MAS3902: Bayesian Inference

Dr. Lee Fawcett

School of Mathematics, Statistics & Physics

Semester 2, 2022/23

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** take place on Tuesdays (3pm, Herschel LT2) and Thursdays (9am, online – fortnightly only; 12pm, Herschel LT1)
- Two of these classes will be **lectures**, and the other session will be a **problems class/drop-in**
  - Lectures will *ordinarily* take place in the in-person slots in the lecture theatres
  - Problems classes/drop-ins will *ordinarily* take place in the online slot
- Due to outreach/recruitment activities, I will not be at work on Thursday 2nd February; Thursday 9th February and Thursday 11th May. Fear not! **Replacement lectures** have been booked:
  - Friday 3rd February (11am, Herschel LT2)
  - Friday 17th February (11am, Herschel LT2)
  - Friday 3rd March (11am, Herschel LT2)
- **Office hours** will be scheduled soon, but just come along and give me a knock!

Timetable and Administrative arrangements

- **Classes** take place on Tuesdays (3pm, Herschel LT2) and Thursdays (9am, online – fortnightly only; 12pm, Herschel LT1)
- Two of these classes will be **lectures**, and the other session will be a **problems class/drop-in**
  - Lectures will *ordinarily* take place in the in-person slots in the lecture theatres
  - Problems classes/drop-ins will *ordinarily* take place in the online slot
- Due to outreach/recruitment activities, I will not be at work on Thursday 2nd February; Thursday 9th February and Thursday 11th May. Fear not! **Replacement lectures** have been booked:
  - Friday 3rd February (11am, Herschel LT2)
  - Friday 17th February (11am, Herschel LT2)
  - Friday 3rd March (11am, Herschel LT2)
- **Office hours** will be scheduled soon, but just come along and give me a knock!

Timetable and Administrative arrangements

- **Classes** take place on Tuesdays (3pm, Herschel LT2) and Thursdays (9am, online – fortnightly only; 12pm, Herschel LT1)
- Two of these classes will be **lectures**, and the other session will be a **problems class/drop-in**
  - Lectures will *ordinarily* take place in the in-person slots in the lecture theatres
  - Problems classes/drop-ins will *ordinarily* take place in the online slot
- Due to outreach/recruitment activities, I will not be at work on Thursday 2nd February; Thursday 9th February and Thursday 11th May. Fear not! **Replacement lectures** have been booked:
  - Friday 3rd February (11am, Herschel LT2)
  - Friday 17th February (11am, Herschel LT2)
  - Friday 3rd March (11am, Herschel LT2)
- **Office hours** will be scheduled soon, but just come along and give me a knock!

Timetable and Administrative arrangements

- **Classes** take place on Tuesdays (3pm, Herschel LT2) and Thursdays (9am, online – fortnightly only; 12pm, Herschel LT1)
- Two of these classes will be **lectures**, and the other session will be a **problems class/drop-in**
  - Lectures will *ordinarily* take place in the in-person slots in the lecture theatres
  - Problems classes/drop-ins will *ordinarily* take place in the online slot
- Due to outreach/recruitment activities, I will not be at work on Thursday 2nd February; Thursday 9th February and Thursday 11th May. Fear not! **Replacement lectures** have been booked:
  - Friday 3rd February (11am, Herschel LT2)
  - Friday 17th February (11am, Herschel LT2)
  - Friday 3rd March (11am, Herschel LT2)
- **Office hours** will be scheduled soon, but just come along and give me a knock!

Timetable and Administrative arrangements

- **Classes** take place on Tuesdays (3pm, Herschel LT2) and Thursdays (9am, online – fortnightly only; 12pm, Herschel LT1)
- Two of these classes will be **lectures**, and the other session will be a **problems class/drop-in**
  - Lectures will *ordinarily* take place in the in-person slots in the lecture theatres
  - Problems classes/drop-ins will *ordinarily* take place in the online slot
- Due to outreach/recruitment activities, I will not be at work on Thursday 2nd February; Thursday 9th February and Thursday 11th May. Fear not! **Replacement lectures** have been booked:
  - Friday 3rd February (11am, Herschel LT2)
  - Friday 17th February (11am, Herschel LT2)
  - Friday 3rd March (11am, Herschel LT2)
- **Office hours** will be scheduled soon, but just come along and give me a knock!

Assessment

Assessment is by:

- 2-hour end-of-semester exam in May/June (**80%**)
- In course assessment (**20%**), comprising four short written assignments:

Date set	Deadline	Marks/weighting
Friday 10th February	4pm, Friday 24th February	20/5%
Friday 24th February	4pm, Friday 10th March	20/5%
Friday 10th March	4pm, Friday 24th March	20/5%
Friday 24th March	4pm, Friday 5th May	20/5%

Assessment

Assessment is by:

- 2-hour end-of-semester exam in May/June (**80%**)
- In course assessment (**20%**), comprising four short written assignments:

Date set	Deadline	Marks/weighting
Friday 10th February	4pm, Friday 24th February	20/5%
Friday 24th February	4pm, Friday 10th March	20/5%
Friday 10th March	4pm, Friday 24th March	20/5%
Friday 24th March	4pm, Friday 5th May	20/5%

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 available for use in lectures – you can 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 Canvas page after each topic is finished, along with any other course material – such as problems sheets, model solutions to assignment questions, supplementary handouts etc.

Other stuff

- Notes (with gaps) will be available for use in lectures – you can 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 Canvas page after each topic is finished, along with any other course material – such as problems sheets, model solutions to assignment questions, supplementary handouts etc.

Other stuff

- Notes (with gaps) will be available for use in lectures – you can 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 Canvas page 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

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 θ

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

considered as a function of θ 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 θ

Bayes Theorem

The posterior probability (density) function for θ 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 θ, 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**

Bayes Theorem

The posterior probability (density) function for θ 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 θ, 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 θ in time
- This means the data are from a Poisson process and so are a random sample from a Poisson distribution with rate θ
- An expert was consulted to construct a suitable prior distribution for the rate θ

Example 1.1

(a) What might be an appropriate probability model for the rate θ? Why is this termed the *prior distribution*?

(b) Suppose the epidemiologist tells you that she expects around two cases of foodbourne botulism per year in England and Wales. After further discussions with the epidemiologist, you are also able to elicit that $SD(\theta) \approx \sqrt{2}$. Using your answer to part (a) find your full prior distribution for θ, and write down its density.

(c) Determine the posterior distribution for θ.

Solution

. . . Solution...

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: $\theta|\boldsymbol{x} \sim Ga(8, 9)$

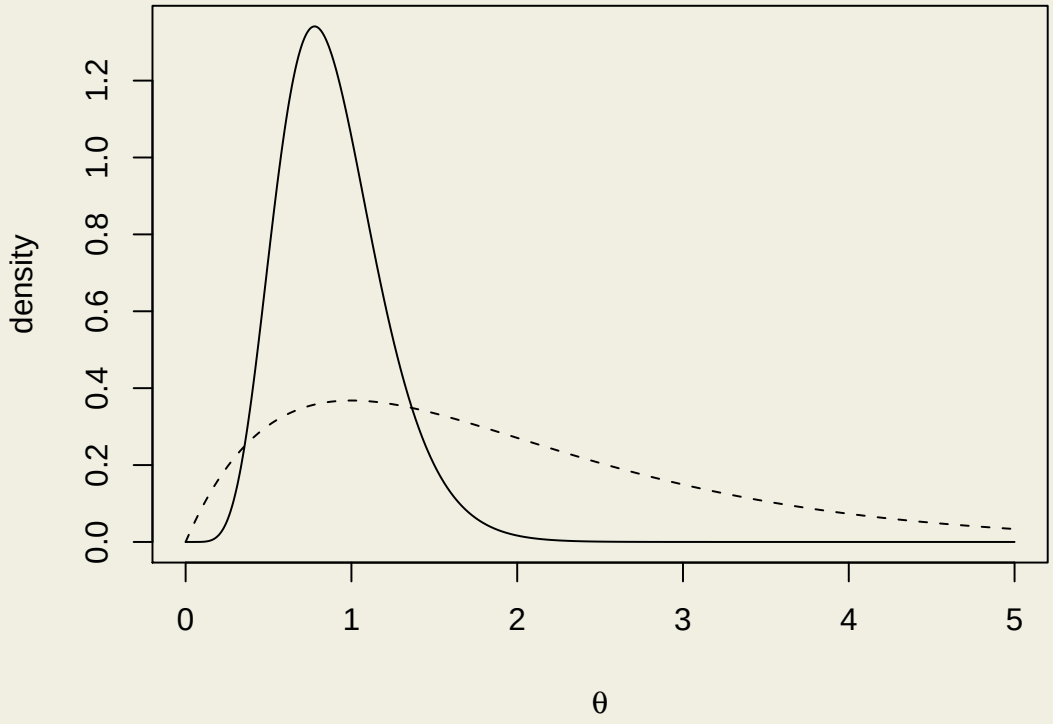


Figure: Prior (dashed) and posterior (solid) densities for θ

	Prior	Likelihood	Posterior
$Mode(\theta)$	1.00	0.75	0.78
$E(\theta)$	2.00	–	0.89
$SD(\theta)$	1.41	–	0.31

Table: Changes in beliefs about θ

- 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 θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{1.1}$$

Determine the posterior distribution for θ.

Solution

...

(1.2)

(1.3)

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 (1.1)	Likelihood (1.2)	Posterior (1.3)
$Mode(\theta)$	$(g-1)/h$	$\bar{x}$	$(g + n\bar{x} - 1)/(h + n)$
$E(\theta)$	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 θ

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 μ but known precision τ: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent).

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

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\}. \tag{1.4}$$

Determine the posterior distribution for μ.

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 μ.

Solution

. . .

(1.5)

(1.6)

(1.7)

Summary

- Model: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent), with τ 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 (1.4)	Likelihood (1.5)	Posterior (1.7)
$Mode(\mu)$	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 μ

- 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

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

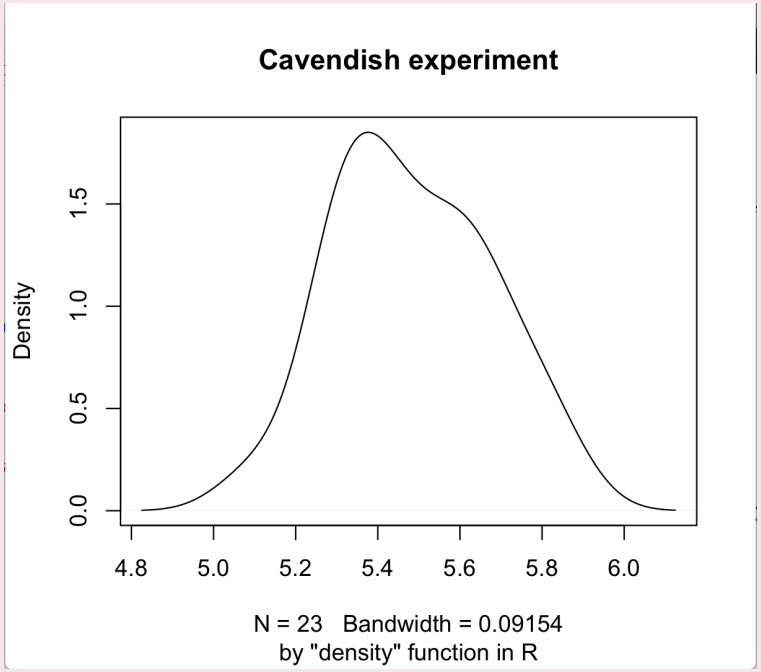
- Standard deviation of the posterior distribution is smaller than that of the prior distribution, that is

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

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 shown below.

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 μ, that is, $X_i|\mu \sim N(\mu, 0.2^2)$, $i = 1, 2, \ldots, 23$
- Use a normal prior distribution for μ with mean $5.41\,g/cm^3$ and standard deviation $0.4\,g/cm^3$
- Derive the posterior distribution for μ

Solution

...

Summary

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

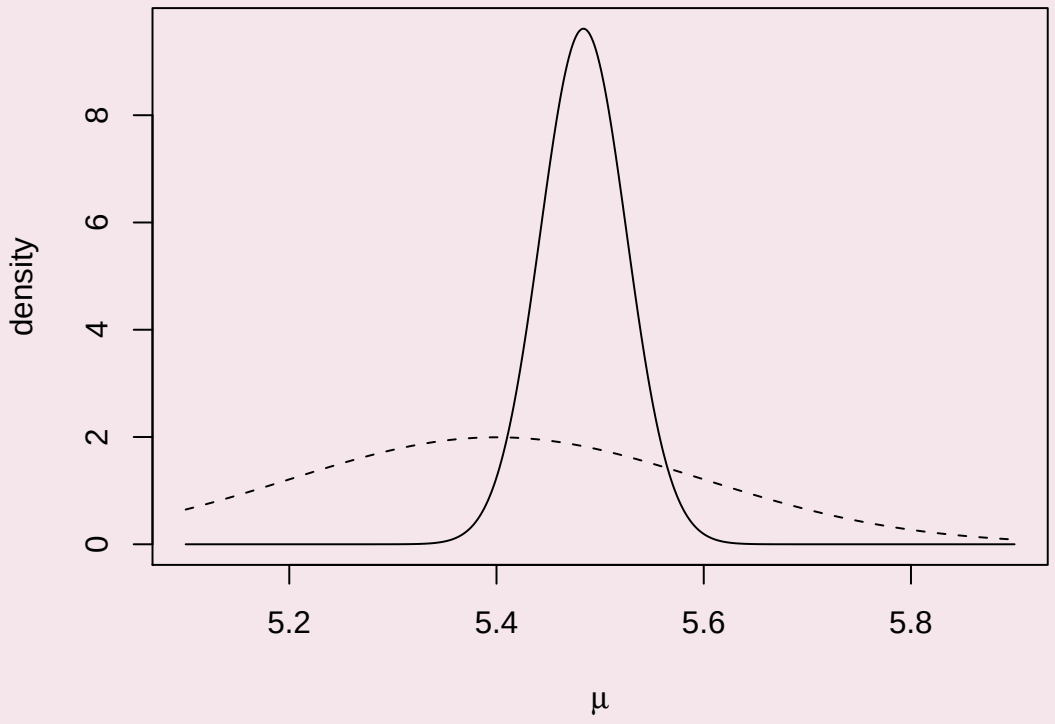


Figure: Prior (dashed) and posterior (solid) densities for the earth's density

Summary

- The actual mean density of the earth is $5.515\,g/cm^3$ (Wikipedia)
- We can determine the (posterior) probability that the mean density is within 0.1 of this value as follows...
- The posterior distribution is $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and so

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9510,$$

calculated using the R command

```
pnorm(5.615,5.484,0.0415)-pnorm(5.415,5.484,0.0415)
```

Summary

- The actual mean density of the earth is $5.515\,g/cm^3$ (Wikipedia)
- We can determine the (posterior) probability that the mean density is within 0.1 of this value as follows...
- The posterior distribution is $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and so

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9510,$$

calculated using the R command

```
pnorm(5.615,5.484,0.0415)-pnorm(5.415,5.484,0.0415)
```

Summary

- The actual mean density of the earth is $5.515\,g/cm^3$ (Wikipedia)
- We can determine the (posterior) probability that the mean density is within 0.1 of this value as follows...
- The posterior distribution is $\mu|\boldsymbol{x} \sim N(5.484, 0.0415^2)$ and so

$$Pr(5.415 < \mu < 5.615|\boldsymbol{x}) = 0.9510,$$

calculated using the R command

```
pnorm(5.615,5.484,0.0415)-pnorm(5.415,5.484,0.0415)
```

Summary

- Without the data, the only basis for determining the earth's density is via the prior distribution
- Here the prior distribution is $\mu \sim N(5.4, 0.4^2)$ and so the (prior) probability that the mean density is within 0.1 of the (now known) true value is

$$Pr(5.415 < \mu < 5.615) = 0.1896,$$

calculated using the R command

```
pnorm(5.615,5.4,0.4)-pnorm(5.415,5.4,0.4)
```

Summary

- Without the data, the only basis for determining the earth's density is via the prior distribution
- Here the prior distribution is $\mu \sim N(5.4, 0.4^2)$ and so the (prior) probability that the mean density is within 0.1 of the (now known) true value is

$$Pr(5.415 < \mu < 5.615) = 0.1896,$$

calculated using the R command

```
pnorm(5.615,5.4,0.4)-pnorm(5.415,5.4,0.4)
```

1.2 Different levels of prior knowledge

1.2.1 Substantial Prior Knowledge

- We have *substantial prior information* for θ 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:

- 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

1.2.2 Limited prior information

Pragmatic approach:

- 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 θ 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

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

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

that is, the Gamma family of distributions

1.2.3 Vague prior knowledge/prior ignorance

- Have very little or no prior information about θ
- 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 θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Vague prior: 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
- Prior ignorance: Uniform prior? Jeffreys prior?

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right], \tag{1.8}$$

and then:

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.9}$$

1.2.3 Vague prior knowledge/prior ignorance

- Have very little or no prior information about θ
- 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 θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Vague prior: 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
- Prior ignorance: Uniform prior? Jeffreys prior?

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right], \tag{1.8}$$

and then:

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.9}$$

1.2.3 Vague prior knowledge/prior ignorance

- Have very little or no prior information about θ
- 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 θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Vague prior: 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
- Prior ignorance: Uniform prior? Jeffreys prior?

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right], \tag{1.8}$$

and then:

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.9}$$

1.2.3 Vague prior knowledge/prior ignorance

- Have very little or no prior information about θ
- 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 θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Vague prior: 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
- Prior ignorance: Uniform prior? Jeffreys prior?

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right], \tag{1.8}$$

and then:

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.9}$$

1.2.3 Vague prior knowledge/prior ignorance

- Have very little or no prior information about θ
- 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 θ will be passed through to the posterior distribution via the data, and so we have $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Vague prior: 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
- Prior ignorance: Uniform prior? Jeffreys prior?

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right], \tag{1.8}$$

and then:

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.9}$$

Example 1.5

In constructing the prior in Example 1.1, did we assume (i) substantial prior knowledge; (ii) limited prior knowledge; or (iii) prior ignorance?

Solution

...

Example 1.6

Suppose the parameter θ represents a probability. What might be the simplest prior to use here, if we are prior ignorant?

Solution

...

Example 1.7

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

Solution

...

Example 1.8

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 θ.

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 $n \to \infty$.

Example 1.9

Suppose we have a random sample from an exponential distribution, that is, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent). Determine the Jeffreys prior for this model.

Solution

$\cdots$

What have we done so far?

1. **Posterior $\propto$ Prior $\times$ Likelihood**

   - **Example: Poisson distribution** with gamma prior (food poisoning, and generic/algebraic example)
   - **Example: Normal distribution** (known precision) with Normal prior for the mean (generic/algebraic example, and Cavendish experiment). (Memorise the results from the algebraic example!)
   - **Example: Binomial distribution** with beta prior (homework)
   - **Example: Exponential distribution** with Normal prior (homework)
   - Make sure you're familiar with simple functions in `R` for computing probabilities (e.g. `pnorm`), and make sure you can sketch/plot distributions!

What have we done so far?

1. **Posterior $\propto$ Prior $\times$ Likelihood**

   - **Example: Poisson distribution** with gamma prior (food poisoning, and generic/algebraic example)
   - **Example: Normal distribution** (known precision) with Normal prior for the mean (generic/algebraic example, and Cavendish experiment). (Memorise the results from the algebraic example!)
   - **Example: Binomial distribution** with beta prior (homework)
   - **Example: Exponential distribution** with Normal prior (homework)
   - Make sure you're familiar with simple functions in `R` for computing probabilities (e.g. `pnorm`), and make sure you can sketch/plot distributions!

What have we done so far?

1. **Posterior $\propto$ Prior $\times$ Likelihood**

- **Example: Poisson distribution** with gamma prior (food poisoning, and generic/algebraic example)
- **Example: Normal distribution** (known precision) with Normal prior for the mean (generic/algebraic example, and Cavendish experiment). (Memorise the results from the algebraic example!)
- **Example: Binomial distribution** with beta prior (homework)
- **Example: Exponential distribution** with Normal prior (homework)
- Make sure you're familiar with simple functions in `R` for computing probabilities (e.g. `pnorm`), and make sure you can sketch/plot distributions!

What have we done so far?

1. **Posterior $\propto$ Prior $\times$ Likelihood**

- **Example: Poisson distribution** with gamma prior (food poisoning, and generic/algebraic example)
- **Example: Normal distribution** (known precision) with Normal prior for the mean (generic/algebraic example, and Cavendish experiment). (Memorise the results from the algebraic example!)
- **Example: Binomial distribution** with beta prior (homework)
- **Example: Exponential distribution** with Normal prior (homework)
- Make sure you're familiar with simple functions in `R` for computing probabilities (e.g. `pnorm`), and make sure you can sketch/plot distributions!

What have we done so far?

1. **Posterior $\propto$ Prior $\times$ Likelihood**

- **Example: Poisson distribution** with gamma prior (food poisoning, and generic/algebraic example)
- **Example: Normal distribution** (known precision) with Normal prior for the mean (generic/algebraic example, and Cavendish experiment). (Memorise the results from the algebraic example!)
- **Example: Binomial distribution** with beta prior (homework)
- **Example: Exponential distribution** with Normal prior (homework)
- Make sure you're familiar with simple functions in `R` for computing probabilities (e.g. `pnorm`), and make sure you can sketch/plot distributions!

What have we done so far?

2. Different levels of prior information

- **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
- **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
- **Vague** prior: Use conjugate prior with maximised variance
- **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
- For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Lots of **Examples**: Examples 1.5–1.9 in notes

3. Posterior distribution when the sample is very large

4. Bayesian inference – confidence intervals and prediction

What have we done so far?

2. **Different levels of prior information**
   - **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
   - **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
   - **Vague** prior: Use conjugate prior with maximised variance
   - **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
   - For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
   - Lots of **Examples**: Examples 1.5–1.9 in notes

3. **Posterior distribution when the sample is very large**

4. **Bayesian inference – confidence intervals and prediction**

What have we done so far?

2. Different levels of prior information

- **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
- **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
- **Vague** prior: Use conjugate prior with maximised variance
- **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
- For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Lots of **Examples**: Examples 1.5–1.9 in notes

3. Posterior distribution when the sample is very large

4. Bayesian inference – confidence intervals and prediction

What have we done so far?

2. Different levels of prior information

- **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
- **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
- **Vague** prior: Use conjugate prior with maximised variance
- **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
- For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Lots of **Examples**: Examples 1.5–1.9 in notes

3. Posterior distribution when the sample is very large

4. Bayesian inference – confidence intervals and prediction

What have we done so far?

2. Different levels of prior information

- **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
- **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
- **Vague** prior: Use conjugate prior with maximised variance
- **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
- For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
- Lots of **Examples**: Examples 1.5–1.9 in notes

3. Posterior distribution when the sample is very large

4. Bayesian inference – confidence intervals and prediction

What have we done so far?

2. **Different levels of prior information**
   - **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
   - **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
   - **Vague** prior: Use conjugate prior with maximised variance
   - **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
   - For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
   - Lots of **Examples**: Examples 1.5–1.9 in notes

3. **Posterior distribution when the sample is very large**

4. **Bayesian inference – confidence intervals and prediction**

What have we done so far?

2. **Different levels of prior information**
   - **Substantial** prior information: $\pi(\theta|\boldsymbol{x}) \sim \pi(\theta)$, but often difficult to construct
   - **Limited** prior information: Use conjugate prior with parameters determined by the information provided by the expert
   - **Vague** prior: Use conjugate prior with maximised variance
   - **Ignorance** prior: Use a Jeffreys prior, or perhaps a $U(0,1)$?
   - For vague/ignorance priors, $\pi(\theta|\boldsymbol{x}) \sim f(\boldsymbol{x}|\theta)$
   - Lots of **Examples**: Examples 1.5–1.9 in notes

3. **Posterior distribution when the sample is very large**

4. **Bayesian inference – confidence intervals and prediction**

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 μ and known (finite) variance σ^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 θ. 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 1.10

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known). Determine the asymptotic posterior distribution for μ.
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 θ to date
- It can answer the questions: How to estimate the value of θ, 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 θ is any region C_α that satisfies $Pr(\theta \in C_\alpha | \boldsymbol{x}) = 1 - \alpha$
- If θ 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 θ, that is, we take the largest region C_α 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

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 θ is any region C_α that satisfies $Pr(\theta \in C_\alpha | \boldsymbol{x}) = 1 - \alpha$
- If θ 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 θ, that is, we take the largest region C_α 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 θ is the region

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

  where γ 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 θ
- 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

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

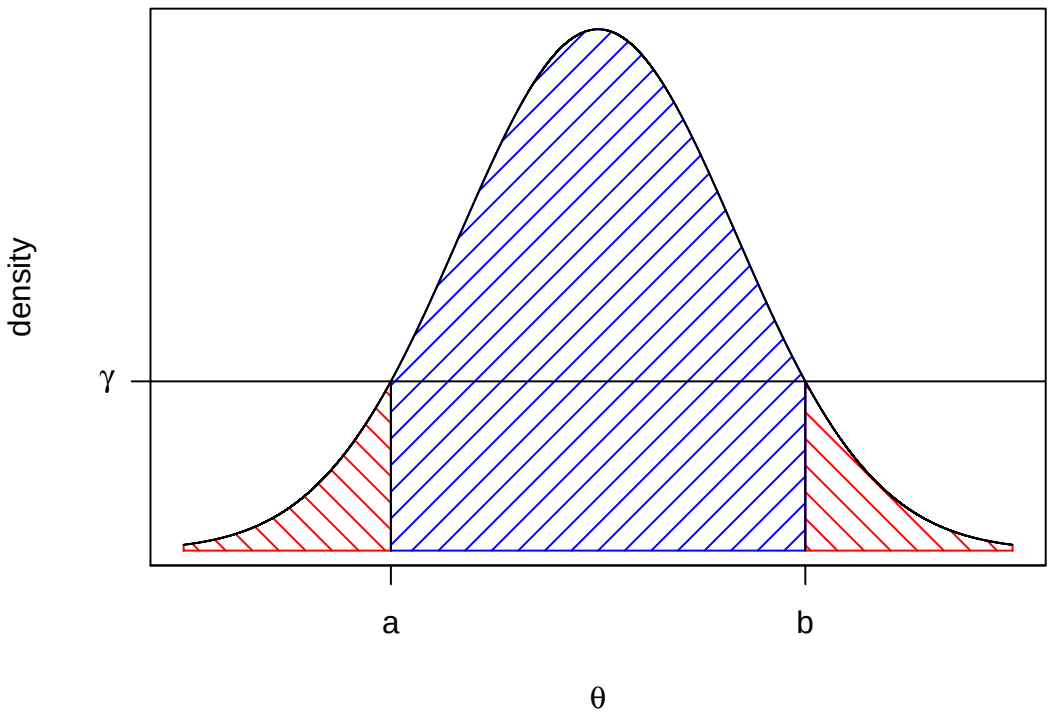


Figure: HDI for θ

Example 1.11

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $N(\mu, 1/\tau)$ distribution (where τ 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 μ.

Solution

...

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 θ

C_F is a 95% frequentist confidence interval for θ

These intervals do not have the same interpretation:

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

Example 1.12

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

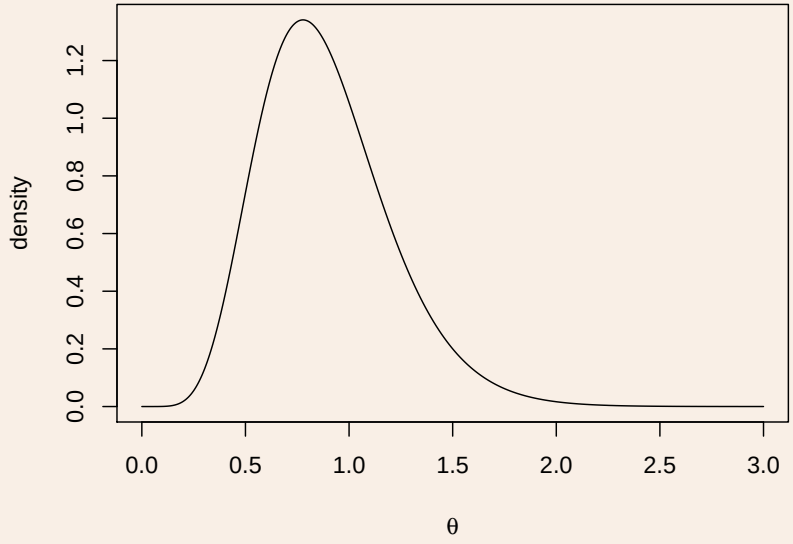


Figure: Posterior density for θ

Determine the 95% HDI for θ.

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 θ
- 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 θ were known, say θ_0, then any prediction can do no better than one based on $f(y|\theta = \theta_0)$
- What if θ 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 θ 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 θ 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 γ is chosen so that $Pr(Y \in C_\alpha|\boldsymbol{x}) = 1-\alpha$

Example 1.13

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 θ.

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 $Y|\mathbf{x} \sim NegBin(8, 0.9) - 8$
- Note that, unfortunately, `R` also calls the distribution of W a negative binomial distribution with parameters r and p: `dnbinom(r,p)`

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 $Y|\mathbf{x} \sim NegBin(8, 0.9) - 8$
- Note that, unfortunately, `R` also calls the distribution of W a negative binomial distribution with parameters r and p: `dnbinom(r,p)`

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)`

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 $NegBin_{\mathbb{R}}(r, p)$ distribution – it has mean $r(1-p)/p$ and variance $r(1-p)/p^2$
- Thus $Y|\boldsymbol{x} \sim NegBin_{\mathbb{R}}(8, 0.9)$

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

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
≥ 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 θ 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

- 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 θ 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 $NegBin_{\mathbb{R}}(8, 0.9)$ 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

- 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 θ 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 $NegBin_{\mathbb{R}}(8, 0.9)$ 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

- 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 θ 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

- 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 θ 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

- 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 θ 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 θ is a continuous quantity and $\boldsymbol{X}$ and Y are independent given θ
- Using Bayes Theorem, the posterior distribution for θ 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 ...

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 θ is a continuous quantity and $\boldsymbol{X}$ and Y are independent given θ
- Using Bayes Theorem, the posterior distribution for θ 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 ...

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 θ is a continuous quantity and $\boldsymbol{X}$ and Y are independent given θ
- Using Bayes Theorem, the posterior distribution for θ 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 ...

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 θ is a continuous quantity and $\boldsymbol{X}$ and Y are independent given θ
- Using Bayes Theorem, the posterior distribution for θ 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 θ but it doesn't: all terms in θ 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 1.14

Rework Example 1.**??** using Candidate's formula to determine the number of cases in 2006.

Solution

. . .