_{i=1}^n x_i + 1, n + \frac{1}{2}). \end{aligned}$$

$$\text{Evidence: } f(x | M_1) = \int f(\theta) f(x | \theta) d\theta$$

Model  $M_1$ 

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

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

$$\begin{aligned} \text{Posterior: } f(\theta | x) &\propto f(\theta) f(x | \theta) \\ &= \frac{1}{2} e^{-\theta/2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-n\theta} \theta^{\sum_{i=1}^n x_i} = \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} \\ &\equiv \text{Gamma}(\sum_{i=1}^n x_i + 1, n + \frac{1}{2}). \end{aligned}$$

$$\begin{aligned} \text{Evidence: } f(x | M_1) &= \int f(\theta) f(x | \theta) d\theta \\ &= \int_0^\infty \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} d\theta \\ &= \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \int_0^\infty e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} d\theta \end{aligned}$$

Model  $M_1$ 

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

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

$$\begin{aligned} \text{Posterior: } f(\theta | x) &\propto f(\theta) f(x | \theta) \\ &= \frac{1}{2} e^{-\theta/2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-n\theta} \theta^{\sum_{i=1}^n x_i} = \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} \\ &\equiv \text{Gamma}(\sum_{i=1}^n x_i + 1, n + \frac{1}{2}). \end{aligned}$$

$$\begin{aligned} \text{Evidence: } f(x | M_1) &= \int f(\theta) f(x | \theta) d\theta \\ &= \int_0^\infty \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} d\theta \\ &= \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \int_0^\infty e^{-(n+\frac{1}{2})\theta} \theta^{\sum_{i=1}^n x_i} d\theta \\ &= \frac{1}{2} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \frac{\Gamma(\sum_{i=1}^n x_i + 1)}{(n+\frac{1}{2})^{\sum_{i=1}^n x_i + 1}} \end{aligned}$$

Model  $M_2$ 

Likelihood:

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

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

$$\begin{aligned} \text{Posterior: } f(\theta_1, \theta_2, t | x) &\propto f(\theta_1) f(\theta_2) f(t) f(x | \theta_1, \theta_2, t) \\ &= \frac{1}{4(n-1)} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^t x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^n x_i} \end{aligned}$$

Model  $M_2$ 

Likelihood:

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

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

$$\begin{aligned} \text{Posterior: } f(\theta_1, \theta_2, t | x) &\propto f(\theta_1) f(\theta_2) f(t) f(x | \theta_1, \theta_2, t) \\ &= \frac{1}{4(n-1)} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^t x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^n x_i} \end{aligned}$$

Conditional Posteriors of  $\theta_1$  and  $\theta_2$  given  $t$ :

$$\begin{aligned} f(\theta_1 | x, t) &\propto f(\theta_1, \theta_2, t | x) \propto e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^t x_i} \\ &\equiv \text{Gamma}(\sum_{i=1}^t x_i + 1, t + \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} f(\theta_2 | x, t) &\propto f(\theta_1, \theta_2, t | x) \propto e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^n x_i} \\ &\equiv \text{Gamma}(\sum_{i=t+1}^n x_i + 1, n - t + \frac{1}{2}) \end{aligned}$$

Model  $M_2$ Marginal Posterior of  $t$ :

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



# Bayesian Model Comparison

Evidence of  $M_2$ :

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

## Bayesian Model Comparison

Evidence of  $M_2$ :

$$\begin{aligned} f(x | M_2) &= \sum_{t=1}^{n-1} \int_{\theta_1} \int_{\theta_2} f(\theta_1) f(\theta_2) f(t) f(x | \theta_1, \theta_2, t) d\theta_2 d\theta_1 = \\ &= \sum_{t=1}^{n-1} \int \int \frac{1}{4(n-1)} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^t x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^n x_i} d\theta_2 d\theta_1 \\ &= \sum_{t=1}^{n-1} \left\{ \frac{1}{4(n-1)} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \frac{\Gamma(\sum_{i=1}^t x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^t x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^n x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^n x_i + 1}} \right\} \\ &= \frac{1}{4(n-1)} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \sum_{t=1}^{n-1} \left\{ \frac{\Gamma(\sum_{i=1}^t x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^t x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^n x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^n x_i + 1}} \right\} \end{aligned}$$

# Bayesian Model Comparison

Evidence of  $M_2$ :

$$\begin{aligned} f(x | M_2) &= \sum_{t=1}^{n-1} \int_{\theta_1} \int_{\theta_2} f(\theta_1) f(\theta_2) f(t) f(x | \theta_1, \theta_2, t) d\theta_2 d\theta_1 = \\ &= \sum_{t=1}^{n-1} \int \int \frac{1}{4^{(n-1)}} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] e^{-(t+\frac{1}{2})\theta_1} \theta_1^{\sum_{i=1}^t x_i} e^{-(n-t+\frac{1}{2})\theta_2} \theta_2^{\sum_{i=t+1}^n x_i} d\theta_2 d\theta_1 \\ &= \sum_{t=1}^{n-1} \left\{ \frac{1}{4^{(n-1)}} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \frac{\Gamma(\sum_{i=1}^t x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^t x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^n x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^n x_i + 1}} \right\} \\ &= \frac{1}{4^{(n-1)}} \left[ \prod_{i=1}^n \frac{1}{x_i!} \right] \sum_{t=1}^{n-1} \left\{ \frac{\Gamma(\sum_{i=1}^t x_i + 1)}{(t+\frac{1}{2})^{\sum_{i=1}^t x_i + 1}} \frac{\Gamma(\sum_{i=t+1}^n x_i + 1)}{(n-t+\frac{1}{2})^{\sum_{i=t+1}^n x_i + 1}} \right\} \end{aligned}$$

Posterior model probabilities:

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

$$\Pr(M_2 | x) = 1 - \Pr(M_1 | x).$$