# MAS3902: Bayesian Inference

Dr. Lee Fawcett

*School of Mathematics, Statistics & Physics*

**Semester 1, 2019/20**

# Contact details


**My name:** Lee Fawcett

**Office:** Room 2.07 Herschel Building

**Phone:** 0191 208 7228

**Email:** lee.fawcett@ncl.ac.uk

**www:** www.mas.ncl.ac.uk/∼nlf8

# Timetable and Administrative arrangements

- **Classes** are on Mondays at 3, Tuesdays at 1 and Thursdays at 2, all in LT2 of the Herschel Building
- Two of these classes will be **lectures**, and the other session will be a **problems class/drop-in**
  - Lectures will *ordinarily* be in the Monday/Tuesday slots, and PCs/DIs in the Thursday slot
  - PCs will take place in even teaching weeks, DIs in odd teaching weeks
  - For the first two weeks all slots will be used as lectures
- **Tutorials** will take place on some Thursdays to support project work. I will remind you about these sessions in advance – they all take place in the Herschel Learning Lab
- **Office hours** will be scheduled soon, but just come along and give me a knock!

# Assessment

Assessment is by:

- End of semester exam in May/June (**85%**)
- In course assessment (**15%**), including:
  - One group project (10%)
  - Three group homework exercises (5% in total)
- The homework exercises will be taken from the questions at the back of the lectures notes; we will work through unassessed questions in problems classes

# Recommended textbooks

- "*Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*" - James Stone
- "*Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*" - John Krushke
- "*Bayesian Statistics: An Introduction*" - Peter Lee

"Bayes' rule" is a good introduction to the main concepts in Bayesian statistics but doesn't cover everything in this course. The other books are broader references which go well beyond the contents of this course.

# Other stuff

- Notes (with gaps) will be handed out in lectures – you should fill in the gaps during lectures
- A (very!) summarised version of the notes will be used in lectures as slides
- These notes and slides will be posted on the **course website** and/or BlackBoard after each topic is finished, along with any other course material – such as problems sheets, model solutions to assignment questions, supplementary handouts etc.

# Chapter 1

## Single parameter problems

# Chapter 1. Single parameter problems

## 1.1 Prior and posterior distributions

Data: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$

Model: pdf/pf $f(\boldsymbol{x}|\theta)$ depends on a single parameter $\theta$

$\rightarrow$ Likelihood function $f(\boldsymbol{x}|\theta)$

considered as a function of $\theta$ for known $\boldsymbol{x}$

Prior beliefs: pdf/pf $\pi(\theta)$

Combine using Bayes Theorem

Posterior beliefs: pdf/pf $\pi(\theta|\boldsymbol{x})$

Posterior distribution summarises all our current knowledge about the parameter $\theta$

## Bayes Theorem

The posterior probability (density) function for $\theta$ is

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})}$$

where

$$f(\boldsymbol{x}) = \begin{cases} \int_\Theta \pi(\theta)\, f(\boldsymbol{x}|\theta)\, d\theta & \text{if } \theta \text{ is continuous,} \\ \\ \sum_\Theta \pi(\theta)\, f(\boldsymbol{x}|\theta) & \text{if } \theta \text{ is discrete.} \end{cases}$$

As $f(\boldsymbol{x})$ is not a function of $\theta$, Bayes Theorem can be rewritten as

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \times f(\boldsymbol{x}|\theta)$$

i.e. posterior $\propto$ prior $\times$ likelihood

## Example 1.1

- Data

| Year | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|---|---|
| Cases | 2 | 0 | 0 | 0 | 1 | 0 | 2 | 1 |

Table: Number of cases of foodbourne botulism in England and Wales, 1998–2005

- Assume that cases occur at random at a constant rate $\theta$ in time
- This means the data are from a Poisson process and so are a random sample from a Poisson distribution with rate $\theta$
- Prior $\theta \sim Ga(2, 1)$, with density

$$\pi(\theta) = \theta\, e^{-\theta}, \quad \theta > 0, \tag{1.1}$$

and mean $E(\theta) = 2$ and variance $Var(\theta) = 2$

- Determine the posterior distribution for $\theta$

## Solution

. . .

(1.2)

(1.3)

# Summary

- Model: $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, 8$ (independent)
- Prior: $\theta \sim Ga(2, 1)$
- Data: in Table above
- Posterior: $\theta|\boldsymbol{x} \sim Ga(8, 9)$

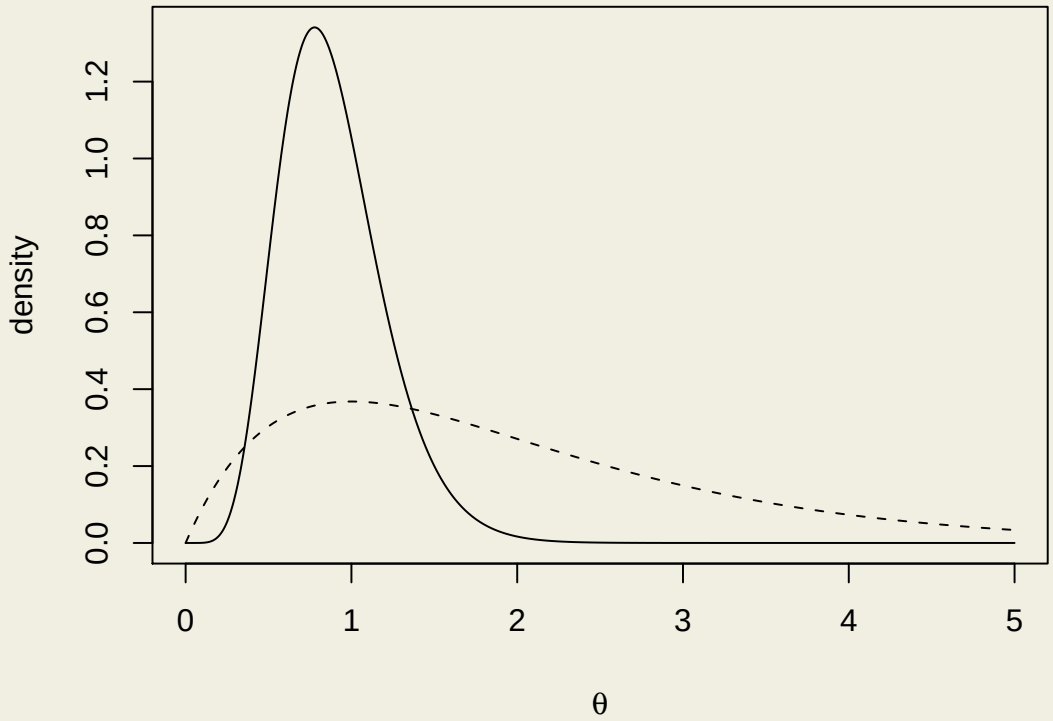



Figure: Prior (dashed) and posterior (solid) densities for $\theta$

| | Prior (1.1) | Likelihood (1.2) | Posterior (1.3) |
|---|---|---|---|
| $Mode(\theta)$ | 1.00 | 0.75 | 0.78 |
| $E(\theta)$ | 2.00 | – | 0.89 |
| $SD(\theta)$ | 1.41 | – | 0.31 |

Table: Changes in beliefs about $\theta$

- Likelihood mode $<$ prior mode $\rightarrow$ posterior mode moves in direction of likelihood mode $\rightarrow$ posterior mode $<$ prior mode
- Reduction in variability from the prior to the posterior

## Example 1.2

We now consider the general case (of Example 1.1). Suppose $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent) and our prior beliefs about $\theta$ are summarised by a $Ga(g, h)$ distribution (with $g$ and $h$ known), with density

$$\pi(\theta) = \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{1.4}$$

Determine the posterior distribution for $\theta$.

## Solution

. . .

(1.5)

(1.6)

# Summary

- Model: $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent)
- Prior: $\theta \sim Ga(g, h)$
- Data: observe $\boldsymbol{x}$
- Posterior: $\theta|\boldsymbol{x} \sim Ga(g + n\bar{x}, h + n)$
- Taking $g \geq 1$

| | Prior (1.4) | Likelihood (1.5) | Posterior (1.6) |
|---|---|---|---|
| $Mode(\theta)$ | $(g-1)/h$ | $\bar{x}$ | $(g + n\bar{x} - 1)/(h + n)$ |
| $E(\theta)$ | $g/h$ | – | $(g + n\bar{x})/(h + n)$ |
| $SD(\theta)$ | $\sqrt{g}/h$ | – | $\sqrt{g + n\bar{x}}/(h + n)$ |

Table: Changes in beliefs about $\theta$

## Comments

- Posterior mean is greater than the prior mean if and only if the likelihood mode is greater than the prior mean, that is,

$$E(\theta|\boldsymbol{x}) > E(\theta) \quad \Longleftrightarrow \quad Mode_{\theta}\{f(\boldsymbol{x}|\theta)\} > E(\theta)$$

- Standard deviation of the posterior distribution is smaller than that of the prior distribution if and only if the sample mean is not too large, that is

$$SD(\theta|\boldsymbol{x}) < SD(\theta) \quad \Longleftrightarrow \quad Mode_{\theta}\{f(\boldsymbol{x}|\theta)\} < \left(2 + \frac{n}{h}\right) E(\theta),$$

and that this will be true in large samples.

## Example 1.3

Suppose we have a random sample from a normal distribution with unknown mean $\mu$ but known precision $\tau$: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent).

Suppose our prior beliefs about $\mu$ can be summarised by a $N(b, 1/d)$ distribution, with probability density function

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \tag{1.7}$$

Determine the posterior distribution for $\mu$.

Hint:

$$d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2 = (d + n\tau)\left\{\mu - \left(\frac{db + n\tau\bar{x}}{d + n\tau}\right)\right\}^2 + c$$

where $c$ does not depend on $\mu$.

## Solution

. . .

(1.8)

(1.9)

(1.10)

# Summary

- Model: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent), with $\tau$ known
- Prior: $\mu \sim N(b, 1/d)$
- Data: observe $\boldsymbol{x}$
- Posterior: $\mu|\boldsymbol{x} \sim N(B, 1/D)$, where

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad D = d + n\tau$$

| | Prior (1.7) | Likelihood (1.8) | Posterior (1.10) |
|---|---|---|---|
| $Mode(\mu)$ | $b$ | $\bar{x}$ | $(db + n\tau\bar{x})/(d + n\tau)$ |
| $E(\mu)$ | $b$ | – | $(db + n\tau\bar{x})/(d + n\tau)$ |
| $Precision(\mu)$ | $d$ | – | $d + n\tau$ |

Table: Changes in beliefs about $\mu$

- Posterior mean is greater than the prior mean if and only if the likelihood mode (sample mean) is greater than the prior mean, that is

$$E(\mu|\boldsymbol{x}) > E(\mu) \quad \Longleftrightarrow \quad Mode_{\mu}\{f(\boldsymbol{x}|\mu)\} > E(\mu)$$

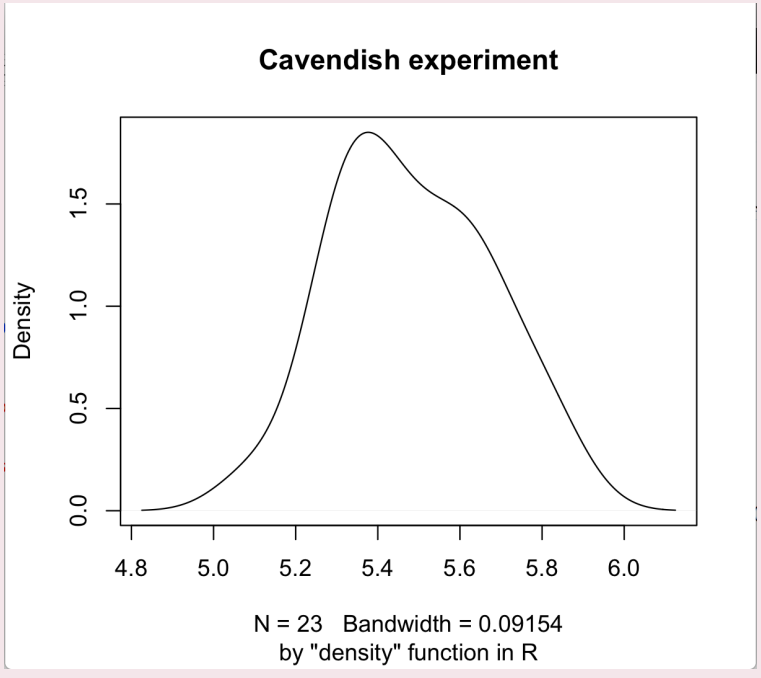
- Standard deviation of the posterior distribution is smaller than that of the prior distribution, that is

$$SD(\mu|\boldsymbol{x}) < SD(\mu)$$

## Example 1.4

The 18th century physicist Henry Cavendish made 23 experimental determinations of the earth's density, and these data (in $g/cm^3$) are

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5.36 | 5.29 | 5.58 | 5.65 | 5.57 | 5.53 | 5.62 | 5.29 |
| 5.44 | 5.34 | 5.79 | 5.10 | 5.27 | 5.39 | 5.42 | 5.47 |
| 5.63 | 5.34 | 5.46 | 5.30 | 5.78 | 5.68 | 5.85 | |



## Example 1.4

- Cavendish asserts that the error standard deviation of these measurements is $0.2\,g/cm^3$
- Assume that they are normally distributed with mean equal to the true earth density $\mu$, that is, $X_i|\mu \sim N(\mu, 0.2^2)$, $i = 1, 2, \ldots, 23$
- Use a normal prior distribution for $\mu$ with mean $5.41\,g/cm^3$ and standard deviation $0.4\,g/cm^3$
- Derive the posterior distribution for $\mu$

## Solution

. . .

# Summary

- Small increase in mean from prior to posterior
- Large decrease in uncertainty from prior to posterior

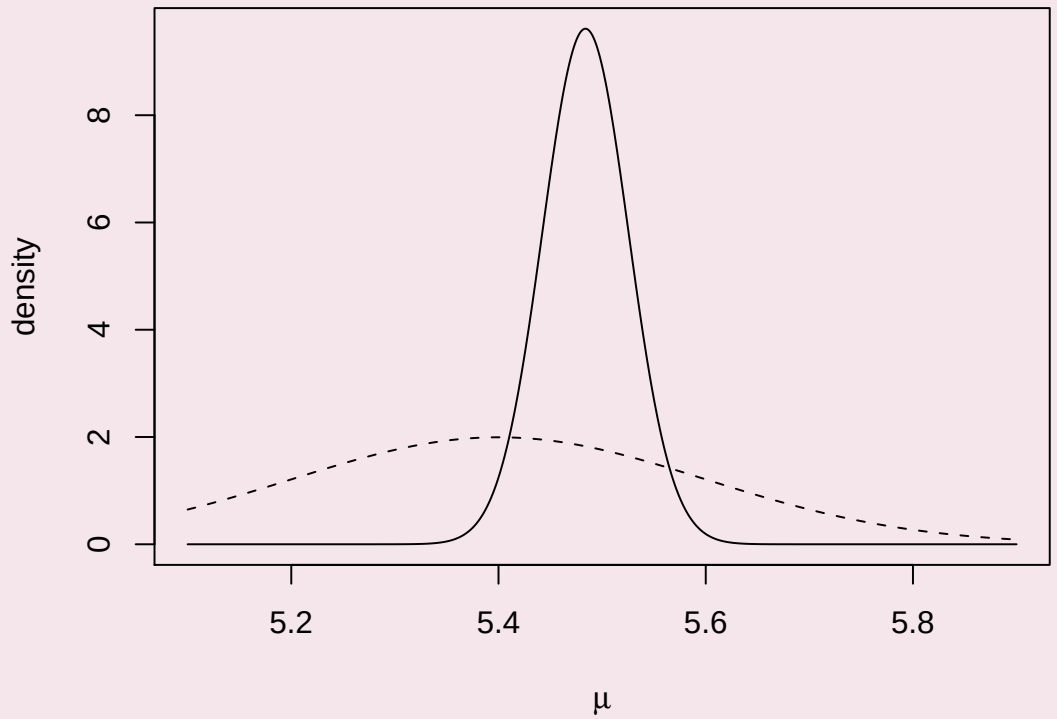



Figure: Prior (dashed) and posterior (solid) densities for the earth's density

# 1.2 Different levels of prior knowledge

## 1.2.1 Substantial Prior Knowledge

- We have *substantial prior information* for $\theta$ when the prior distribution *dominates* the likelihood function, that is $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$

- Difficulties:
  1. Intractability of mathematics in deriving the posterior
  2. Practical formulation of the prior distribution – coherently specifying prior beliefs in the form of a probability distribution is far from straightforward, let alone reconciling differences between experts!

## 1.2.2 Limited prior information

Pragmatic approach:

- Uniform distribution
- Choose a distribution which makes the Bayes updating from prior to posterior mathematically straightforward
- Use what prior information is available to determine the parameters of this distribution

Previous examples:

1. Poisson random sample, Gamma prior distribution $\longrightarrow$ Gamma posterior distribution
2. Normal random sample (known variance), Normal prior distribution $\longrightarrow$ Normal posterior distribution

In these examples, the prior distribution and the posterior distribution come from the same family

## Definition 1.1 (Conjugate priors)

Suppose that data $\boldsymbol{x}$ are to be observed with distribution $f(\boldsymbol{x}|\theta)$. A family $\mathfrak{F}$ of prior distributions for $\theta$ is said to be *conjugate* to $f(\boldsymbol{x}|\theta)$ if for every prior distribution $\pi(\theta) \in \mathfrak{F}$, the posterior distribution $\pi(\theta|\boldsymbol{x})$ is also in $\mathfrak{F}$.

## Comment

The conjugate family depends crucially on the model chosen for the data $\boldsymbol{x}$.

For example, the only family conjugate to the model "random sample from a Poisson distribution" is the Gamma family. Here, the likelihood is $f(\boldsymbol{x}|\theta) \propto \theta^{n\bar{x}} e^{-n\theta}$, $\theta > 0$. Therefore we need a family with density $f(\theta|\boldsymbol{a})$ and parameters $\boldsymbol{a}$ such that

$$f(\theta|\boldsymbol{A}) \propto f(\theta|\boldsymbol{a}) \times \theta^{n\bar{x}} e^{-n\theta}, \quad \theta > 0$$

$$\Longrightarrow \quad f(\theta|\boldsymbol{a}) \propto \theta^{a_1} e^{-a_2\theta}, \quad \theta > 0$$

that is, the Gamma family of distributions

## 1.2.3 Vague Prior Knowledge

- Have very little or no prior information about $\theta$
- Still must choose a prior distribution
- Sensible to choose a prior distribution which is not concentrated about any particular value, that is, one with a very large variance
- Most of the information about $\theta$ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Improper uniform distribution with support in an unbounded region
- Represent vague prior knowledge by using a prior distribution which is conjugate to the model for $\boldsymbol{x}$ and which has as large a variance as possible

## Example 1.5

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with $\tau$ known). Determine the posterior distribution assuming a vague prior for $\mu$.

## Solution

. . .

## Example 1.6

Suppose we have a random sample from an Poisson distribution, that is, $X_i|\theta \sim Po(\theta)$, $i = 1, 2, \ldots, n$ (independent). Determine the posterior distribution assuming a vague prior for $\theta$.

## Solution

- The conjugate prior distribution is a Gamma distribution
- The $Ga(g, h)$ distribution has mean $m = g/h$ and variance $v = g/h^2$
- Rearranging gives $g = m^2/v$ and $h = m/v$
- Clearly $g \to 0$ and $h \to 0$ as $v \to \infty$ (for fixed $m$)
- We have seen that, for this model, using a $Ga(g, h)$ prior distribution results in a $Ga(g + n\bar{x}, h + n)$ posterior distribution
- Therefore, taking a vague prior distribution will give a $Ga(n\bar{x}, n)$ posterior distribution

Note that the posterior mean is $\bar{x}$ (the likelihood mode) and that the posterior variance $\bar{x}/n \to 0$ and $n \to \infty$.

# 1.3 Asymptotic Posterior Distribution

## Background

- There are many asymptotic results in Statistics
- The **Central Limit Theorem** is a statement about the asymptotic distribution of $\bar{X}_n$ as the sample size $n \to \infty$, where the $X_i$ are i.i.d. with known mean $\mu$ and known (finite) variance $\sigma^2$
- Under different distributions of $X_i$, $\bar{X}_n$ has the same moments, $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$, but its distribution varies
- The CLT says that as $n \to \infty$, regardless of the distribution of $X_i$,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} N(0, 1)$$

## Theorem

Suppose we have a statistical model $f(\boldsymbol{x}|\theta)$ for data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, together with a prior distribution $\pi(\theta)$ for $\theta$. Then

$$\sqrt{J(\hat{\theta})}\,(\theta - \hat{\theta})|\boldsymbol{x} \xrightarrow{\mathcal{D}} N(0,1) \qquad \text{as } n \to \infty,$$

where $\hat{\theta}$ is the likelihood mode and $J(\theta)$ is the observed information

$$J(\theta) = -\frac{\partial^2}{\partial \theta^2} \log f(\boldsymbol{x}|\theta)$$

## Usage

In large samples,

$$\theta|\boldsymbol{x} \sim N\left(\hat{\theta}, J(\hat{\theta})^{-1}\right), \qquad \text{approximately.}$$

- In large samples ($n$ large), how the prior distribution is specified does not matter.

## Example 1.7

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with $\tau$ known). Determine the asymptotic posterior distribution for $\mu$.
Recall that

$$f(\boldsymbol{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\},$$

and therefore

$$\log f(\boldsymbol{x}|\mu) = \frac{n}{2}\log\tau - \frac{n}{2}\log(2\pi) - \frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Rightarrow \quad \frac{\partial}{\partial\mu}\log f(\boldsymbol{x}|\mu) = -\frac{\tau}{2}\times\sum_{i=1}^{n}-2(x_i-\mu) = \tau\sum_{i=1}^{n}(x_i-\mu) = n\tau(\bar{x}-\mu)$$

$$\Rightarrow \quad \frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu) = -n\tau \qquad \Rightarrow \quad J(\mu) = -\frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{x}|\mu) = n\tau.$$

## Solution

. . .

# 1.4 Bayesian inference

- The posterior distribution $\pi(\theta|\boldsymbol{x})$ summarises all our information about $\theta$ to date
- It can answer the questions: How to estimate the value of $\theta$, and what is the uncertainty of the estimator?

## 1.4.1 Estimation

### Point estimates

Many useful summaries, such as

- the mean: $E(\theta|\boldsymbol{x})$
- the mode: $Mode(\theta|\boldsymbol{x})$
- the median: $Median(\theta|\boldsymbol{x})$

## Interval Estimates

- A more useful summary of the posterior distribution is one which also reflects its variation
- A $100(1-\alpha)\%$ *Bayesian confidence interval* for $\theta$ is any region $C_\alpha$ that satisfies $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$
- If $\theta$ is continuous with posterior density $\pi(\theta|\boldsymbol{x})$ then

$$\int_{C_\alpha} \pi(\theta|\boldsymbol{x})\, d\theta = 1-\alpha$$

- The usual correction is made for discrete $\theta$, that is, we take the largest region $C_\alpha$ such that $Pr(\theta \in C_\alpha|\boldsymbol{x}) \leq 1-\alpha$
- Bayesian confidence intervals are sometimes called *credible regions* or *plausible regions*
- Clearly these intervals are not unique, since there will be many intervals with the correct probability coverage for a given posterior distribution

## Highest density intervals

- A $100(1-\alpha)\%$ *highest density interval* (HDI) for $\theta$ is the region

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

  where $\gamma$ is chosen so that $Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$
- It is a $100(1-\alpha)\%$ Bayesian confidence interval but only includes the most likely values of $\theta$
- This region is sometimes called a *most plausible Bayesian confidence interval*
- If the posterior distribution has many modes then it is possible that the HDI will be the union of several disjoint regions

## Highest density intervals

- If the posterior distribution is unimodal (has one mode) and symmetric about its mean then the HDI is an equi-tailed interval, that is, takes the form $C_\alpha = (a, b)$, where

$$Pr(\theta < a|\boldsymbol{x}) = Pr(\theta > b|\boldsymbol{x}) = \alpha/2$$

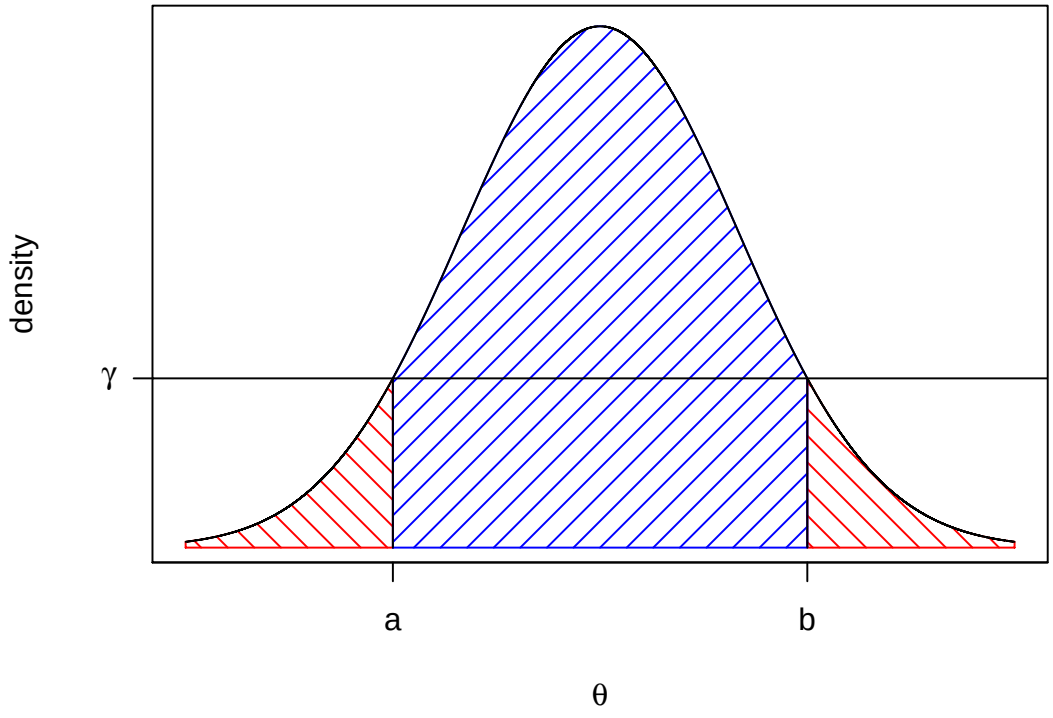



Figure: HDI for $\theta$

## Example 1.8

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $N(\mu, 1/\tau)$ distribution (where $\tau$ is known). We have seen that, assuming vague prior knowledge, the posterior distribution is $\mu|\boldsymbol{x} \sim N\{\bar{x}, 1/(n\tau)\}$. Determine the $100(1-\alpha)\%$ HDI for $\mu$.

## Solution

$\ldots$

## Comment

Note that this interval is numerically identical to the 95% frequentist confidence interval for the (population) mean of a normal random sample with known variance. However, the interpretation is very different.

## Interpretation of confidence intervals

$C_B$ is a 95% Bayesian confidence interval for $\theta$

$C_F$ is a 95% frequentist confidence interval for $\theta$

These intervals do not have the same interpretation:

- the probability that $C_B$ contains $\theta$ is 0.95
- the probability that $C_F$ contains $\theta$ is either 0 or 1
- the interval $C_F$ covers the true value $\theta$ on 95% of occasions — in repeated applications of the formula

## Example 1.9

Recall Example 1.1 on the number of cases of foodbourne botulism in England and Wales. The data were modelled as a random sample from a Poisson distribution with mean $\theta$. Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\boldsymbol{x} \sim Ga(8, 9)$, with density

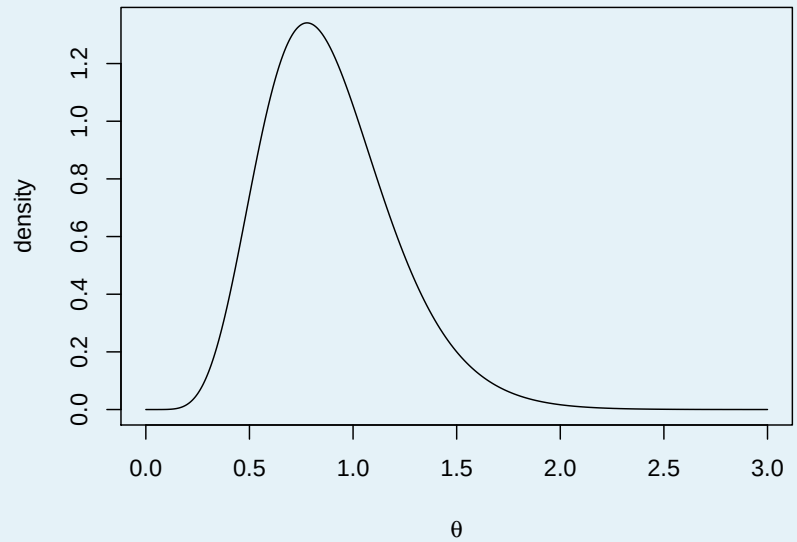



Figure: Posterior density for $\theta$

Determine the 95% HDI for $\theta$.

## Solution

. . .

## Simple way to calculate the HDI

- Use the R function `hdiGamma` in the package `nclbayes`
- It calculates the HDI for any Gamma distribution
- Here we use

```
library(nclbayes)
hdiGamma(p=0.95,a=8,b=9)
```

- The package also has functions
  - `hdiBeta` for the Beta distribution
  - `hdiInvchi` for the Inv-Chi distribution (introduced in Chapter 3)

# 1.4.2 Prediction

- Much of statistical inference (both Frequentist and Bayesian) is aimed towards making statements about a parameter $\theta$
- Often the inferences are used as a yardstick for similar future experiments
- For example, we may want to predict the outcome when the experiment is performed again

## Predicting the future

- There will be uncertainty about the future outcome of an experiment
- Suppose this future outcome $Y$ is described by a probability (density) function $f(y|\theta)$
- If $\theta$ were known, say $\theta_0$, then any prediction can do no better than one based on $f(y|\theta = \theta_0)$
- What if $\theta$ is unknown?

## Frequentist solution

- Get estimate $\hat{\theta}$ and use $f(y|\theta = \hat{\theta})$ but this ignores uncertainty on $\hat{\theta}$
- Better: use $E_{\hat{\theta}}\{f(y|\theta = \hat{\theta})\}$ to average over uncertainty on $\hat{\theta}$

## Bayesian solution

- Use the *predictive distribution*, with density

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

  when $\theta$ is a continuous quantity
- Notice this could be rewritten as

$$f(y|\boldsymbol{x}) = E_{\theta|\boldsymbol{x}}\{f(y|\theta)\}$$

- Uses $f(y|\theta)$ but weights each $\theta$ by our posterior beliefs

## Prediction interval

- Useful range of plausible values for the outcome of a future experiment
- Similar to a HDI interval
- A $100(1-\alpha)\%$ *prediction interval* for $Y$ is the region

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

where $\gamma$ is chosen so that $Pr(Y \in C_\alpha|\boldsymbol{x}) = 1-\alpha$

## Example 1.10

Recall Example 1.1 on the number of cases of foodbourne botulism in England and Wales.

The data for 1998–2005 were modelled as a random sample from a Poisson distribution with mean $\theta$.

Using a $Ga(2, 1)$ prior distribution, we found the posterior distribution to be $\theta|\boldsymbol{x} \sim Ga(8, 9)$.

Determine the predictive distribution for the number of cases for the following year (2006).

## Solution

. . .

## Comments

- This predictive probability function is related to that of a negative binomial distribution
- If $Z \sim NegBin(r, p)$ then

$$Pr(Z = z) = \binom{z-1}{r-1} p^r (1-p)^{z-r}, \quad z = r, r+1, \ldots$$

and so $W = Z - r$ has probability function

$$Pr(W = w) = Pr(Z = w+r) = \binom{w+r-1}{r-1} p^r (1-p)^w, \quad w = 0, 1, \ldots$$

- This is the same probability function as our predictive probability function, with $r = 8$ and $p = 0.9$
- Therefore $Y|\boldsymbol{x} \sim NegBin(8, 0.9) - 8$
- Note that, unfortunately, `R` also calls the distribution of $W$ a negative binomial distribution with parameters $r$ and $p$: `dnbinom(r,p)`

- To distinguish between this distribution and the $NegBin(r, p)$ distribution used above, we shall denote the distribution of $W$ as a $NegBin_{\mathrm{R}}(r, p)$ distribution – it has mean $r(1-p)/p$ and variance $r(1-p)/p^2$
- Thus $Y|\boldsymbol{x} \sim NegBin_{\mathrm{R}}(8, 0.9)$

## Comparison between predictive and naive predictive distributions

- Can compare the predictive distribution $Y|\boldsymbol{x}$ with a naive predictive $Y|\theta = \hat{\theta} \sim Po(0.75)$ where

$$f(y|\theta = \hat{\theta}) = \frac{0.75^y \, e^{-0.75}}{y!}, \quad y = 0, 1, \ldots \ .$$

- Probability functions:

| $y$ | correct $f(y\|\boldsymbol{x})$ | naive $f(y\|\theta = \hat{\theta})$ |
|---|---|---|
| 0 | 0.430 | 0.472 |
| 1 | 0.344 | 0.354 |
| 2 | 0.155 | 0.133 |
| 3 | 0.052 | 0.033 |
| 4 | 0.014 | 0.006 |
| 5 | 0.003 | 0.001 |
| $\geq 6$ | 0.005 | 0.002 |

- The naive predictive distribution is a predictive distribution which uses a degenerate posterior distribution $\pi^*(\theta|\boldsymbol{x})$
- Here $Pr_{\pi^*}(\theta = 0.75|\boldsymbol{x}) = 1$ and standard deviation $SD_{\pi^*}(\theta|\boldsymbol{x}) = 0$
- The correct posterior standard deviation of $\theta$ is $SD_{\pi}(\theta|\boldsymbol{x}) = \sqrt{8}/9 = 0.314$
- Using a degenerate posterior distribution results in the naive predictive distribution having too small a standard deviation:

$$SD(Y|x = 1) = \begin{cases} 0.994 & \text{using the correct } \pi(\theta|\boldsymbol{x}) \\ 0.866 & \text{using the naive } \pi^*(\theta|\boldsymbol{x}), \end{cases}$$

  these values being calculated from $NegBin_{\mathrm{R}}(8, 0.9)$ and $Po(0.75)$ distributions
- $\{0, 1, 2\}$ is a 92.9% prediction set/interval
- $\{0, 1, 2\}$ is a 95.9% prediction set/interval using the the more "optimistic" naive predictive distribution

## Predictive distribution (general case)

- Generally requires calculation of a non-trivial integral (or sum)

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

- Easier method available when using a conjugate prior distribution
- Suppose $\theta$ is a continuous quantity and $\boldsymbol{X}$ and $Y$ are independent given $\theta$
- Using Bayes Theorem, the posterior distribution for $\theta$ given $\boldsymbol{x}$ and $y$ is

$$\begin{aligned}\pi(\theta|\boldsymbol{x}, y) &= \frac{\pi(\theta) f(\boldsymbol{x}, y|\theta)}{f(\boldsymbol{x}, y)} \\ &= \frac{\pi(\theta) f(\boldsymbol{x}|\theta) f(y|\theta)}{f(\boldsymbol{x}) f(y|\boldsymbol{x})} \qquad \text{since } \boldsymbol{X} \text{ and } Y \text{ are indep given } \theta \\ &= \frac{\pi(\theta|\boldsymbol{x})\, f(y|\theta)}{f(y|\boldsymbol{x})}.\end{aligned}$$

Rearranging, we obtain ...

## Candidate's formula

- The predictive p(d)f is

$$f(y|\boldsymbol{x}) = \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)}$$

- The RHS looks as if it depends on $\theta$ but it doesn't: all terms in $\theta$ cancel
- For this formula to be useful, we have to be able to work out $\theta|\boldsymbol{x}$ and $\theta|\boldsymbol{x}, y$ fairly easily
- This is the case when using conjugate priors

## Example 1.11

Rework Example 1.10 using Candidate's formula to determine the number of cases in 2006.

## Solution

. . .

# 1.5 Mixture Prior Distributions

Sometimes prior beliefs cannot be adequately represented by a simple distribution, for example, a normal distribution or a beta distribution. In such cases, mixtures of distributions can be useful.

## Example 1.12

Investigations into infants suffering from severe *idiopathic respiratory distress syndrome* have shown that whether the infant survives may be related to their weight at birth. Suppose that the distribution of birth weights (in kg) of infants who survive is a normal $N(2.3, 0.52^2)$ distribution and that of infants who die is a normal $N(1.7, 0.66^2)$ distribution. Also the proportion of infants that survive is 0.6. What is the distribution of birth weights of infants suffering from this syndrome?

## Solution

. . .

This distribution is a mixture of two normal distributions

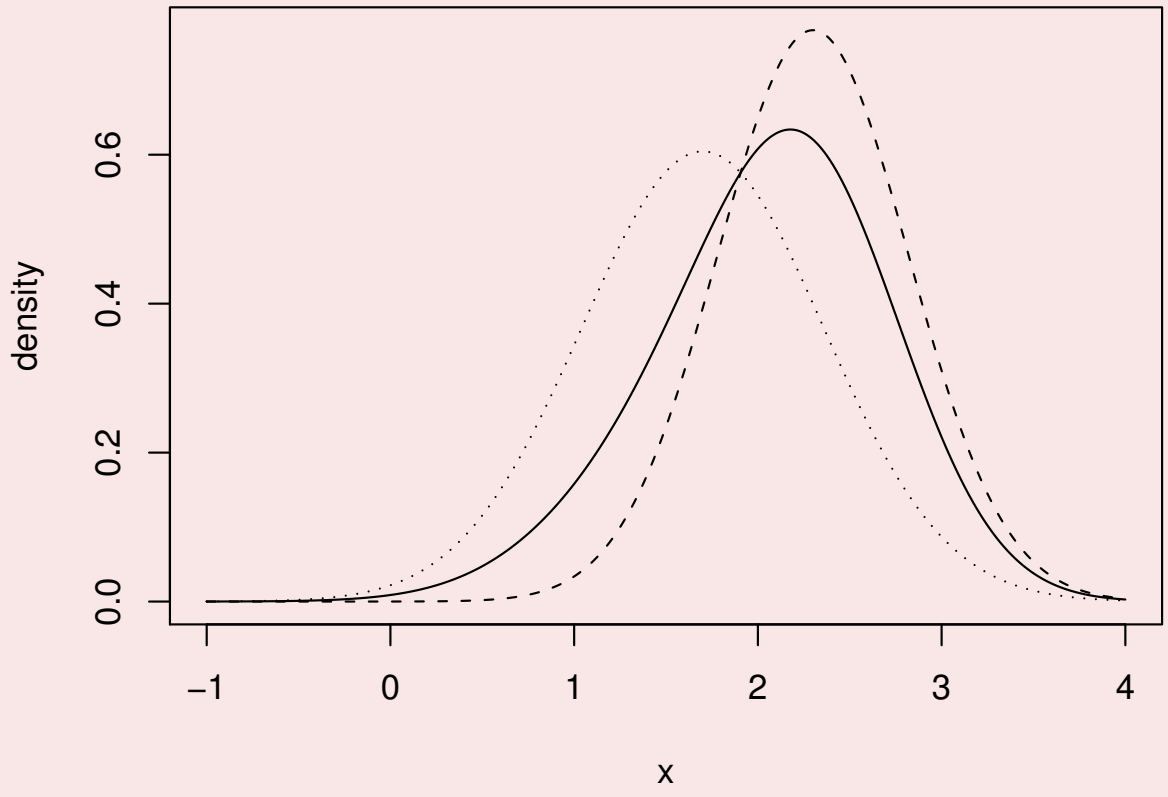



Figure: Plot of the mixture density (solid) with its component densities (survive – dashed; die – dotted)

## Definition 1.2 (Mixture distribution)

A *mixture* of the distributions $\pi_i(\theta)$ with *weights* $p_i$ $(i = 1, 2, \ldots, m)$ has probability (density) function

$$\pi(\theta) = \sum_{i=1}^{m} p_i \pi_i(\theta) \tag{1.11}$$

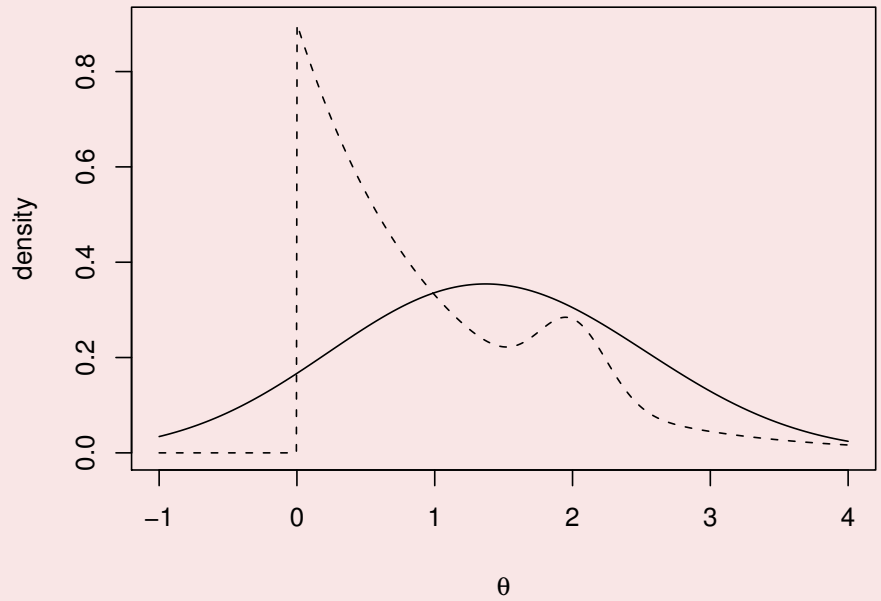



Figure: Plot of two mixture densities: solid is $0.6N(1, 1) + 0.4N(2, 1)$; dashed is $0.9Exp(1) + 0.1N(2, 0.25^2)$

## Properties of mixture distributions

In order for a mixture distribution to be proper, we must have

$$\begin{aligned} 1 &= \int_{\Theta} \pi(\theta)\, d\theta \\ &= \int_{\Theta} \sum_{i=1}^{m} p_i \pi_i(\theta)\, d\theta \\ &= \sum_{i=1}^{m} p_i \int_{\Theta} \pi_i(\theta)\, d\theta \\ &= \sum_{i=1}^{m} p_i, \end{aligned}$$

that is, the sum of the weights must be one

## Mean and Variance

Suppose the mean and variance of the distribution for $\theta$ in component $i$ are

$$E_i(\theta) = \int_\Theta \theta\, \pi_i(\theta)\, d\theta \quad \text{and} \quad Var_i(\theta) = \int_\Theta \{\theta - E_i(\theta)\}^2\, \pi_i(\theta)\, d\theta$$

Then

$$E(\theta) = \sum_{i=1}^{m} p_i E_i(\theta), \tag{1.12}$$

$$\begin{aligned} E(\theta^2) &= \sum_{i=1}^{m} p_i E_i(\theta^2) \\ &= \sum_{i=1}^{m} p_i \left\{ Var_i(\theta) + E_i(\theta)^2 \right\} \end{aligned} \tag{1.13}$$

and use $Var(\theta) = E(\theta^2) - E(\theta)^2$

## Posterior distribution

Using Bayes Theorem gives

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &= \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})} \\ &= \sum_{i=1}^{m} \frac{p_i \pi_i(\theta)\, f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})}\end{aligned} \tag{1.14}$$

where $f(\boldsymbol{x})$ is a constant with respect to $\theta$.

## Component posterior distributions

If the prior distribution were $\pi_i(\theta)$ (instead of the mixture distribution) then, using Bayes Theorem, the posterior distribution would be

$$\pi_i(\theta|\boldsymbol{x}) = \frac{\pi_i(\theta)\, f(\boldsymbol{x}|\theta)}{f_i(\boldsymbol{x})}$$

where $f_i(\boldsymbol{x})$ $i = 1, 2, \ldots, m$ are constants with respect to $\theta$.

Substituting this in to (1.14) gives

$$\pi(\theta|\boldsymbol{x}) = \sum_{i=1}^{m} \frac{p_i f_i(\boldsymbol{x})}{f(\boldsymbol{x})}\, \pi_i(\theta|\boldsymbol{x}).$$

Thus the posterior distribution is a mixture distribution of component distributions $\pi_i(\theta|\boldsymbol{x})$ with weights $p_i^* = p_i f_i(\boldsymbol{x})/f(\boldsymbol{x})$. Now

$$\sum_{i=1}^{m} p_i^* = 1 \quad \Rightarrow \quad \sum_{i=1}^{m} \frac{p_i f_i(\boldsymbol{x})}{f(\boldsymbol{x})} = 1 \quad \Rightarrow \quad f(\boldsymbol{x}) = \sum_{i=1}^{m} p_i f_i(\boldsymbol{x})$$

and so

$$p_i^* = \frac{p_i f_i(\boldsymbol{x})}{\sum_{j=1}^{m} p_j f_j(\boldsymbol{x})}, \qquad i = 1, 2, \ldots, m.$$

# Summary

- Likelihood: $f(\boldsymbol{x}|\theta)$
- Prior: $\pi(\theta) = \sum_{i=1}^{m} p_i \pi_i(\theta)$
- Posterior: $\pi(\theta|\boldsymbol{x}) = \sum_{i=1}^{m} p_i^* \pi_i(\theta|\boldsymbol{x})$, where

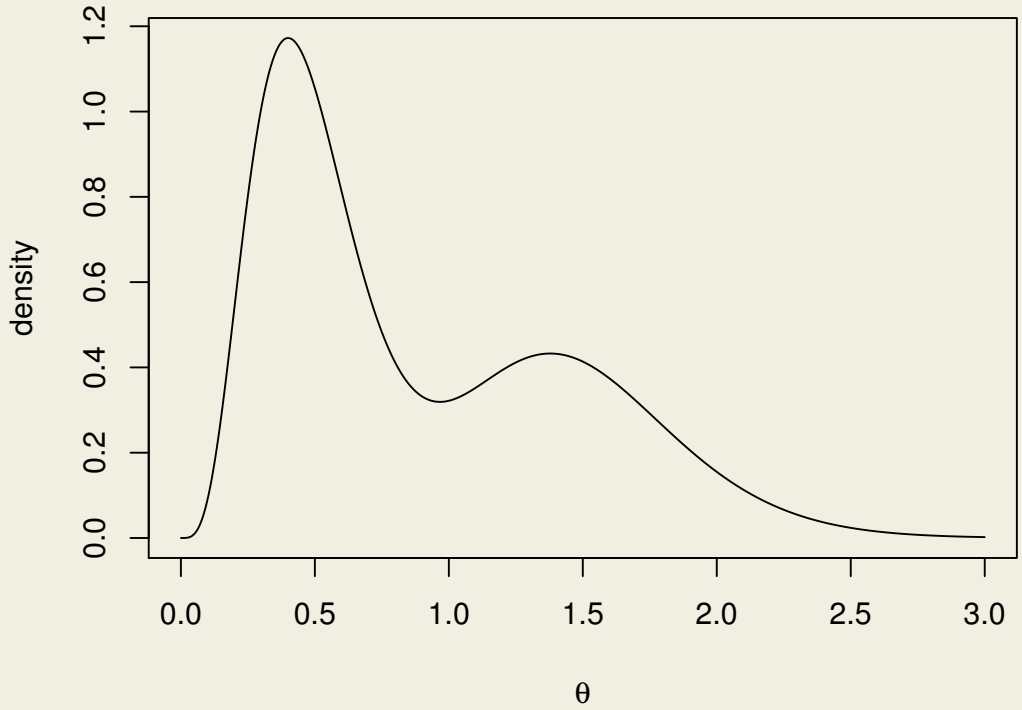
$$\pi_i(\theta) \xrightarrow{\boldsymbol{x}} \pi_i(\theta|\boldsymbol{x}) \qquad \text{and} \qquad p_i^* = \frac{p_i f_i(\boldsymbol{x})}{\sum_{j=1}^{m} p_j f_j(\boldsymbol{x})}$$

## Example 1.13

- Model: $X_j|\mu \sim Exp(\theta)$, $j = 1, 2, \ldots, 20$ (independent)
- Prior: mixture distribution with density

$$\pi(\theta) = 0.6\, Ga(5, 10) + 0.4\, Ga(15, 10)$$

Here the component distributions are $\pi_1(\theta) = Ga(5, 10)$ and $\pi_2(\theta) = Ga(15, 10)$, with weights $p_1 = 0.6$ and $p_2 = 0.4$



## Component posterior distributions (General case)

- Model: $X_j|\theta \sim Exp(\theta)$, $j = 1, 2, \ldots, n$ (independent)
- Prior: $\theta \sim Ga(g_i, h_i)$
- Data: observe $\boldsymbol{x}$
- Posterior: $\theta|\boldsymbol{x} \sim Ga(g_i + n, h_i + n\bar{x})$

## In this example ($n = 20$)

- Component priors:

$$\pi_1(\theta) = Ga(5, 10) \quad \text{and} \quad \pi_2(\theta) = Ga(15, 10)$$

- Component posteriors:

$$\pi_1(\theta|\boldsymbol{x}) = Ga(25, 10 + 20\bar{x}) \quad \text{and} \quad \pi_2(\theta|\boldsymbol{x}) = Ga(35, 10 + 20\bar{x})$$

## Weights

- We have

$$p_1^* = \frac{0.6 f_1(\boldsymbol{x})}{0.6 f_1(\boldsymbol{x}) + 0.4 f_2(\boldsymbol{x})} \qquad \Rightarrow \qquad (p_1^*)^{-1} - 1 = \frac{0.4 f_2(\boldsymbol{x})}{0.6 f_1(\boldsymbol{x})}$$

- In general, the functions

$$f_i(\boldsymbol{x}) = \int_\Theta \pi_i(\theta)\, f(\boldsymbol{x}|\theta)\, d\theta$$

  are potentially complicated integrals (solved either analytically or numerically)
- However, as with Candidates formula, these calculations become much simpler when we have a conjugate prior distribution

- Rewriting Bayes Theorem, we obtain

$$f(\boldsymbol{x}) = \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{\pi(\theta|\boldsymbol{x})}$$

- So when the prior and posterior densities have a simple form (as they do when using a conjugate prior), it is straightforward to determine $f(\boldsymbol{x})$ using algebra rather than having to use calculus

## In this example . . .

- The gamma distribution is the conjugate prior distribution: random sample of size $n$ with mean $\bar{x}$ and $Ga(g, h)$ prior $\rightarrow$ $Ga(g + n, h + n\bar{x})$ posterior, and so

$$f(\boldsymbol{x}) = \frac{\pi(\theta)\, f(\boldsymbol{x}|\theta)}{\pi(\theta|\boldsymbol{x})}$$

$$= \frac{\dfrac{h^g \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^n e^{-n\bar{x}\theta}}{\dfrac{(h + n\bar{x})^{g+n} \theta^{g+n-1} e^{-(h+n\bar{x})\theta}}{\Gamma(g + n)}}$$

$$= \frac{h^g\, \Gamma(g + n)}{\Gamma(g)(h + n\bar{x})^{g+n}}$$

Notice that all terms in $\theta$ have cancelled

- Therefore

$$
\begin{aligned}
(p_1^*)^{-1} - 1 &= \frac{p_2 f_2(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x})} \\
&= \frac{0.4 \times 10^{15}\,\Gamma(35)}{\Gamma(15)(10+20\bar{x})^{35}} \Big/ \frac{0.6 \times 10^{5}\,\Gamma(25)}{\Gamma(5)(10+20\bar{x})^{25}} \\
&= \frac{2\Gamma(35)\Gamma(5)}{3\Gamma(25)\Gamma(15)(1+2\bar{x})^{10}} \\
&= \frac{611320}{7(1+2\bar{x})^{10}}
\end{aligned}
$$

$$
\Longrightarrow \quad p_1^* = \frac{1}{1 + \dfrac{611320}{7(1+2\bar{x})^{10}}}, \qquad p_2^* = 1 - p_1^*
$$

## Results for Gamma distribution

If $\theta \sim Ga(g, h)$ then $E(\theta) = \frac{g}{h}$ and

$$E(\theta^2) = Var(\theta) + E(\theta)^2 = \frac{g}{h^2} + \frac{g^2}{h^2} = \frac{g(g+1)}{h^2}$$

## Summaries for $\pi(\theta) = 0.6\, Ga(5, 10) + 0.4\, Ga(15, 10)$

- Mean: $E(\theta) = \sum_{i=1}^{2} p_i E_i(\theta) = 0.6 \times \frac{5}{10} + 0.4 \times \frac{15}{10} = 0.9$
- Second moment:

$$E(\theta^2) = \sum_{i=1}^{2} p_i E_i(\theta^2) = 0.6 \times \frac{5 \times 6}{10^2} + 0.4 \times \frac{15 \times 16}{10^2} = 1.14$$

- Variance: $Var(\theta) = E(\theta^2) - E(\theta)^2 = 1.14 - 0.9^2 = 0.33$
- Standard deviation: $SD(\theta) = \sqrt{Var(\theta)} = \sqrt{0.33} = 0.574$

## Posterior distribution

The posterior distribution is the mixture distribution

$$\frac{1}{1+\frac{611320}{7(1+2\bar{x})^{10}}} \times Ga(25, 10+20\bar{x}) + \left(1 - \frac{1}{1+\frac{611320}{7(1+2\bar{x})^{10}}}\right) \times Ga(35, 10+20\bar{x})$$

| $\bar{x}$ | $\hat{\theta} = 1/\bar{x}$ | Posterior mixture distribution | $E(\theta\|\boldsymbol{x})$ | $SD(\theta\|\boldsymbol{x})$ |
|---|---|---|---|---|
| No data | | $0.6\, Ga(5, 10) + 0.4\, Ga(15, 10)$ | 0.9 | 0.574 |
| 4 | 0.25 | $0.99997\, Ga(25, 90) + 0.00003\, Ga(35, 90)$ | 0.278 | 0.056 |
| 2 | 0.5 | $0.9911\, Ga(25, 50) + 0.0089\, Ga(35, 50)$ | 0.502 | 0.102 |
| 1.2 | 0.8 | $0.7027\, Ga(25, 34) + 0.2973\, Ga(35, 34)$ | 0.823 | 0.206 |
| 1 | 1.0 | $0.4034\, Ga(25, 30) + 0.5966\, Ga(35, 30)$ | 1.032 | 0.247 |
| 0.8 | 1.25 | $0.1392\, Ga(25, 26) + 0.8608\, Ga(35, 26)$ | 1.293 | 0.260 |
| 0.5 | 2.0 | $0.0116\, Ga(25, 20) + 0.9884\, Ga(35, 20)$ | 1.744 | 0.300 |

Table: Posterior distributions (with summaries) for various sample means $\bar{x}$

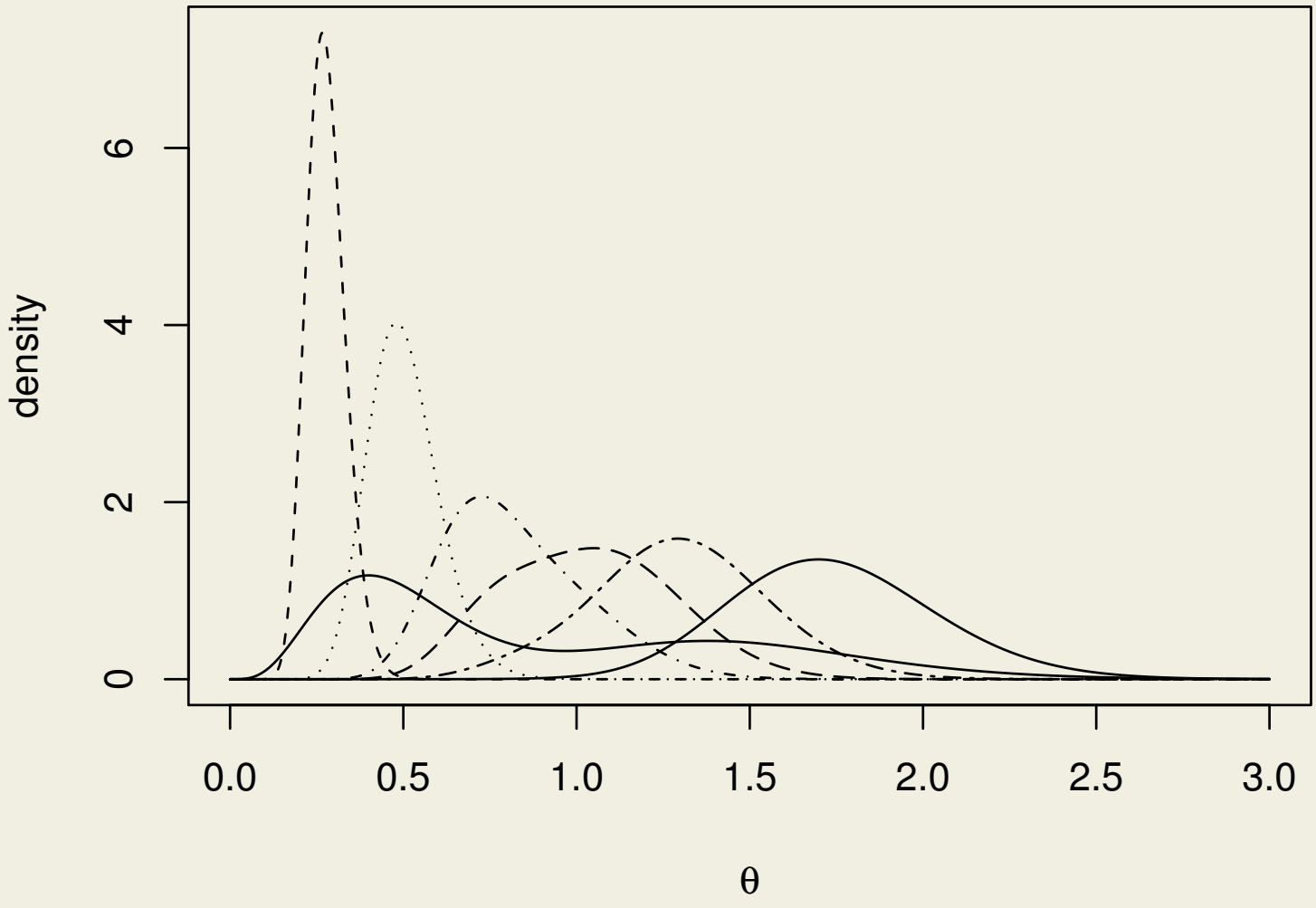



Figure: Plot of the prior distribution (dotted) and various posterior distributions

# Comments

- Likelihood mode is $1/\bar{x}$

  $\rightarrow$ large values of $\bar{x}$ indicate that $\theta$ is small and *vice versa*
- Component posterior means:

$$E_1(\theta|\boldsymbol{x}) = \frac{25}{10 + 20\bar{x}} \quad \text{and} \quad E_2(\theta|\boldsymbol{x}) = \frac{35}{10 + 20\bar{x}}$$

- Component 1 has smallest posterior mean

Recall $p_1^* = 1 \Big/ \left(1 + \frac{611320}{7(1 + 2\bar{x})^{10}}\right)$

- $p_1^*$ is increasing in $\bar{x}$
  Comment: as $\bar{x}$ increases, the posterior gives more weight to the component with the smallest mean
- $p_1^* \rightarrow 1$ as $\bar{x} \rightarrow \infty$

## Posterior summaries

- Mean:

$$
\begin{aligned}
E(\theta|\boldsymbol{x}) &= \sum_{i=1}^{2} p_i^* E_i(\theta|\boldsymbol{x}) \\
&= \frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}} \times \frac{25}{10+20\bar{x}} \\
&\qquad + \left(1 - \frac{1}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}}\right) \times \frac{35}{10+20\bar{x}} \\
&= \cdots \\
&= \frac{1}{2(1+2\bar{x})} \left\{ 7 - \frac{2}{1+\dfrac{611320}{7(1+2\bar{x})^{10}}} \right\}
\end{aligned}
$$

- Second moment: $E(\theta^2|\boldsymbol{x}) = \sum_{i=1}^{2} p_i^* E_i(\theta^2|\boldsymbol{x}) = \cdots$
- Variance: $Var(\theta|\boldsymbol{x}) = E(\theta^2|\boldsymbol{x}) - E(\theta|\boldsymbol{x})^2 = \cdots$
- Standard deviation: $SD(\theta|\boldsymbol{x}) = \sqrt{Var(\theta|\boldsymbol{x})} = \cdots$

# 1.6 Learning objectives

By the end of this chapter, you should be able to:

1. Determine the likelihood function using a random sample from **any** distribution
2. Combine this likelihood function with **any** prior distribution to obtain the posterior distribution
3. Name the posterior distribution if it is a "standard" distribution listed in these notes or on the exam paper – this list may well include distributions that are standard within the subject but which you have not met before. If the posterior is not a "standard" distribution then it is okay just to give its density (or probability function) up to a constant.
4. Do all the above for a particular data set or for a general case with random sample $x_1, \ldots, x_n$

5. Describe the different levels of prior information; determine and use conjugate priors and vague priors
6. Determine the asymptotic posterior distribution
7. Determine the predictive distribution, particularly when having a random sample from any distribution and a conjugate prior via Candidate's formula
8. Describe and calculate the confidence intervals, HDIs and prediction intervals
9. Calculate the mean and variance of a mixture distribution
10. Determine posterior distributions when the prior is a mixture of conjugate distributions, including component distributions and weights