MAS2905: Probability and Bayesian Methods with R

Dr. Lee Fawcett

Semester 2, 2020—2021

Contact details

My name: Dr. 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`

Classes

As you now know, things are very different this academic year!

The aim was for you to have a blended learning expereince with some **present-in-person** (PiP) activities on campus and some online-only activities.

The academic year has not worked out like this!

In semester 2 we will continue with our online provision, until the situation improves enough for us to have some PiP activities scheduled back on campus.

Classes

Currently, this is what the timetable for our classes looks like:

- Fortnightly small group problems classes (PC). These will take place on Thursdays at the following times:
  - Group 1: 11–12
  - Group 2: 12–1
  - Group 3: 1:30–2:30
- Fortnightly **synchronous online** (SO) classes, alternating with the small group problems classes. These will take place on Fridays 12–1.
- As per semester 1, **non-synchronous online** (NSO) activities will be posted to Canvas every week, covering what would have ordinarily been covered in standard face-to-face lectures.

Schedule

Week	w/b	Activities	Date	Time	Deadline?
10	1 Feb	Problems class	Thurs 4 Feb	*—11:00/12:00/13:30: Check timetable—*	
		NSO activities	*Some pre-class reading/videos*		Wed 3 Feb
11	8 Feb	SO Lecture	Fri 12 Feb	12:00–13:00	—
		NSO activities	*Some pre-class reading/videos*		Thu 11 Feb
		Assignment 1 released	Thu 11 Feb	16:00	16:00 Thu 11 Mar
12	15 Feb	Problems class	Thurs 18 Feb	*—11:00/12:00/13:30 Check timetable—*	
		NSO activities	*Some pre-class reading/videos*		Wed 17 Feb
13	22 Feb	SO Lecture	Fri 26 Feb	12:00–13:00	—
		NSO activities	*Some pre-class reading/videos*		Thu 25 Feb
—	1 Mar	*—Buffer week, no new work scheduled—*			
14	8 Mar	Problems class	Thurs 11 Mar	*—11:00/12:00/13:30 Check timetable—*	
		NSO activities	*Some pre-class reading/videos*		Wed 10 Mar
		Assignment 1 due	Thu 11 Mar	16:00	**16:00 Thu 11 Mar**
15	15 Mar	SO Lecture	Fri 19 Mar	12:00–13:00	—
		NSO activities	*Some pre-class reading/videos*		Thu 18 Mar
		Assignment 2 released	Thu 18 Mar	16:00	16:00 Thu 29 Apr
16	22 Mar	Problems class	Thurs 25 Mar	*—11:00/12:00/13:30 Check timetable—*	
		NSO activities	*Some pre-class reading/videos*		Wed 24 Mar
—		*—Easter holidays—*			
17	26 Apr	SO Lecture	Fri 30 Apr	12:00–13:00	—
		NSO activities	*Some pre-class reading/videos*		Thu 29 Apr
		Assignment 2 due	Thu 29 Apr	16:00	**16:00 Thu 29 Apr**
18	3 May	Problems class	Thurs 6 May	*—11:00/12:00/13:30 Check timetable—*	
		NSO activities	*Some pre-class reading/videos*		Wed 5 May
19	10 May	Revision class	Tues 11 May	13:00–14:00	—
		Revision class	Fri 14 May	12:00–13:00	—

Assessment: Reminder

The full 20 credit MAS2905 module will be assessed via a formal written examination in summer 2021 (possibly a '24 hour take-home paper') and some in-course assessment.

The make-up is 50% summer exam, 50% in-course assessment. Further details:

- **Summer exam, 100 marks in total**
  - Semester 1 material: Section A (20 marks)
  - Semester 2 material: Section B (80 marks)
  - Answer *all* questions in *both* parts
- **In-course assessment, 100 marks in total**
  - Semester 1 material: Test 1 (20 marks), mini project (30 marks), test 2 (30 marks)
  - Semester 2 material: Two short homeworks (10 marks each)

Course materials

All course materials will be posted to the dedicated Canvas page for MAS2905. Here, you will find:

- A full copy of these lecture notes, with gaps
- A copy of the lecture notes with the gaps filled in, after the associated SO/NSO activities
- All slides used to support SO/NSO activities after these have taken place
- All video recordings
- Solutions to assessed exercises
- Any other materials, including additional handouts/worksheets and specimen exam papers and solutions, . . .

Books and other resources

Some useful books for semester 2 include:

- "*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.

Preface to lecture notes:

What is Bayesian Statistics?

Statistics: An alternative approach

- Up until now, you have been taught Probability and Statistics according to a particular way of thinking
- The **Bayesian paradigm** offers another way of seeing things
- Your ideas about Probability and Statistics are deeply entrenched...perhaps so much so that at first, *Bayesian Statistics* might take you a while to grasp!
- Not because it's hard! Just because you have become so conditioned to think in a particular way.
- We want to broaden your horizons a bit!

Statistics: An alternative approach

There are two main approaches to statistics: **frequentist** (or classical) statistics and **Bayesian** statistics.

All of the statistics teaching you've encountered so far is likely to be about frequentist methods. Bayesian methods are substantially different and can feel quite strange to start with.

So, before starting the main course material, this **non-examinable** preface explains the concepts behind Bayesian statistics and how they differ from frequentist approaches.

Statistics: An alternative approach

One way of defining Statistics is as a way to learn about the world from some data which is subject to random variation.

For example, in **climate science** we want to learn about the climate given some imperfect measurements.

Some statistical questions that we might ask in this field are:

- What is our best estimate of the world temperature this year? This is a point estimation problem.
- What is a plausible range of values? This is an interval estimation (or uncertainty quantification) problem.
- What will the temperature be in 100 years?This is a prediction problem.
- Is the climate warming?This is a hypothesis testing problem.

Statistics: An alternative approach

Frequentist and Bayesian methods address all these types of problem: point estimation, interval estimation, prediction and hypothesis testing.

But they use different approaches to do so.

Some typical frequentist approaches include:

- Least squares (point estimation)
- Maximum likelihood (point estimation)
- Confidence intervals (interval estimation)
- Test statistics and p-values (hypothesis testing)

Bayesian Statistics doesn't use any of these familiar methods!

Bayes goes to war!

- The work of Thomas Bayes (see later) was published in 1764, 3 years after his death
- Over the course of the next 100–150 years, it received little attention
- In fact, some key figures in Statistics – e.g. **R.A. Fisher** – outrightly rejected the idea of Bayesian statistics
- By the start of WW2, Bayes' rule was **virtually taboo** in the world of Statistics!
- During WW2, some of the world's leading mathematicians resurrected Bayes' rule in deepest secrecy to crack the coded messages of the Germans

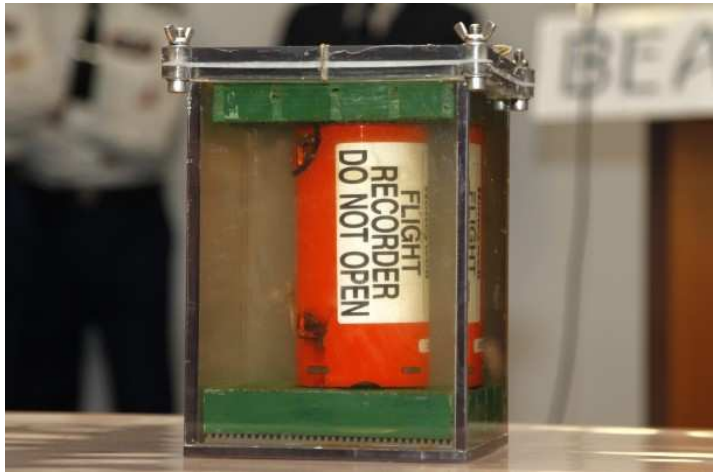
Bayes goes to war!

- **Alan Turing** – mathematician working at Bletchley Park
- Designed the ‘bombe’ – an electro–mechanical machine for testing every possible permutation of a message produced on the Enigma machine — could take up to **4 days** to decode a message
- New system: **Banburismus** – where Bayesian methods were used to quantify the belief in guesses of a stretch of letters in an Enigma message
- Certain permutations – unlikely to be the original message – were ‘thrown out’ before they were even tested
- Greatly reduced the time it took to crack Enigma codes

Air France 447

12:56:47

Heure Time	Provenance From	Vol Flight			Terminal	
1100	Los Angeles Minneapolis	AF	065	DL 8553	2E	Landed
1115	Rio Janeiro Intl	AF	447		2E	Delayed
1120	Bogota	AF	423		2E	Arrived 11:34
1125	Detroit Wayne Co	AF	377	DL 8573	2E	Arrived 11:29
1130	Damascus	AF	511		2E	Arrived 11:31
1130	Washington Dulles	AF	027	DL 8331	2E	Arrived 11:39
1135	Montreal	AF	349		2E	Arrived 11:41
1135	Sao Paulo Guarul	AF	459			Transfered 2C
1140	Istanbul	AF	2391	NW 4367	2E	Arrived 11:54
1140	Tunis	AF	1685	DL 8595	2E	Arrived 11:28
1150	Birmingham	AF	5133	MK 9331	2E	Arrived 11:42
1150	London-City	AF	5059	MK 9391	2E	Arrived 11:29
1210	Dublin	AF	5001	UX 3574	2E	Arrived 12:05



Air France 447

- Black boxes could be anywhere within an area of the South Atlantic the size of Switzerland (**6,500 square miles**)
- Mid–Atlantic ridge – between two tectonic plates – just as Mountainous as Switzerland!
- So remote – scientists have not yet charted the sea–bed
- Search method: sonar – detectors emitting sound waves which would bounce back once they hit something
- After two years of meticulously searching an area north of the plane's trajectory (after analysing debris drift) $\longrightarrow$ **nothing**

Air France 447

- April 2011: *Metron, Inc.*, of Retson, Virginia, hired to launch a **Bayesian review** of the entire search effort
- Included in the analysis were:
  - Data from 9 previous airline crashes involving loss of pilot control – reduced search area to **1,600 square miles**
  - **Expert opinions** on the credibility of the flight data
  - **Expert opinions** about whether or not the black box 'pingers' might have become damaged on impact
  - Positions/recovery times of bodies found drifting – **expert opinions** assigned to the reliability of this data because of the turbulent equatorial waters
  - **Expert information** from oceanographers: Sea state, visibility, underwater geography,...
- All information combined through **Bayes Theorem**: After one week, black boxes found!

Body Mass Index

Suppose we are interested in whether or not there is any real difference between the Body Mass Index (BMI) of Maths & Stats students and Food & Human Nutrition students.

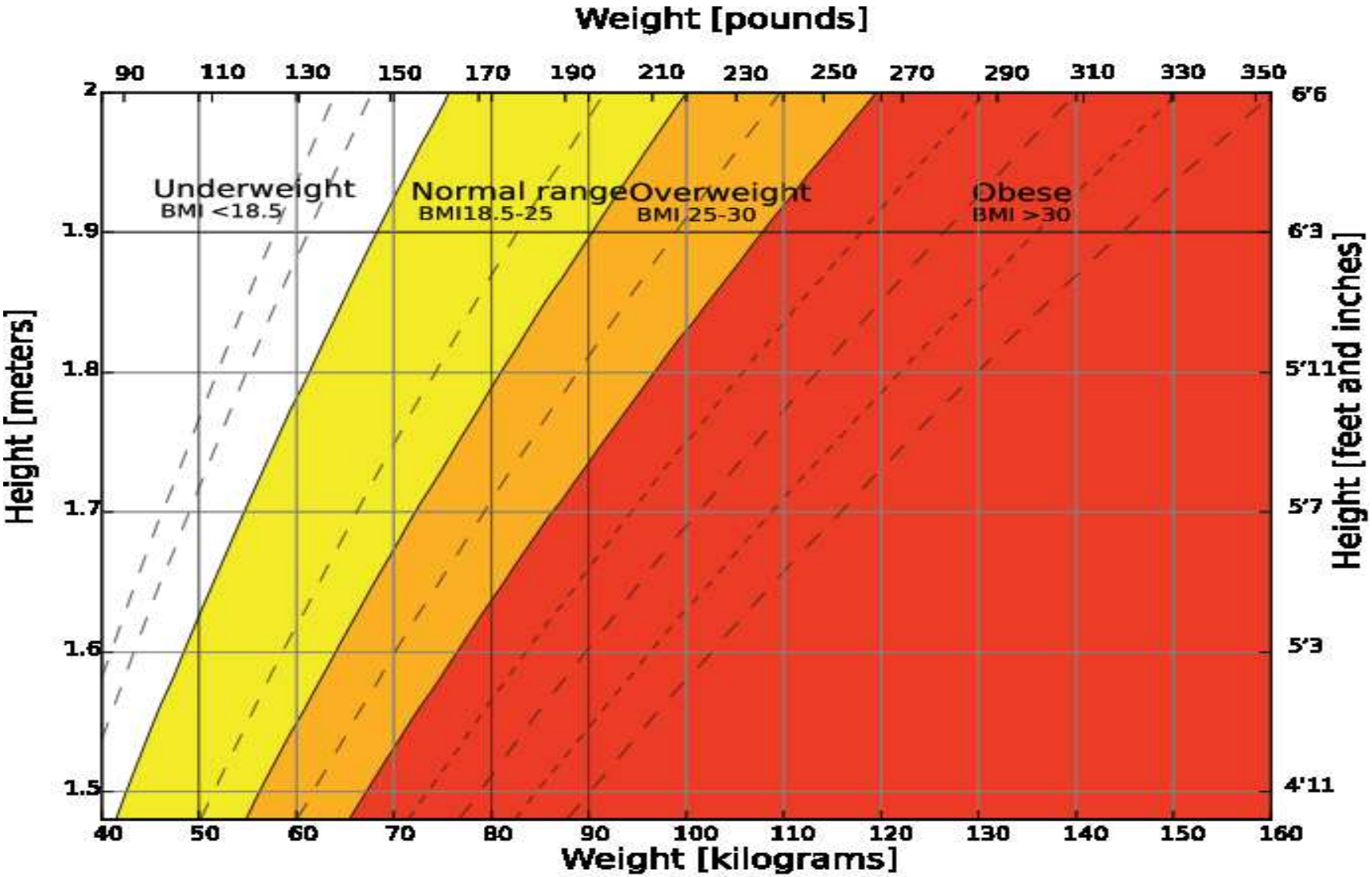
BMI is often used as a measure of ‘fatness’ or ‘thinness’: it is the ratio of a person’s weight (measured in kg) to their squared height (measured in metres).

In a two–sample t–test, we often test the null hypothesis that there is no difference between the population means of the two groups:

$$H_0 : \mu_M = \mu_F,$$

How can we test this hypothesis?

Body Mass Index



Body Mass Index

Suppose we obtain the following BMI values for a random sample of students from both groups:

Maths & Stats	23.1	28.3	35.3	24.0	32.4	30.5	30.1	21.9	22.1
	29.9	27.2	33.2						
Food & Human Nutrition	25.5	27.3	21.0	23.3	19.1	22.2	29.5	27.5	23.2

From this we can obtain the summaries:

$$n_M = 12 \qquad \bar{x}_M = 28.17 \qquad s_M = 4.54$$

$$n_F = 9 \qquad \bar{x}_F = 24.29 \qquad s_F = 3.40$$

Body Mass Index

The pooled standard deviation is

$$
\begin{aligned}
s &= \sqrt{\frac{(n_M - 1)s_M^2 + (n_F - 1)s_F^2}{n_M + n_F - 2}} \\
&= \sqrt{\frac{11 \times 4.54^2 + 8 \times 3.4^2}{12 + 9 - 2}} \\
&= 4.099
\end{aligned}
$$

Then the test statistic is

$$
\begin{aligned}
t &= \frac{|\bar{x}_M - \bar{x}_F|}{s \times \sqrt{\frac{1}{n_M} + \frac{1}{n_F}}} \\
&= 2.147.
\end{aligned}
$$

Body Mass Index

From t_{19}–tables, we get:

p–value	10%	5%	1%
critical value	1.729	2.093	2.861

Our test statistic $t = 2.147$ indicates that our p–value is less than 5%:

- We have a significant result at the 5% level of significance
- We **reject** H_0
- We conclude that there *is* a significant difference in BMI between M&S and FHN students

Body Mass Index

Notice from R that the *exact* p–value is 0.004488 = 4.488%.

- As we have already concluded, we would **reject** H_0 at the 5% level;
- However, if we work at the 1% level of significance, we would **retain** H_0!

Convention tells us to work at the 5% level, but why? Some practitioners work at the 1% level. Do we need to collect more data and re-analyse?

Body Mass Index

In MAS1604 you were introduced to **Bayes' Theorem**.

As we shall see, Bayes' Theorem gives us a way of combining subjective assessments with observed data.

For example, what if – prior to observing the data – we believed that Food & Human Nutrition students were likely to have a **considerably lower BMI** than Maths & Stats students?

What if, from a previous study, we knew that Maths & Stats students were quite likely to have a **BMI somewhere between 25 and 33**?

Body Mass Index

Would it not be sensible to build this information into our analysis before forming our conclusions using the data alone?

This could be a good idea – we have relatively small samples, which could be biased!

A **Bayesian analysis** can do this!

Some people argue against this on the grounds that it is subjective... But as we have just seen, the usual approach to hypothesis testing is also subjective!

Not only are hypothesis tests tricky to interpret... but so are **confidence intervals**.

Body Mass Index

Obtain a 95% confidence interval for the population mean BMI for Maths & Stats students.

We use

$$\bar{x}_M \pm t\frac{s_M}{\sqrt{n_M}}, \qquad \text{giving}$$

$$28.17 \quad \pm \quad 2.201 \times \frac{4.54}{\sqrt{12}} \qquad \text{i.e.}$$

$$28.17 \quad \pm \quad 2.743,$$

giving (25.28, 31.05).

Body Mass Index

What is $\Pr(25.28 < \mu_M < 31.05)$?

NOT 0.95!! In the frequentist approach, population parameters (in this case μ_M) are NOT random variables!

This means μ_M is a fixed (but unknown) quantity. So it's either *in* the interval, or it's *not*, i.e.

$$\Pr(25.28 < \mu_M < 31.05) = 1 \qquad \text{or}$$

$$\Pr(25.28 < \mu_M < 31.05) = 0.$$

Nothing else!

Equivalent intervals in the **Bayesian framework** have more natural interpretations!

Frequentist/Subjectivist debate

One of the main arguments *against* working within the Bayesian paradigm is that it is **subjective**.

Although we have not yet thought about *how* the Bayesian framework combines subjective assessments with the data, we have said that this is what it does.

Surely we should strive to be as *objective* as we can?

Frequentist/Subjectivist debate

Actually, to be able to incorporate a person's beliefs into an analysis of sample data is surely the **right thing to do** — especially if that person is an expert!

This is in line with how scientific experiments are carried out:

- The experimenter usually **knows something**;
- She then carries out the experiment in which **data are collected**;
- the experimenter then **updates her beliefs** from these results.

Frequentist/Subjectivist debate

In other words, the data are used to refine what the experimenter knows.

And that is what MAS2905 is all about – we will consider how

- **prior beliefs** can be combined with
- **experimental data** to form
- **posterior beliefs** – which include both the prior knowledge and what we have learnt from the data

Bayesian Statistics: Leading players

Thomas Bayes (1702–1761)

- Presbyterian minister in Tunbridge Wells, Kent
- Bayes' solution to a problem of "inverse probability" was presented in the *Essay towards Solving a Problem in the Doctrine of Chances* (1764)
- Work published posthumously by his friend Richard Price in the *Philosophical Transactions of the Royal Society of London*.
- This work gives us **Bayes' Theorem**, a key result on which most of the work in this course rests.

Bayesian Statistics: Leading players

LII. *An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Bayesian Statistics: Leading players

Bruno di Finetti (1906–1985)

- Italian mathematical probabilist, developed his ideas on subjective probability in the 1920s.
- Famed for saying

  "Probability does not exist"

- By this, he meant it has no *objective* existence – i.e. it is not a feature of the world around us.
- It is a measure of degree–of–belief – your belief, my belief, someone else's belief – all of which could be different.

Bayesian Statistics: Leading players

Dennis Lindley (1923—2013)

- Worked hard to find a mathematical basis for the subject of Statistics
- With Leonard Savage, he found a deeper justification for Statistics in **Bayesian theory**
- Turned into a critic of the classical statistical inference he had hoped to justify
- Quoted as saying:

  "Uncertainty is a personal matter. It is not the uncertainty, but your uncertainty"

Recent developments

One of the difficulties in early applications of Bayesian methodology was the maths:

Combining prior beliefs with a probability model for the data often resulted in maths that was just **too difficult/impossible to solve by hand**.

Recent developments

Then – in the early 1990s – a technique called "**Markov chain Monte Carlo**" (MCMC for short) was developed.

MCMC is a computer–intensive simulation–based procedure that gets around the problem of hard maths.

If you choose to take **MAS3902: Bayesian Inference** next year, you will be introduced to this technique.

MCMC has **revolutionised the use of Bayesian Statistics**, to the extent that Bayesian data analyses are now routinely used by non–Statisticians in fields as diverse as biology, civil engineering and sociology.

Newcastle's contribution

Number of academic staff publishing advanced research using Bayesian methods:

- 1985: **1** (Professor Boys)
- 2018: **Lots!**
  - *Systems Biology:* (Boys, Farrow, Gillespie, Golightly, Heaps, Henderson, Prangle, Wilkinson, Wilson)
  - *Environmental Extremes:* (Fawcett, Walshaw)

Newcastle's contribution: Environmental Extremes

Use Statistical methods to help plan for environmental extremes:

- **Hurricane–induced waves**: how high should we build a sea–wall to protect New Orleans against another *Hurricane Katrina*?
- **Extreme wind speeds**: how strong should we design buildings so they will not fail in storm–strength wind speeds?
- **Extreme cold spells**: How much fuel should we stockpile to cater for an extremely cold winter?

Newcastle's contribution: Environmental Extremes

We work within the Bayesian framework to combine

- expert information from hydrologists/oceanographers, with
- extreme rainfall data/hurricane–induced sea surge data,

to help estimate how likely an **extreme flooding event is to occur**, or to aid the **design of sea wall defences** in storm–prone regions.

If you're interested, you can see my webpage for more details.

Quick quiz

Suppose A, θ and μ_M are constants and Z is a random variable, such that $A = 5$, $\theta = 10$, $\mu_M = 33.5$ and $Z \sim N(0, 1)$.

Write down

1. $\Pr(Z > 0)$
2. $\Pr(-1.96 < Z < 1.96)$
3. $\Pr(2 < A < 4)$
4. $\Pr(7 < \theta < 12)$
5. $\Pr(25.28 < \mu_M < 31.05)$

Chapter 1

Introduction

Introduction

Consider the following three experiments.

- **Experiment 1: Fisher's tea lady**
  The tea lady claims to know whether milk or tea is poured in first: for 10 pairs of cups of tea she makes the correct choice each time.
- **Experiment 2: Music expert**
  The expert claims he can distinguish between a page from a Haydn score and a page from a Mozart score: he does so correctly 10 times.
- **Experiment 3: The Drunk**
  A somewhat inebriated friend at a party claims they can predict the outcome of the toss of a coin: they do so correctly 10 times.

Introduction

Let $\theta = \Pr(\text{correct choice})$.

Let's suppose the tea lady, the music expert and the drunk *cannot* do as they claim.

Then, just by guessing, we could expect each of them to 'get it right' 5 times out of 10, i.e. $\theta = 1/2$.

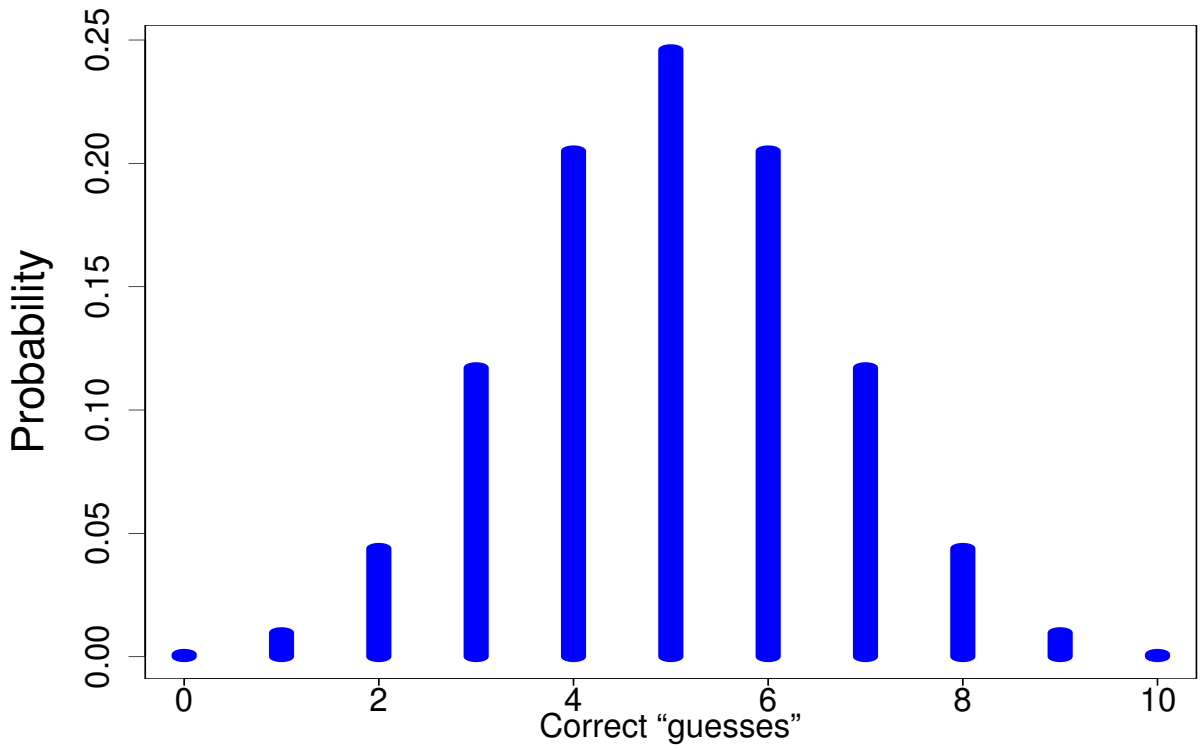
Introduction

We could then test the **null hypothesis**

$$H_0 : \text{they were guessing}, \qquad \text{i.e.}$$

$$H_0 : \theta = 1/2;$$

Graphically, assuming H_0 is true, this gives:



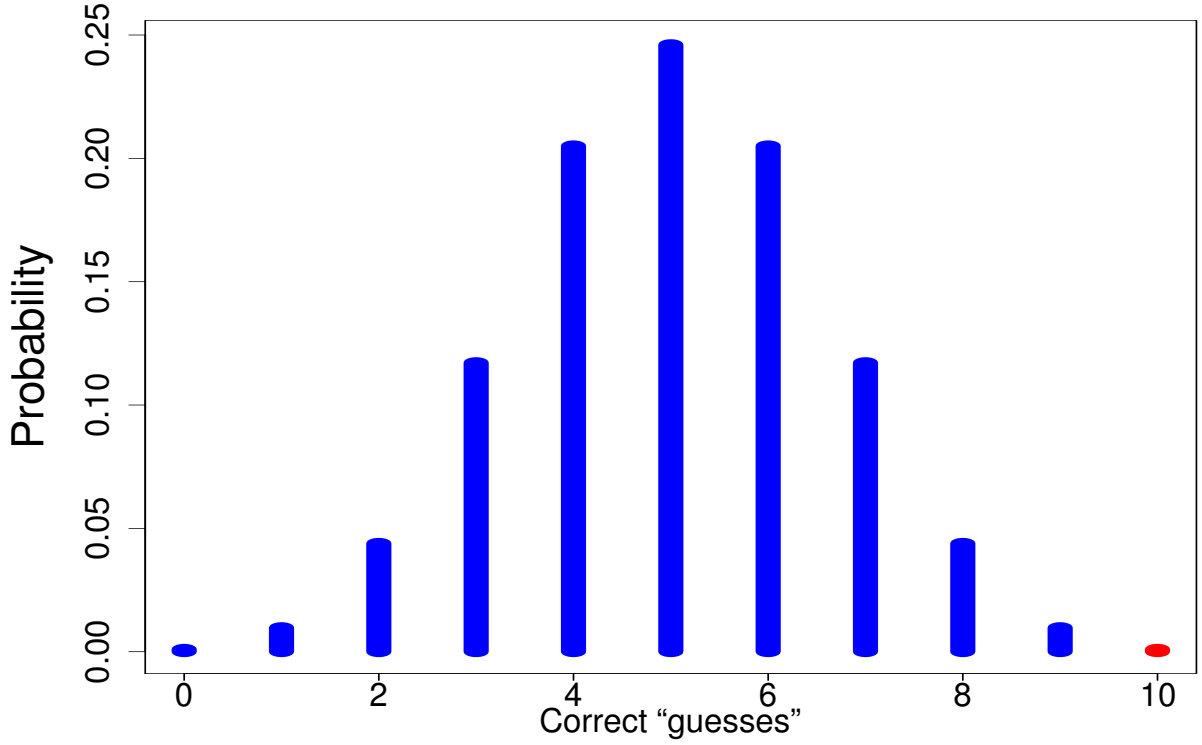
Introduction

We could then test the **null hypothesis**

$$H_0 : \text{they were guessing}, \qquad \text{i.e.}$$

$$H_0 : \theta = 1/2;$$

Graphically, assuming H_0 is true, this gives:



Introduction

This gives us:

$$p\text{−value} = \left(\frac{1}{2}\right)^{10} = 0.00098 < 0.1\%.$$

From this $p-$value, we would:

- conclude that we had **very strong evidence** against H_0
- conclude that the choices were **not just guesses**
- perhaps feel justified in **validating each claim**

But does this make sense?

Surely, we have some additional information about what values of θ are **plausible** for each experiment?

Introduction

Prior to each experiment, our beliefs about θ may be

- **Experiment 1:** $\theta > 0.5$ – folklore (and science!) suggests this may be possible;
- **Experiment 2:** $0.9 < \theta < 1.0$ – we expect an "expert" to be correct;
- **Experiment 3:** $\theta = 0.5$ – no way of guessing correctly with a "fair" coin.

Introduction

The traditional approach to Statistics, sometimes called **Frequentist Statistics** or **Classical Statistics**, may try to take this prior information into account:

For example, in Experiment 2, the test may be of $H_0 : \theta \geq 0.9$ against $H_1 : \theta < 0.9$.

However, in **Bayesian Statistics**, we attempt to calibrate our prior information about unknown quantities.

We do this by constructing a **probability distribution** which describes how likely we believe different values are to occur.

Introduction

This **prior information** is then combined with that from **experimental data** using Bayes Theorem.

The key ingredients of a Bayesian analysis are

- a statistical model for the experimental data;
- quantifiable prior information about any unknown parameters.

Before we consider any detailed descriptions of Bayesian analyses, we recap the various interpretations of probability and highlight the **subjective approach**.

Probability

The concept of probability (chance) has been around for a very long time, particularly in the area of **gambling**.

Games of chance have been played since about **3500 B.C.**

The Egyptians started using cubical dice around **2000 B.C.**

The mathematical theory of probability was started around the 17th century by **Galilei**, **Pascal** and **Fermat** to solve (again) gambling problems.

There are three main ways of understanding and thinking about probability.

Frequency interpretation

The probability of an outcome is the **relative frequency** with which the outcome would be obtained if the experiment were repeated a large number of times *under similar conditions*.

For example, if a coin is tossed 1,000,000 times and a head appears n times then

$$\Pr(\text{Head}) = \frac{n}{1,000,000}.$$

We would expect this probability to be about 0.5.

Classical interpretation

This is based on the concept of **equally likely outcomes** resulting from ideas of **symmetry**.

If the outcome of an experiment must be one of n different outcomes and these n outcomes are equally likely, then the probability of each outcome is $1/n$.

Subjective interpretation

Your **subjective probability** for an outcome A represents your own judgement of the likelihood that the outcome will occur.

This judgement will be based on the beliefs and information H you have at the time.

One way of determining (or quantifying) a subjective value for $\Pr_H(A)$ is to consider a series of possible bets with outcome

$$\text{win £}c \text{ if } A \text{ occurs and £0 if } A^c \text{ occurs.}$$

How much would you be prepared to pay (stake) for placing such a bet?

Subjective interpretation (1/1)

In terms of expected winnings, you should be prepared to stake £cp if you believe that $\Pr_H(A) = p$. Why?

Suppose you would win £10 if A occurs and £0 if it does not. Suppose further that you believe A is likely to occur with probability $p = 0.3$.

Then the **expected monetary value** of the bet is
£$10 \times 0.3 =$ £3;

that is, on average, we can expect to win £3 from this bet. Thus, staking any more than £3 would not be wise!

Subjective interpretation

One problem with this approach is that, in general, p will depend on c.

A person who is willing to bet £1 on the spin of a coin to win £2 if it lands heads may refuse to bet if the stakes are raised to £1000: most people are **risk–averse**.

Therefore, we shall restrict our attention to the $c = 1$ case: pay £p for the bet

win £1 if A occurs and £0 if A^c occurs.

Subjective interpretation

You can make sure that your bet is "honest" by randomising between whether you "**host**" the bet or "**place**" the bet.

For example, suppose you believe that $\Pr_H(A) = 0.5$.

An "honest" bet would mean that you would buy the bet for a maximum stake of £0.50.

However, if you weren't honest you might try to buy the bet for any amount less than £0.50, say £0.20.

If you were hosting the bet, you would take the bet for any amount more than £p, say for £0.80.

These conflicting interests can be offset if, when choosing p, it is equally likely that you are hosting the bet or placing the bet.

Frequency interpretation: drawbacks

1. It does not say how many times the experiment should be repeated
2. “Similar conditions” is a vague concept
3. It is not appropriate for many probability calculations of one-off events
4. Standard statistical methods using the frequentist approach are not totally objective since they require subjective judgements about the validity of probability models, choice of hypotheses and interpretation of results (for instance, see BMI example in the preface to these lecture notes)

Classical interpretation: drawbacks

1. Only applies to equally likely outcomes
2. Depends on a subjective assessment of whether symmetry arguments apply
3. It is not appropriate for many probability calculations of one–off events

Subjective interpretation: drawbacks

1. It is not objective – but perhaps it is more obvious (honest) about when subjective beliefs are used
2. It requires people to be **coherent**: they will not make any wagers which they are certain to lose; also, they will not prefer to suffer a given penalty when there is the option of another penalty which is certainly smaller

   Being coherent results in, *inter alia*, that

$$\Pr(A_1|H) > \Pr(A_2|H) \quad \text{and} \quad \Pr(A_2|H) > \Pr(A_3|H) \implies \Pr(A_1|H) > \Pr(A_3|H).$$

Probability

Each of these interpretations use quite different methods of reasoning.

In this course – unlike any other course you have taken so far – we will concentrate on the subjective interpretation and describe how, if carefully used, it can be a more useful approach than the other two methods.

Everything we do rests on **Bayes' Theorem**, and we review this now.

Bayes' Theorem

Before we state **Bayes' Theorem**, we need a recap of **conditional probability**.

Definition (Conditional Probability)

Consider two events E and F, where $\Pr(F) > 0$.

The **conditional probability** of E given that F has occurred is

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

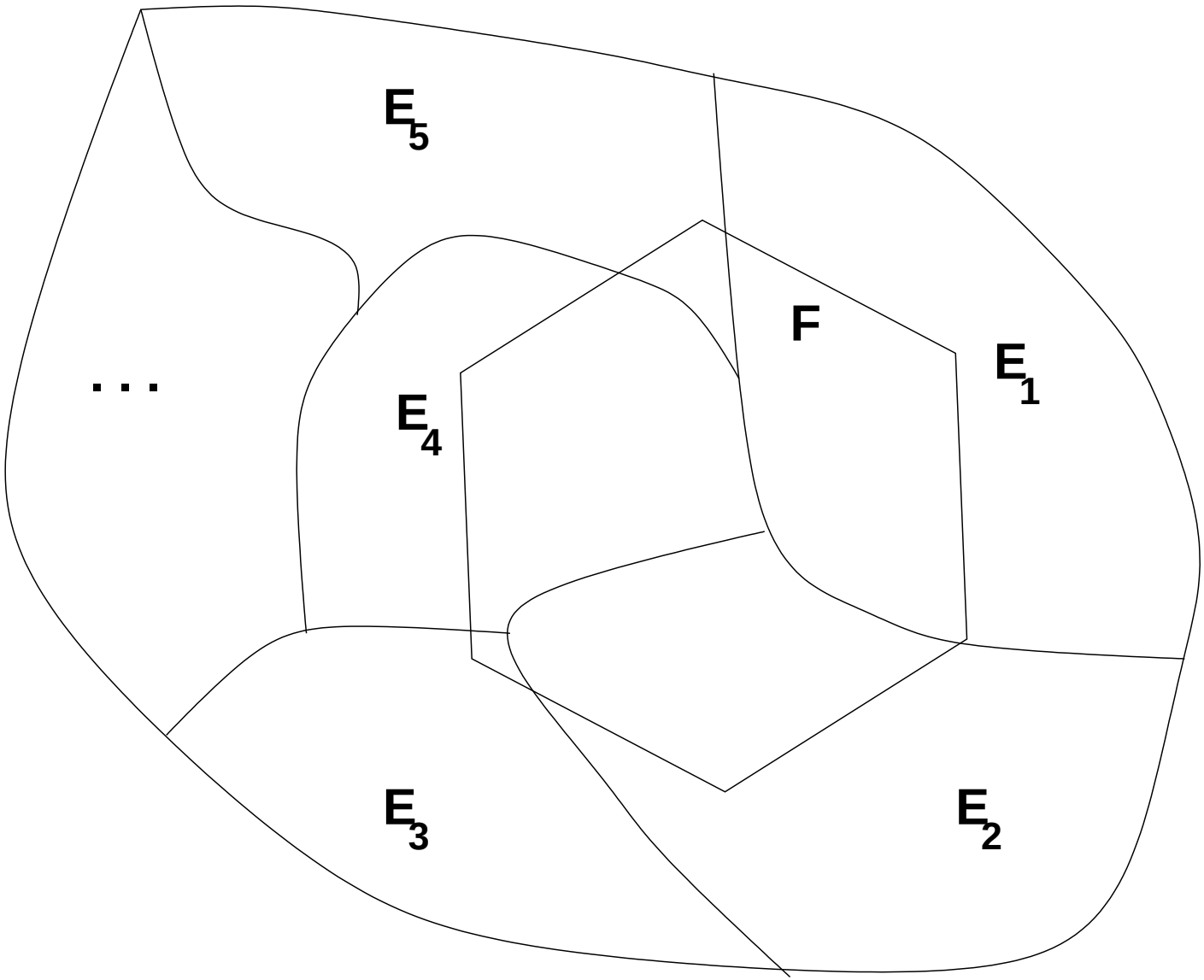
Bayes' Theorem

Definition (Partition)

The events $E_1, E_2, \ldots, E_n$ form a **partition** of the sample space $\mathcal{S}$ if they are disjoint events ($E_i \cap E_j = \emptyset$, $i \neq j$) with $\Pr(E_i) > 0$, $i = 1, 2, \ldots, n$, and $\cup_{i=1}^{n} E_i = \mathcal{S}$.

Figure 1.1 gives a diagram of a typical partition with an additional event F.

Bayes' Theorem



Bayes' Theorem

[Law of Total Probability] If $E_1, E_2, \ldots, E_n$ are a **partition** of $\mathcal{S}$ and F is any event then

$$\Pr(F) = \sum_{i=1}^{n} \Pr(F|E_i)\Pr(E_i).$$

Law of Total Probability: Proof

As $E_1, E_2, \ldots, E_n$ are a **partition** of $\mathcal{S}$, we have

$$\Pr(F) = \Pr(F \cap E_1) + \Pr(F \cap E_2) + \ldots + \Pr(F \cap E_n)$$

$$= \Pr(F|E_1)\Pr(E_1) + \Pr(F|E_2)\Pr(E_2) + \ldots + \Pr(F|E_n)\Pr(E_n)$$

(by conditional probability)

$$= \sum_{i=1}^{n} \Pr(F|E_i)\Pr(E_i).$$

Bayes' Theorem

Theorem (Bayes' Theorem)

If $E_1, E_2, \ldots, E_n$ *are a* **partition** *of* $\mathcal{S}$ *and* F *is any event with* $\Pr(F) > 0$ *then*

$$\Pr(E_i|F) = \frac{\Pr(F|E_i)\Pr(E_i)}{\sum_{j=1}^{n} \Pr(F|E_j)\Pr(E_j)}, \qquad i = 1, 2, \ldots, n.$$

Bayes' Theorem: Proof (slide 1/1)

Using the definition of conditional probability, for $i = 1, 2, \ldots, n$,

$$\Pr(E_i|F) = \frac{\Pr(E_i \cap F)}{\Pr(F)}$$

$$= \frac{\Pr(F \cap E_i)}{\Pr(F)}$$

$$= \frac{\Pr(F|E_i)\Pr(E_i)}{\Pr(F)} \quad \text{again using the definition}$$

$$= \frac{\Pr(F|E_i)\Pr(E_i)}{\sum_{j=1}^{n} \Pr(F|E_j)\Pr(E_j)} \quad \text{using the Law of Total Probability.}$$

Example 1.2

A laboratory blood test is 95% effective in detecting a certain disease when it is present.

However, the test also yields a “false positive” result for 1% of healthy people tested.

Also, 0.5% of the population actually have the disease.

(a) Calculate the probability that a person who tests positive actually has the disease.

(b) Find the probability that a person who tests negative does *not* have the disease.

Solution to Example 1.2 (slide 1/3)

Let $D =$ person has the disease, +ve $=$ result is postive and –ve $=$ result is negative.

We require (a) $\Pr(D|\text{+ve})$, and (b) $\Pr(D^c|\text{–ve})$.

From the question we have

- $\Pr(\text{+ve}|D) = 0.95$
- $\Pr(\text{+ve}|D^c) = 0.01$
- $\Pr(D) = 0.005$ and
- $\Pr(D^c) = 0.995$

Solution to Example 1.2 (slide 2/3)

Also, D and D^c form a partition of $\mathcal{S}$. Therefore, by Bayes' Theorem

$$\Pr(D|\text{+ve}) = \frac{\Pr(\text{+ve}|D)\Pr(D)}{\Pr(\text{+ve}|D)\Pr(D) + \Pr(\text{+ve}|D^c)\Pr(D^c)}$$

$$= \frac{0.95 \times 0.005}{0.95 \times 0.005 \ + \ 0.01 \times 0.995}$$

$$\simeq 0.323.$$

Thus, only 32.3% of people who test positive actually have the disease – 67.7% of people who test positive do *not* have the disease!

Solution to Example 1.2 (slide 3/3)

Now for part (b),

$$\Pr(D^c|\text{–ve}) = \frac{\Pr(\text{–ve}|D^c)\Pr(D^c)}{\Pr(\text{–ve}|D^c)\Pr(D^c) + \Pr(\text{–ve}|D)\Pr(D)}$$

$$= \frac{0.99 \times 0.995}{0.99 \times 0.995 \ + \ 0.05 \times 0.005}$$

$$\simeq 0.9997.$$

Therefore, nearly everyone who tests negative does not have the disease – very few people who test negative *will* have the disease.

Quick recap

- General intro to Bayesian methods
  - Deciphering Enigma messages
  - Air France 447
  - Body Mass Index
- Recap of probability
  - Frequency, classical and subjective interpretations
  - Limitations
  - Rules: conditional probability, total probability, Bayes
  - Examples: blood test, **car diagnosis**, multiple choice exam

Example 1.3

Suppose that your car suffers from two intermittent problems, one caused by a fault in the engine (θ_1) and the other due to a fault in the gearbox (θ_2).

These occur with probabilities 0.4 and 0.6 respectively.

When examined your car exhibits one of the following symptoms

$$
\begin{aligned}
x_1 &: \text{overheating only,} \\
x_2 &: \text{irregular traction only,} \\
x_3 &: \text{both symptoms.}
\end{aligned}
$$

Example 1.3

Suppose it is known in the garage trade that these symptoms occur with probabilities that depend on the fault.

The probabilities $\Pr(X = x|\theta)$ are given in Table 1.1.

	O/H x_1	I/T x_2	Both x_3
θ_1: fault in engine	0.1	0.4	0.5
θ_2: fault in gearbox	0.5	0.3	0.2

Construct a diagnostic rule for these symptoms and determine the probability of misdiagnosis.

Solution to Example 1.3 (1/3)

First we must calculate the **posterior probabilities** $\Pr(\theta_1|x)$ and $\Pr(\theta_2|x)$ for $x = x_1, x_2, x_3$.

Since θ_1 and θ_2 form a partition, we can use Bayes' Theorem as follows.

We have

$$\Pr(\theta_1|x_1) = \frac{\Pr(X = x_1|\theta_1)\Pr(\theta_1)}{\Pr(X = x_1|\theta_1)\Pr(\theta_1) + \Pr(X = x_1|\theta_2)\Pr(\theta_2)}$$

$$= \frac{0.1 \times 0.4}{0.1 \times 0.4 + 0.5 \times 0.6}$$

$$= \frac{4}{34} = 0.118.$$

Solution to Example 1.3 (2/3)

Also,

$$\Pr(\theta_1|x_2) = \frac{\Pr(X = x_2|\theta_1)\Pr(\theta_1)}{\Pr(X = x_2|\theta_1)\Pr(\theta_1) + \Pr(X = x_2|\theta_2)\Pr(\theta_2)}$$

$$= \frac{0.4 \times 0.4}{0.4 \times 0.4 + 0.3 \times 0.6}$$

$$= \frac{16}{34} = 0.471.$$

Solution to Example 1.3 (3/3)

And

$$\Pr(\theta_1|x_3) = \frac{\Pr(X = x_3|\theta_1)\Pr(\theta_1)}{\Pr(X = x_3|\theta_1)\Pr(\theta_1) + \Pr(X = x_3|\theta_2)\Pr(\theta_2)}$$

$$= \frac{0.5 \times 0.4}{0.5 \times 0.4 + 0.2 \times 0.6}$$

$$= \frac{20}{32} = 0.625.$$

Also, $\Pr(\theta_2|x_i) = 1 - \Pr(\theta_1|x_i)$, $i = 1, 2, 3$, and so we obtain the **posterior distributions** $\Pr(\theta|x)$ given in Table 1.2.

Example 1.3

	O/H x_1	I/T x_2	Both x_3
θ_1: fault in engine	0.118	0.471	0.625
θ_2: fault in gearbox	0.882	0.529	0.375

This table is very informative. For example, it shows that if both symptoms (x_3) are observed, then the probability that the fault is in the engine (θ_1) changes from 0.4 to 0.625.

In terms of **odds**:

$$\text{Prior odds}: \frac{\Pr(\theta_1)}{\Pr(\theta_2)} = \frac{0.4}{0.6} = \frac{2}{3} \quad \text{or 3:2 in favour of } \theta_2$$

$$\text{Posterior odds}: \frac{\Pr(\theta_1|x_3)}{\Pr(\theta_2|x_3)} = \frac{0.625}{0.375} = \frac{5}{3} \quad \text{or 5:3 in favour of } \theta_1.$$

Example 1.3

We are now in a position to design our **diagnostic rule**.

This is simply a rule which diagnoses a symptom (x) as being due to some particular fault (θ).

Consider first that we observe overheating only (x_1). The posterior probabilities are in favour of declaring the fault as in the gearbox (θ_2) since $\Pr(\theta_2|x_1) > \Pr(\theta_1|x_1)$.

In the same way, we can determine the most likely diagnosis having observed irregular traction only (x_2) and both symptoms (x_3), giving the diagnostic rule in Table 1.3.

Symptom	Diagnosis
overheating only (x_1)	fault in gearbox (θ_2)
irregular traction only (x_2)	fault in gearbox (θ_2)
both symptoms (x_3)	fault in engine (θ_1)

Solution to Example 1.3 (1/1)

If we were to carry out this diagnostic rule repeatedly then the probability of misdiagnosing a fault is

$$\begin{aligned}
&\Pr(\text{Misdiagnosis}) \\
&= \Pr(\theta_1, X = x_1) + \Pr(\theta_1, X = x_2) + \Pr(\theta_2, X = x_3) \\
&= \Pr(X = x_1|\theta_1)\Pr(\theta_1) + \Pr(X = x_2|\theta_1)\Pr(\theta_1) \\
&\qquad + \Pr(X = x_3|\theta_2)\Pr(\theta_2) \\
&= (0.1 \times 0.4) + (0.4 \times 0.4) + (0.2 \times 0.6) \\
&= 0.32.
\end{aligned}$$

Therefore, in repeated use of this rule, around a third of the diagnoses will be wrong.

Example 1.4

A student sits a multiple choice exam in which there are m alternative answers to each question.

The student either knows the answer (with probability θ) or guesses randomly (with probability $1 - \theta$).

What is the probability that the student actually knew the answer to a question they answered correctly?

Solution to Example 1.4 (1/2)

Let

$$C = \text{student answers question correctly}$$
$$K = \text{student knows answer.}$$

We require $\Pr(K|C)$. From the question we have

- $\Pr(K) = \theta$
- $\Pr(K^c) = 1 - \theta$
- $\Pr(C|K) = 1$ and
- $\Pr(C|K^c) = \frac{1}{m}$.

Solution to Example 1.4 (2/2)

Also, K and K^c form a partition of $\mathcal{S}$. Therefore, by Bayes' Theorem,

$$\begin{aligned}
\Pr(K|C) &= \frac{\Pr(C|K)\Pr(K)}{\Pr(C|K)\Pr(K) + \Pr(C|K^c)\Pr(K^c)} \\
&= \frac{1 \times \theta}{1 \times \theta \; + \; \frac{1}{m} \times (1-\theta)} \\
&= \frac{\theta}{\frac{m\theta + (1-\theta)}{m}} \\
&= \frac{m\theta}{1 + (m-1)\theta}.
\end{aligned}$$

Example 1.4

Suppose that there are $m = 5$ alternative answers for each question. Then

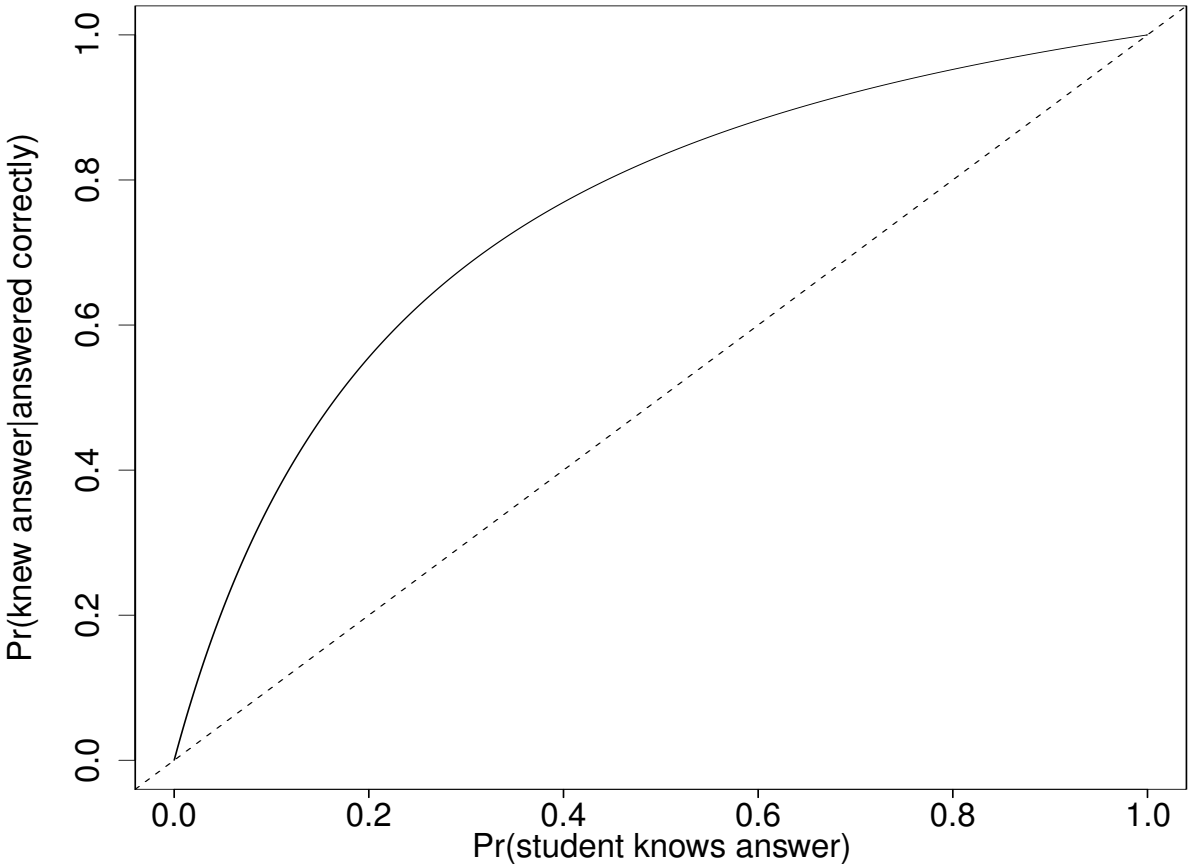
$$\Pr(K|C) = \frac{5\theta}{1 + 4\theta}.$$

We can see the effect of observing a correct answer on our belief that the student actually knows the answer by calculating $\Pr(K|C)$ for various θ – see Table 1.4.

Example 1.4

$\Pr(K)$ $= \theta$	$\Pr(K\|C)$ $= 5\theta/(1+4\theta)$
0.0	0.000
0.1	0.357
0.2	0.556
0.3	0.682
0.4	0.769
0.5	0.833
0.6	0.882
0.7	0.921
0.8	0.952
0.9	0.978
1.0	1.000

Example 1.4


Pr(knew answer|answered correctly)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Pr(student knows answer)

Likelihood

Suppose that an experiment results in data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ and we decide to model the data using a probability (density) function $f(\boldsymbol{x}|\theta)$.

This p(d)f describes how likely different data $\boldsymbol{x}$ are to occur given a value of the (unknown) parameter θ.

However, once we have observed the data, $f(\boldsymbol{x}|\theta)$ tells us how likely different values of the parameters θ are: it is then known as the **likelihood function** for θ.

In other courses you may have seen it written as $L(\theta|\boldsymbol{x})$ or $L(\theta)$ but, whatever the notation used for the likelihood function, it is simply the joint probability (density) function of the data, $f(\boldsymbol{x}|\theta)$, regarded as a function of θ rather than of $\boldsymbol{x}$.

Likelihood

The likelihood function can be simplified if we have further structure in the data.

For example, we may have independent observations, in which case

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta), \tag{1.1}$$

or independent and identically distributed observations (random sample), so that

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f_{X}(x_i|\theta). \tag{1.2}$$

Example 1.5

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ of radioactive particle counts.

A typical model for such data would be $X_i|\theta \sim Poisson(\theta)$, usually abbreviated $X_i|\theta \sim Po(\theta)$, (independent).

Determine the likelihood function for θ.

Solution to Example 1.5 (1/1)

Using Equation (1.2), the likelihood function is

$$
\begin{aligned}
f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \\
&= \frac{e^{-\theta} \times \ldots \times e^{-\theta} \times \theta^{x_1} \times \ldots \times \theta^{x_n}}{\prod_{i=1}^{n} x_i!} \\
&= \frac{e^{-\theta-\theta-\ldots-\theta}\theta^{x_1+\ldots+x_n}}{\prod_{i=1}^{n} x_i!} \\
&= \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!} \\
&= \frac{e^{-n\theta}\theta^{n\bar{x}}}{\prod_{i=1}^{n} x_i!}.
\end{aligned}
$$

Example 1.6

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ of times between radioactive particle emissions.

If the emissions occur randomly in time then a plausible model for such data would be $X_i|\theta \sim \mathit{Exponential}(\theta)$, usually abbreviated $X_i|\theta \sim \mathit{Exp}(\theta)$, (independent).

Determine the likelihood function for θ.

Solution to Example 1.6 (1/1)

Using Equation (1.2), the likelihood function is

$$
\begin{aligned}
f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\
&= \theta \times \ldots \times \theta \times e^{-\theta x_1} \times \ldots \times e^{-\theta x_n} \\
&= \theta^n e^{-\theta x_1 - \theta x_2 - \ldots - \theta x_n} \\
&= \theta^n e^{-\theta(x_1 + \ldots + x_n)} \\
&= \theta^n e^{-\theta n \bar{x}}.
\end{aligned}
$$

Example 1.7

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a Normal distribution: $X_i|\mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ (independent).

Determine the likelihood function for (μ, σ).

Solution to Example 1.7 (1/1)

The (joint) probability density function is

$$
\begin{aligned}
f(\boldsymbol{x}|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)\right\}.
\end{aligned}
$$

Sufficiency

Consider again the Poisson model in Example 1.5. The likelihood function is

$$f(\boldsymbol{x}|\theta) = \frac{\theta^{n\bar{x}} e^{-n\theta}}{\prod_{i=1}^{n} x_i!}$$
$$= \left(\prod_{i=1}^{n} x_i!\right)^{-1} \times \theta^{n\bar{x}} e^{-n\theta}.$$

Notice that this likelihood function depends on the data only through $\prod_{i=1}^{n}(x_i!)^{-1}$ and $\bar{x}$.

Further, in $f(\boldsymbol{x}|\theta)$, θ only "interacts" with $\bar{x}$ — the other term simply scales $f(\boldsymbol{x}|\theta)$ — so that, for example, the point at which $f(\boldsymbol{x}|\theta)$ is maximized is determined only by $\bar{x}$.

Sufficiency

Informally, we think of all the information about θ in the data being contained in $\bar{x}$.

More formally, we can show that the distribution of the data given the value $\bar{x}$ does not depend on θ.

Sufficiency

Definition (Statistic)

A **statistic** is any function of the data (and not of unknown parameters).

Definition (Sufficiency)

The statistic $T(\boldsymbol{X})$ is **sufficient** for θ if $f(\boldsymbol{x}|T(\boldsymbol{X}) = t)$ does not depend on θ.

Example 1.8

Consider again the Poisson model in Example 1.5.

Suppose we had just two observations.

Then $n = 2$ and $X_i|\theta \sim Po(\theta)$, $i = 1, 2$ (independent).

Show that $T = X_1 + X_2$ is sufficient for θ. Note that $T|\theta \sim Po(2\theta)$.

Solution to Example 1.8 (1/2)

Consider the conditional distribution of the data $\boldsymbol{X}$ on $T = t$.

As X_1 and X_2 are discrete random variables, we have

$$\begin{aligned} f(\boldsymbol{x}|T(\boldsymbol{X}) = t) &= \frac{\Pr(X_1 = x_1, X_2 = x_2, T = t)}{\Pr(T = t)} \\ &= \frac{\Pr(X_1 = x_1, X_2 = t - x_1)}{\Pr(T = t)}. \end{aligned}$$

Now, as X_1 and X_2 are independent, for the numerator we have:

$$\begin{aligned} \Pr(X_1 = x_1, X_2 = t - x_1|\theta) &= \Pr(X_1 = x_1|\theta)\Pr(X_2 = t - x_1|\theta) \\ &= \frac{e^{-\theta}\theta^{x_1}}{x_1!} \times \frac{e^{-\theta}\theta^{t-x_1}}{(t - x_1)!} \\ &= \frac{e^{-2\theta}\theta^{x_1+t-x_1}}{x_1!(t - x_1)!} = \frac{e^{-2\theta}\theta^{t}}{x_1!(t - x_1)!}. \end{aligned}$$

Solution to Example 1.8 (2/2)

For the denominator we know that

$$\Pr(T = t|\theta) = \frac{e^{-2\theta}(2\theta)^t}{t!} \qquad \text{(sum of two Poissons).}$$

Therefore

$$\begin{aligned} f(\boldsymbol{x}|T(\boldsymbol{X}) = t) &= \frac{e^{-2\theta}\theta^t}{x_1!(t-x_1)!} \Big/ \frac{e^{-2\theta}(2\theta)^t}{t!} \\ &= \frac{t!e^{-2\theta}\theta^t}{x_1!(t-x_1)!e^{-2\theta}(2\theta)^t} \\ &= \frac{t!}{2^t x_1!(t-x_1)!} \end{aligned}$$

which does not depend on θ. Hence, by Definition 1.4, $T = X_1 + X_2$ is sufficient for θ.

Sufficiency

Definition (Joint sufficiency)

The statistics $\underline{\boldsymbol{T}}(\boldsymbol{X}) = \big(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_k(\boldsymbol{X})\big)^T$ are **jointly sufficient** for θ if $f(\boldsymbol{x}|\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t})$ does not depend on θ.

Theorem (Factorisation Theorem)

Under certain regularity conditions

$$\boldsymbol{T}(\boldsymbol{X}) \text{ is sufficient for } \theta \iff f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\, g(\boldsymbol{t}(\boldsymbol{x}), \theta)$$

for some functions h and g.

Example 1.9

Consider again the Poisson model in Example 1.5 with $n = 2$: $X_i|\theta \sim Po(\theta)$, $i = 1, 2$ (independent). Determine a sufficient statistic for θ.

Solution to Example 1.9 (1/1)

The (joint) probability function is

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \frac{e^{-\theta}\theta^{x_1}}{x_1!} \times \frac{e^{-\theta}\theta^{x_2}}{x_2!} \\ &= \frac{e^{-2\theta}\theta^{x_1+x_2}}{x_1!x_2!} \\ &= \frac{1}{x_1!x_2!} \times e^{-2\theta}\theta^{x_1+x_2} \\ &= h(\boldsymbol{x})\, g(x_1 + x_2, \theta), \end{aligned}$$

where $h(\boldsymbol{x}) = 1/(x_1!x_2!)$ and $g(t, \theta) = e^{-2\theta}\theta^t$.

Therefore, by the Factorisation Theorem, $T = X_1 + X_2$ is sufficient for θ.

Example 1.10

Suppose we have a random sample from an exponential distribution: $X_i|\theta \sim Exp(\theta),\ i = 1, 2, \ldots, n$ (independent).

Determine a sufficient statistic for θ.

Solution to Example 1.10 (1/1)

The (joint) probability density function is

$$
\begin{aligned}
f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\
&= \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \\
&= 1 \times \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \\
&= h(\boldsymbol{x})\, g(\Sigma\, x_i, \theta),
\end{aligned}
$$

where $h(\boldsymbol{x}) = 1$ and $g(t, \theta) = \theta^n e^{-\theta t}$. Therefore, by the Factorisation Theorem, $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Example 1.11

Suppose we have a random sample from a Normal distribution: $X_i|\mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ (independent).

Determine sufficient statistics for (μ, σ).

Solution to Example 1.11 (1/1)

The (joint) probability density function is

$$
\begin{aligned}
f(\boldsymbol{x}|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)\right\} \\
&= h(\boldsymbol{x})\, g(\Sigma\, x_i, \Sigma\, x_i^2, \mu, \sigma),
\end{aligned}
$$

where $h(\boldsymbol{x}) = (2\pi)^{-n/2}$ and
$g(t_1, t_2, \mu, \sigma) = \sigma^{-n}\exp\{-(t_2 - 2\mu t_1 + n\mu^2)/(2\sigma^2)\}$.

Therefore, by the Factorisation Theorem, $T_1 = \sum_{i=1}^{n} X_i$ and $T_2 = \sum_{i=1}^{n} X_i^2$ are (jointly) sufficient for μ and σ.

Quick quiz

If $X_1, X_2, \ldots, X_n$ are independent random variables with probability density function

$$f(x|\theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots,$$

(a) Form the likelihood function $f(\boldsymbol{x}|\theta)$;

(b) Use the Factorisation Theorem to obtain a sufficient statistic for θ.

Quick quiz: solution

The joint probability function is

$$
\begin{aligned}
f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \theta(1-\theta)^{x_i-1} \\
&= \theta^n(1-\theta)^{x_1-1}(1-\theta)^{x_2-1} \times \ldots \times (1-\theta)^{x_n-1} \\
&= \theta^n(1-\theta)^{\sum x_i - n} \\
&= 1 \times \theta^n(1-\theta)^{\sum x_i - n} \\
&= h(\boldsymbol{x})\, g\left(\sum x_i, \theta\right)
\end{aligned}
$$

where $h(\boldsymbol{x}) = 1$ and $g(t, \theta) = \theta^n(1-\theta)^{t-n}$. Therefore, by the Factorisation Theorem, $T = \sum X_i$ is sufficient for θ.

Chapter 2

Bayes' Theorem for Distributions

Bayes' Theorem for Distributions

Suppose we have data $\boldsymbol{x}$ which we model using the probability (density) function $f(\boldsymbol{x}|\theta)$, which depends on a single parameter θ.

Once we have observed the data, $f(\boldsymbol{x}|\theta)$

- is the **likelihood function** for θ
- is a function of θ (for fixed $\boldsymbol{x}$) *rather than*
- a function of $\boldsymbol{x}$ (for fixed θ).

Bayes' Theorem for Distributions

Also, suppose we have prior beliefs about likely values of θ expressed by a probability (density) function $\pi(\theta)$.

We can combine both pieces of information using the following version of Bayes Theorem.

The resulting distribution for θ is called the **posterior distribution** for θ as it expresses our beliefs about θ *after* seeing the data.

It summarises all our current knowledge about the parameter θ.

Bayes' Theorem for Distributions

The posterior probability (density) function for θ is

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

where

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

Notice that, 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 for Distributions

Thus, to obtain the posterior distribution, we need:

(1) data, from which we can form the **likelihood** $f(\boldsymbol{x}|\theta)$, and

(2) a suitable distribution, $\pi(\theta)$, that represents our **prior beliefs** about θ.

Some important continuous distributions

Definition (Continuous Uniform distribution)

The random variable Y follows a Uniform $U(a, b)$ distribution if it has probability density function

$$f(y|a, b) = \frac{1}{b-a}, \qquad a \leq y \leq b.$$

This form of probability density function ensures that all values in the range $[a, b]$ are **equally likely**, hence the name "uniform".

This distribution is sometimes called the **rectangular distribution** because of its shape.

You should remember from MAS1604 that

$$E(Y) = \frac{a+b}{2} \quad \text{and} \quad Var(Y) = \frac{(b-a)^2}{12}.$$

Some important continuous distributions

Definition (Beta distribution)

The random variable Y follows a Beta $Be(a, b)$ distribution ($a > 0$, $b > 0$) if it has probability density function

$$f(y|a, b) = \frac{y^{a-1}(1-y)^{b-1}}{B(a, b)}, \qquad 0 < y < 1. \tag{2.1}$$

The constant term $B(a, b)$, also known as the **beta function**, ensures that the density integrates to one. Therefore

$$B(a, b) = \int_0^1 y^{a-1}(1-y)^{b-1}\, dy. \tag{2.2}$$

Some important continuous distributions

It can be shown that the beta function can be expressed in terms of another function, called the **gamma function** $\Gamma(\cdot)$, as

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}\, dx. \tag{2.3}$$

Some important continuous distributions

Tables are available for both $B(a, b)$ and $\Gamma(a)$.

However, these functions are very simple to evaluate when a and b are integers since the gamma function is a generalisation of the factorial function.

In particular, when a and b are integers, we have

$$\Gamma(a) = (a-1)! \quad \text{and} \quad B(a, b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

Some important continuous distributions

It can be shown, using the identity $\Gamma(a) = (a-1)\Gamma(a-1)$, that

$$E(Y) = \frac{a}{a+b}, \qquad \text{and} \qquad Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Also

$$Mode(Y) = \frac{a-1}{a+b-2}, \quad \text{if } a > 1 \text{ and } b > 1.$$

Some important continuous distributions

Definition (Gamma distribution)

The random variable Y follows a Gamma $Ga(a, b)$ distribution ($a > 0$, $b > 0$) if it has probability density function

$$f(y|a, b) = \frac{b^a y^{a-1} e^{-by}}{\Gamma(a)}, \qquad y > 0,$$

where $\Gamma(a)$ is the gamma function defined in Equation (2.3).

It can be shown that

$$E(Y) = \frac{a}{b} \qquad \text{and} \qquad Var(Y) = \frac{a}{b^2}.$$

Also

$$Mode(Y) = \frac{a-1}{b}, \quad \text{if } a \geq 1.$$

Bayes Theorem for distributions in action

We will now see Bayes' Theorem for distributions "in action".

Recall that we have

$$\pi(\theta|\boldsymbol{x}) \quad \propto \quad \pi(\theta) \times f(\boldsymbol{x}|\theta), \qquad \text{i.e.}$$

$$\textbf{posterior} \quad \propto \quad \textbf{prior} \times \textbf{likelihood}$$

Recall also that, for now, we will assume our prior for θ has been given to us – we will look at how such priors are constructed, or **elicited**, in the next chapter.

Example 2.1

Consider an experiment with a possibly biased coin. Let $\theta = \Pr(\text{Head})$.

Suppose that, before conducting the experiment, we believe that all values of θ are equally likely.

This gives a prior distribution $\theta \sim U(0, 1)$, and so

$$\pi(\theta) = 1, \qquad 0 < \theta < 1. \tag{2.4}$$

Note that with this prior distribution $E(\theta) = 0.5$.

We now toss the coin 5 times and observe 1 head. Determine the posterior distribution for θ given this data.

“Full” solution to Example 2.1 (1/4)

The data is an observation on the random variable $X|\theta \sim Bin(5, \theta)$. This gives a likelihood function

$$f(x = 1|\theta) = {}^5\mathrm{C}_1\theta^1(1-\theta)^4$$

$$= 5\theta(1-\theta)^4 \tag{2.5}$$

which favours values of θ near its maximum $\theta = 0.2$ (we observed 1 head out of 5 tosses).

Therefore, we have a conflict of opinions: the prior distribution (2.4) suggests that θ is probably around 0.5 and the data (2.5) suggest that it is around 0.2.

We can use Bayes’ Theorem to combine these two sources of information in a coherent way.

“Full” solution to Example 2.1 (2/4)

Recall the “full” version of Bayes’ Theorem for distributions:

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

where

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

“Full” solution to Example 2.1 (3/4)

First the denominator:

$$
\begin{aligned}
f(x=1) &= \int_{\Theta} \pi(\theta) f(x=1|\theta)\, d\theta \qquad (2.6)\\
&= \int_0^1 1 \times 5\theta(1-\theta)^4\, d\theta \\
&= \int_0^1 \theta \times 5(1-\theta)^4\, d\theta
\end{aligned}
$$

which, using integration by parts, gives

$$
\begin{aligned}
f(x=1) &= \left[-(1-\theta)^5\,\theta\right]_0^1 + \int_0^1 (1-\theta)^5\, d\theta \\
&= 0 + \left[-\frac{(1-\theta)^6}{6}\right]_0^1 \\
&= \frac{1}{6}.
\end{aligned}
$$

“Full” solution to Example 2.1 (4/4)

Therefore, the posterior density is

$$\begin{aligned}
\pi(\theta|x=1) &= \frac{\pi(\theta)f(x=1|\theta)}{f(x=1)} \\
&= \frac{1 \times 5\theta(1-\theta)^4}{1/6}, \qquad 0 < \theta < 1 \\
&= 30\,\theta(1-\theta)^4, \qquad 0 < \theta < 1 \\
&= \frac{\theta^{2-1}(1-\theta)^{5-1}}{B(2,5)}, \qquad 0 < \theta < 1,
\end{aligned}$$

as

$$B(2,5) = \frac{(2-1)!(5-1)!}{(2+5-1)!} = \frac{24}{720} = \frac{1}{30},$$

and so the posterior distribution is $\theta|x=1 \sim Be(2,5)$ – see Equation 2.1. This distribution has its mode at $\theta = 0.2$, and mean at $E[\theta|x=1] = 2/7 = 0.286$.

Example 2.1

There were two "difficulties" in calculating the posterior distribution:

- Obtaining the $f(x)$ term for the denominator in the "full" version of Bayes' Theorem
- Manipulating the result to realise that we had a $Be(2, 5)$ distribution

Example 2.1: "Quick" solution

However, in many cases we can recognise the posterior distribution without the need to calculate $f(x)$.

In this example, we can calculate the posterior distribution as

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta) f(x = 1|\theta) \\
&\propto 1 \times 5\theta(1-\theta)^4, \qquad 0 < \theta < 1 \\
&= k\theta(1-\theta)^4, \qquad 0 < \theta < 1.
\end{aligned}$$

You should be able to recognise this as a $Be(2, 5)$ distribution – we can re–write the above as

$$\pi(\theta|\boldsymbol{x}) = k\theta^{2-1}(1-\theta)^{5-1},$$

i.e. $\theta|x = 1 \sim Be(2, 5)$.

Example 2.1: Summary

It is possible that we have a biased coin.

Recall that if $Y \sim U(a, b)$,

$$E(Y) = \frac{a+b}{2} \qquad \text{and} \qquad Var(Y) = \frac{(b-a)^2}{12}.$$

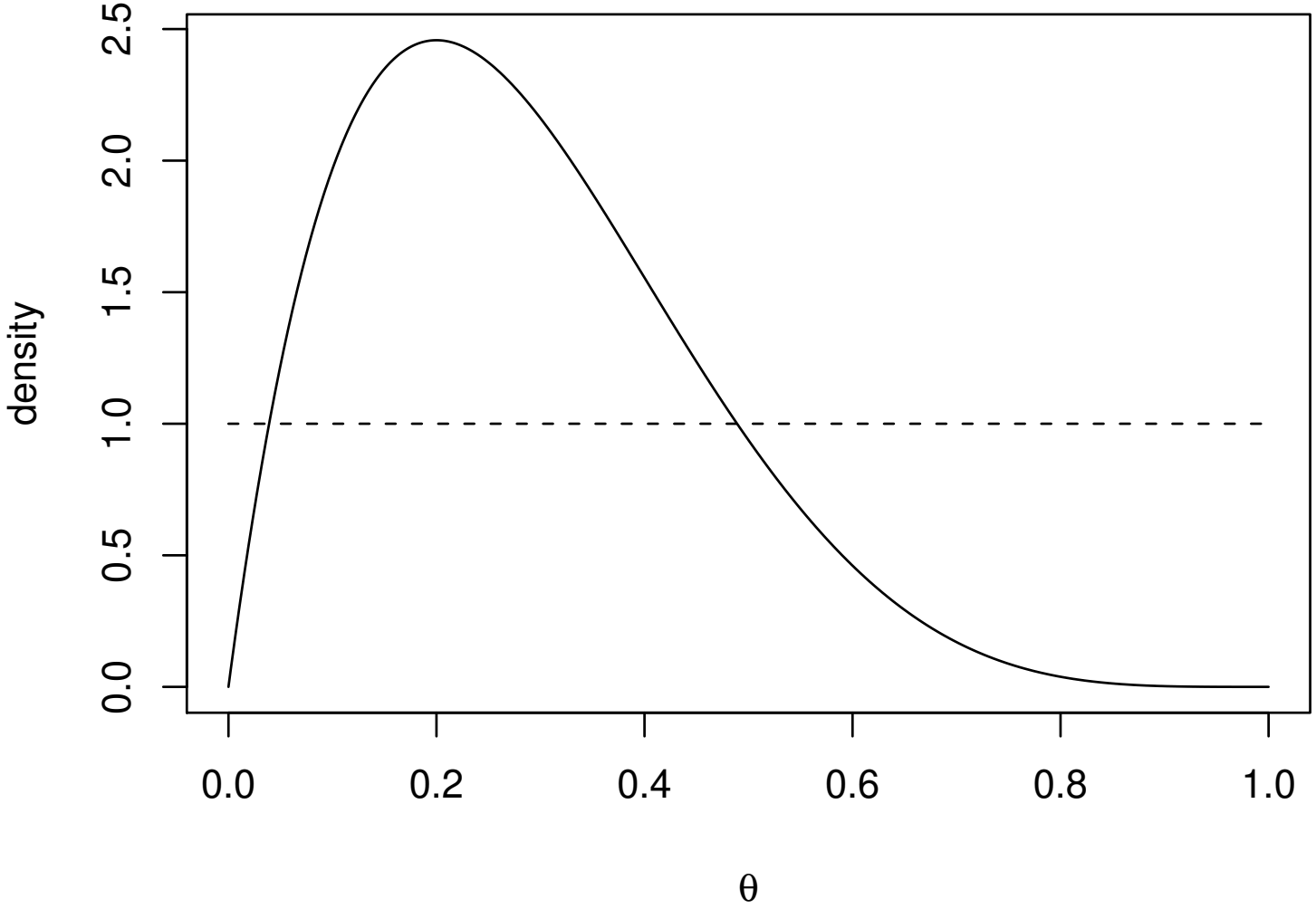
Also, if $Y \sim Be(a, b)$,

$$E(Y) = \frac{a}{a+b} \qquad \text{and} \qquad Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Thus:

	Prior $U(0,1)$	Observed $\theta = 0.2$	Posterior $Be(2,5)$
Mean	0.5	$\longmapsto$	0.286
St. dev.	0.289	$\longmapsto$	0.160

Example 2.1: Summary



Example 2.2

Consider an experiment to determine how good a music expert is at distinguishing between pages from Haydn and Mozart scores. Let $\theta = \Pr(\text{correct choice})$.

Suppose that, before conducting the experiment, we have been told that the expert is very competent.

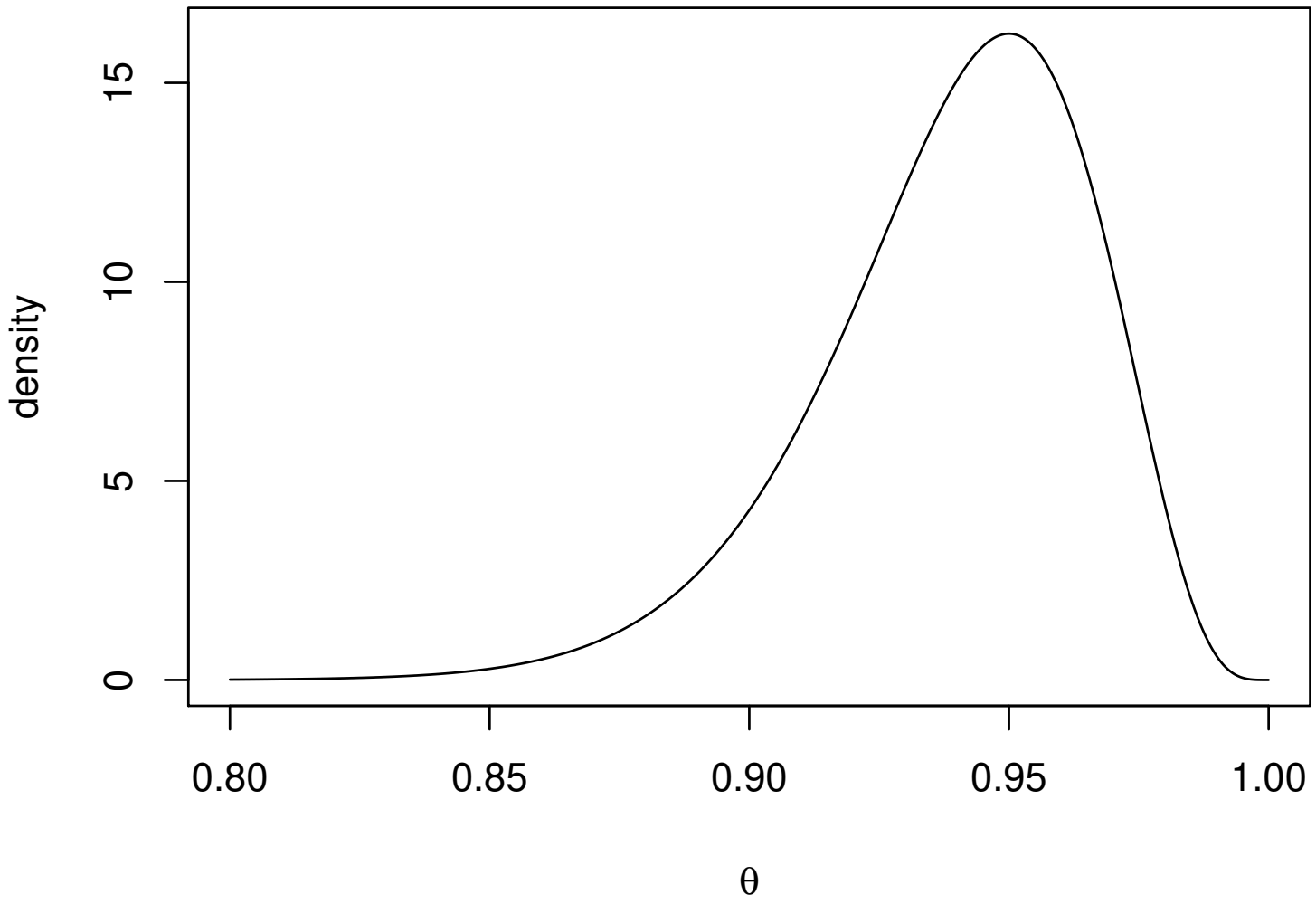
In fact, it is suggested that we should have a prior distribution which has a mode around $\theta = 0.95$ and for which $\Pr(\theta < 0.8)$ is very small.

We choose $\theta \sim Be(77, 5)$ (see Example 3.2, Chapter 3), with probability density function

$$\begin{aligned}\pi(\theta) &= \frac{\theta^{76}(1-\theta)^4}{B(77,5)}, \qquad 0 < \theta < 1 \\ &= 128107980\theta^{76}(1-\theta)^4, \qquad 0 < \theta < 1. \qquad (2.7)\end{aligned}$$

Example 2.2

A graph of this prior density is given in Figure 2.4.



Example 2.2

In the experiment, the music expert makes the correct choice 9 out of 10 times.

Determine the posterior distribution for θ given this information.

Solution to Example 2.2 (1/2)

We have an observation on the random variable $X|\theta \sim Bin(10, \theta)$.

This gives a likelihood function of

$$f(x = 9|\theta) = {}^{10}\mathrm{C}_9 \theta^9 (1-\theta)^1$$

$$= 10\theta^9(1-\theta) \tag{2.8}$$

which favours values of θ near its maximum $\theta = 0.9$.

We combine these two sources of information using Bayes' Theorem.

Solution to Example 2.2 (2/2)

The posterior density function is

$$\pi(\theta|x=9) \propto \pi(\theta) f(x=9|\theta)$$

$$\propto \frac{\theta^{76}(1-\theta)^4}{B(77,5)} \times 10\,\theta^9(1-\theta), \qquad 0<\theta<1$$

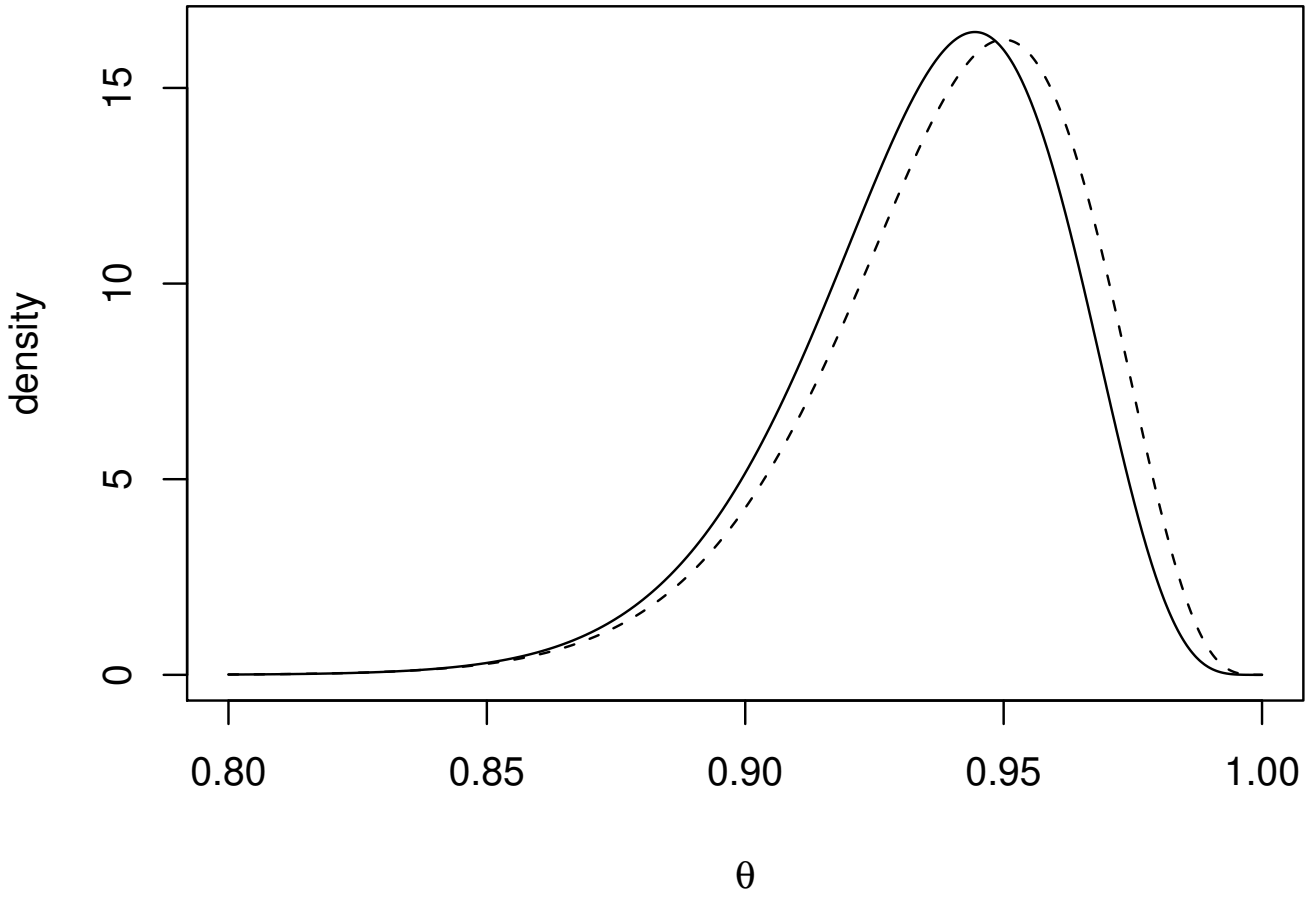
$$= 128107980\theta^{76}(1-\theta)^4 \times 10\,\theta^9(1-\theta), \qquad 0<\theta<1$$

$$= k\theta^{85}(1-\theta)^5, \qquad 0<\theta<1. \tag{2.9}$$

We can recognise this density function as one from the Beta family. In fact, the posterior distribution is $\theta|x=9 \sim Be(86,6)$.

Example 2.2: Summary

The changes in our beliefs about θ are described by the prior and posterior distributions shown in Figure 2.5 and summarised in Table 2.1.



Example 2.2: Summary

	Prior (2.7)	Likelihood (2.8)	Posterior (2.9)
$Mode(\theta)$	0.950	0.900	0.944
$E(\theta)$	0.939	–	0.935
$SD(\theta)$	0.0263	–	0.0256

Notice that, having observed only a 90% success rate in the experiment, the posterior mode and mean are smaller than their prior values.

Also, the experiment has largely confirmed our ideas about θ, with the uncertainty about θ being only very slightly reduced.

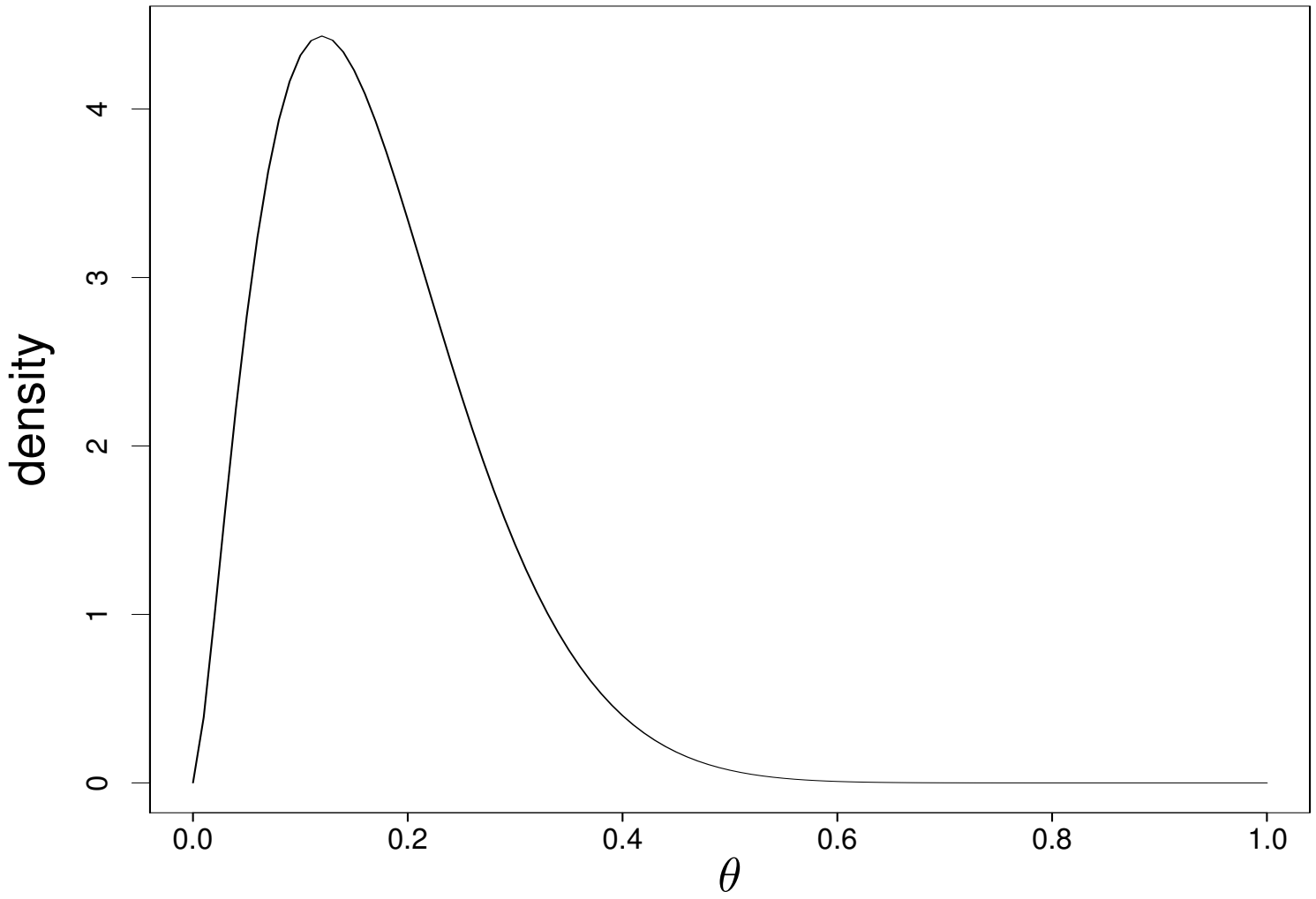
Example 2.3

Max, a video game pirate, is trying to identify the proportion of potential customers θ who might be interested in buying *Call of Duty: Elite* next month.

Based on the proportion of customers who have bought similarly violent games from him in the past, he assumes that $\theta \sim Be(2.5, 12)$ (see Example 3.3, Chapter 3).

Example 2.3

A plot of this prior density is shown in Figure 2.6.



Example 2.3

Max asks five potential customers if they would buy *Call of Duty: Elite* from him, and four say they would.

Using this information, what is Max's posterior distribution for θ?

Solution to Example 2.3 (1/2)

We have been told that the prior for θ is a $Be(2.5, 12)$ distribution – this has density given by

$$\pi(\theta) = \frac{\theta^{2.5-1}(1-\theta)^{12-1}}{B(2.5, 12)}$$

$$= 435.1867\theta^{1.5}(1-\theta)^{11}. \tag{2.10}$$

We have an observation on the random variable $X|\theta \sim Bin(5, \theta)$. This gives a likelihood function of

$$f(x = 4|\theta) = {}^{5}\mathrm{C}_4\theta^4(1-\theta)^1$$

$$= 5\theta^4(1-\theta), \tag{2.11}$$

which favours values of θ near its maximum 0.8.

Solution to Example 2.3 (2/2)

We combine our prior information (2.10) with the data (2.11) – to obtain our posterior distribution – using Bayes' Theorem.

The posterior density function is

$$\pi(\theta|x=4) \propto \pi(\theta)f(x=4|\theta)$$

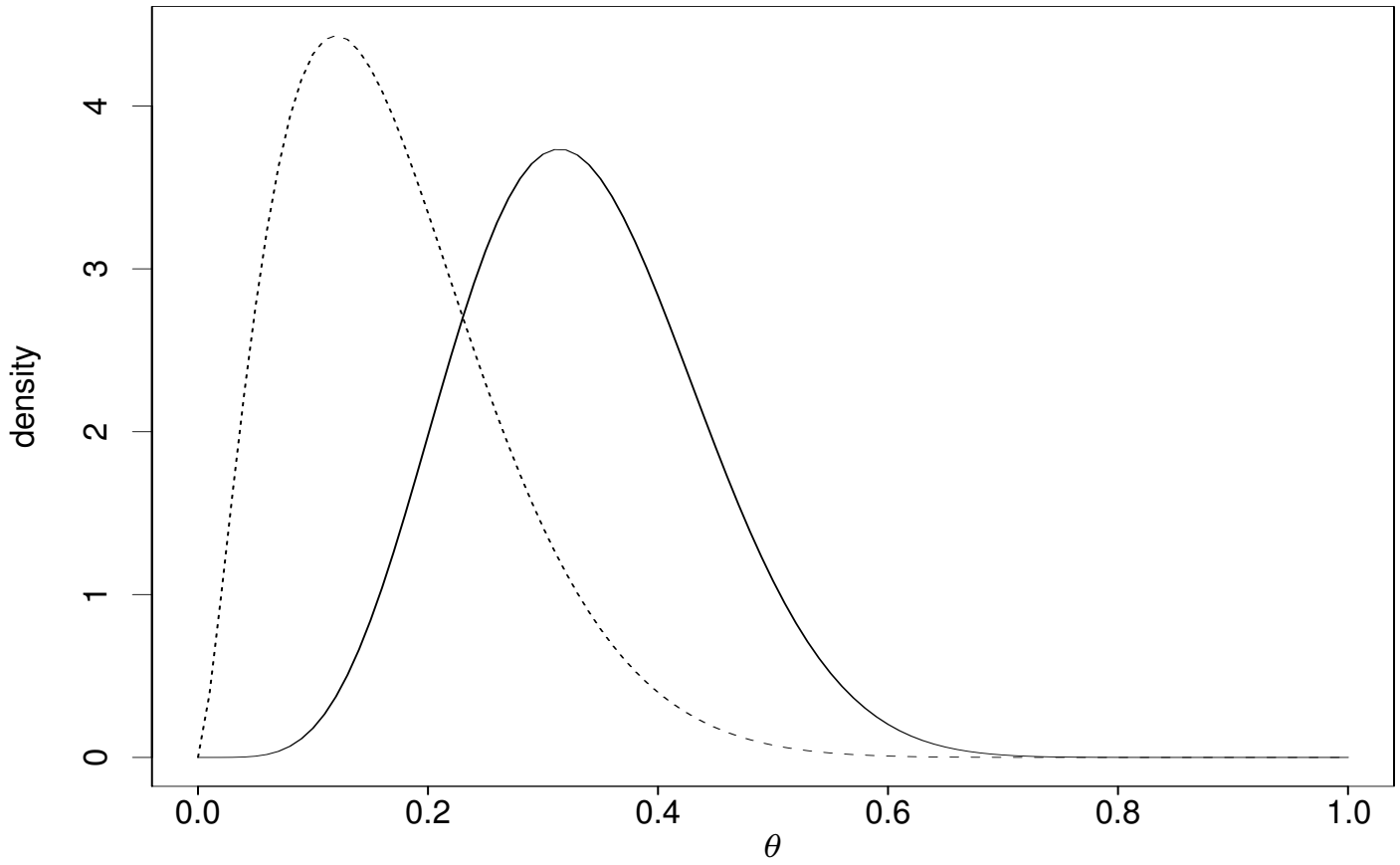
$$\propto 435.1867\theta^{1.5}(1-\theta)^{11} \times 5\theta^{4}(1-\theta), \qquad 0<\theta<1,$$

$$= k\theta^{5.5}(1-\theta)^{12}, \qquad 0<\theta<1. \tag{2.12}$$

You should recognise this density function as one from the beta family. In fact, we have a $Be(6.5, 13)$, i.e. $\theta|x=4 \sim Be(6.5, 13)$.

Example 2.3: Summary

The changes in our beliefs about θ are described by the prior and posterior distributions shown in Figure 2.7 and summarised in Table 2.2.



Example 2.3: Summary

	Prior (2.10)	Likelihood (2.11)	Posterior (2.12)
$Mode(\theta)$	0.12	0.8	0.314
$E(\theta)$	0.172	–	0.333
$SD(\theta)$	0.096	–	0.104

Notice how the posterior has been "pulled" from the prior towards the observed value: the mode has moved up from 0.12 to 0.314, and the mean has moved up from 0.172 to 0.333.

Having just one observation in the likelihood, we see that there is hardly any change in the standard deviation from prior to posterior: we would expect to see a decrease in standard deviation with the addition of more data values.

Announcements

- Assignment drop-in this Thursday, 2pm (questions 4, 9, 10, 11, 12). Good place for hints, tips etc.
- First computer practical this Friday, 1pm, Herschel PC cluster
- ReCap problems...?
- Next week: 2 lectures and a problems class (case study 1)

Example 2.4

Table 2.3 shows some data on the times between serious earthquakes.

An earthquake is included if its magnitude is at least 7.5 on the Richter scale or if over 1000 people were killed.

Recording starts on 16 December 1902 (4500 killed in Turkistan).

The table includes data on 21 earthquakes, that is, 20 “waiting times” between earthquakes.

840	157	145	44	33	121	150	280	434	736
584	887	263	1901	695	294	562	721	76	710

Example 2.4

It is believed that earthquakes happen in a random haphazard kind of way and that times between earthquakes can be described by an exponential distribution.

Data over a much longer period suggest that this exponential assumption is plausible.

Therefore, we will assume that these data are a random sample from an exponential distribution with rate θ (and mean $1/\theta$). The parameter θ describes the rate at which earthquakes occur.

Example 2.4

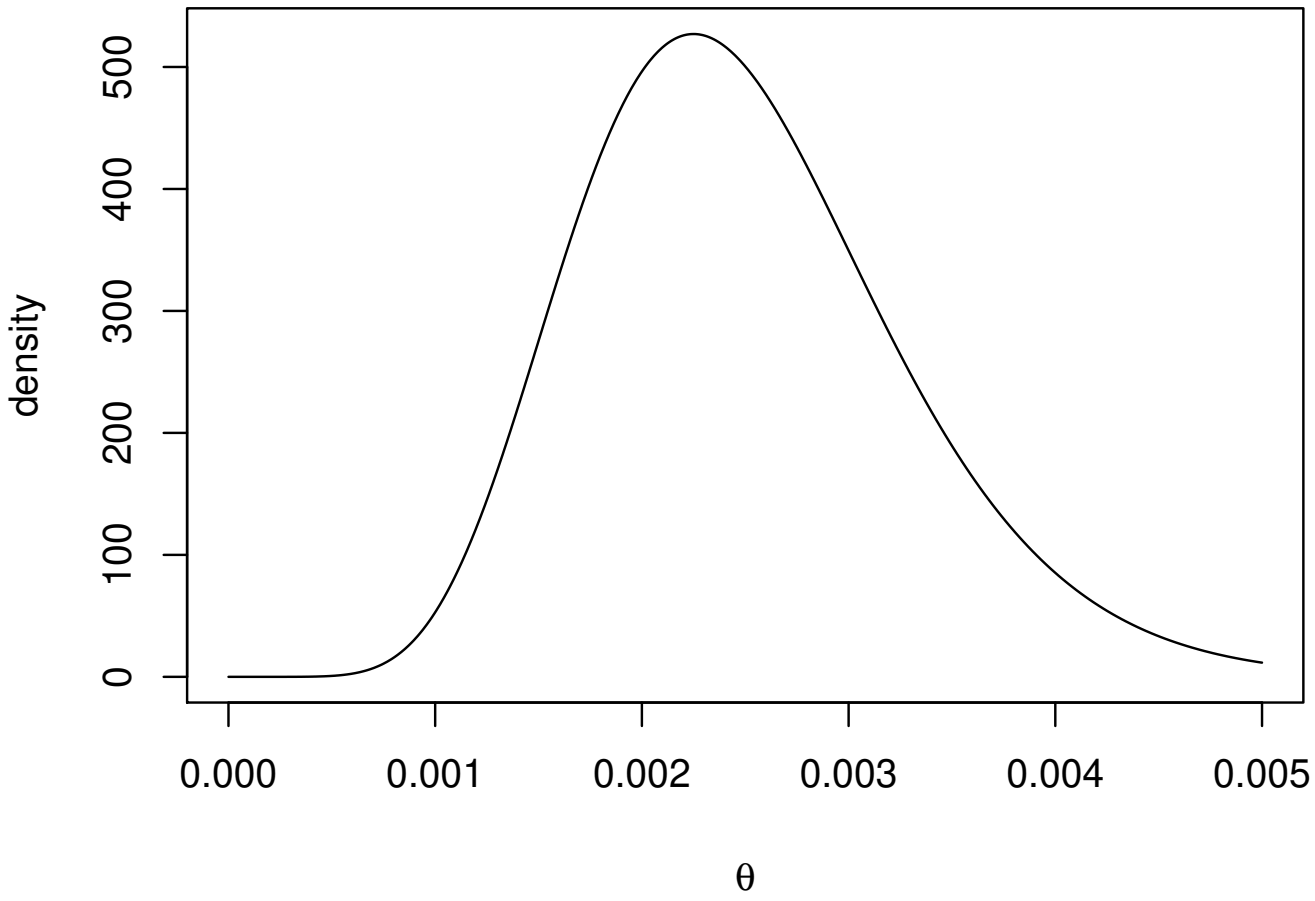
An expert on earthquakes has prior beliefs about the rate of earthquakes, θ, described by a $Ga(10, 4000)$ distribution (see Example 3.1, Chapter 3), which has density density

$$\pi(\theta) = \frac{4000^{10}\,\theta^{9} e^{-4000\theta}}{\Gamma(10)}, \quad \theta > 0, \tag{2.13}$$

and mean $E(\theta) = 0.0025$.

Example 2.4

A plot of this prior distribution can be found in Figure 2.8. As you might expect, the expert believes that only very small values of θ are likely, though larger values are not ruled out!



Solution to Example 2.4 (1/2)

The data are observations on $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, 20$ (independent).

Therefore, the likelihood function for θ is

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{20} \theta e^{-\theta x_i}, \qquad \theta > 0$$

$$= \theta^{20} \exp\left(-\theta \sum_{i=1}^{20} x_i\right), \qquad \theta > 0$$

$$= \theta^{20} e^{-9633\theta}, \qquad \theta > 0. \tag{2.14}$$

Solution to Example 2.4 (2/2)

We now apply Bayes Theorem to combine the expert opinion with the observed data. The posterior density function is

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

$$\propto \frac{4000^{10}\,\theta^{9} e^{-4000\theta}}{\Gamma(10)} \times \theta^{20} e^{-9633\theta}, \qquad \theta > 0$$

$$= k\,\theta^{30-1} e^{-13633\theta}, \qquad \theta > 0. \tag{2.15}$$

The only continuous distribution which takes the form $k\theta^{g-1}e^{-h\theta}$, $\theta > 0$ is the $Ga(g, h)$ distribution.

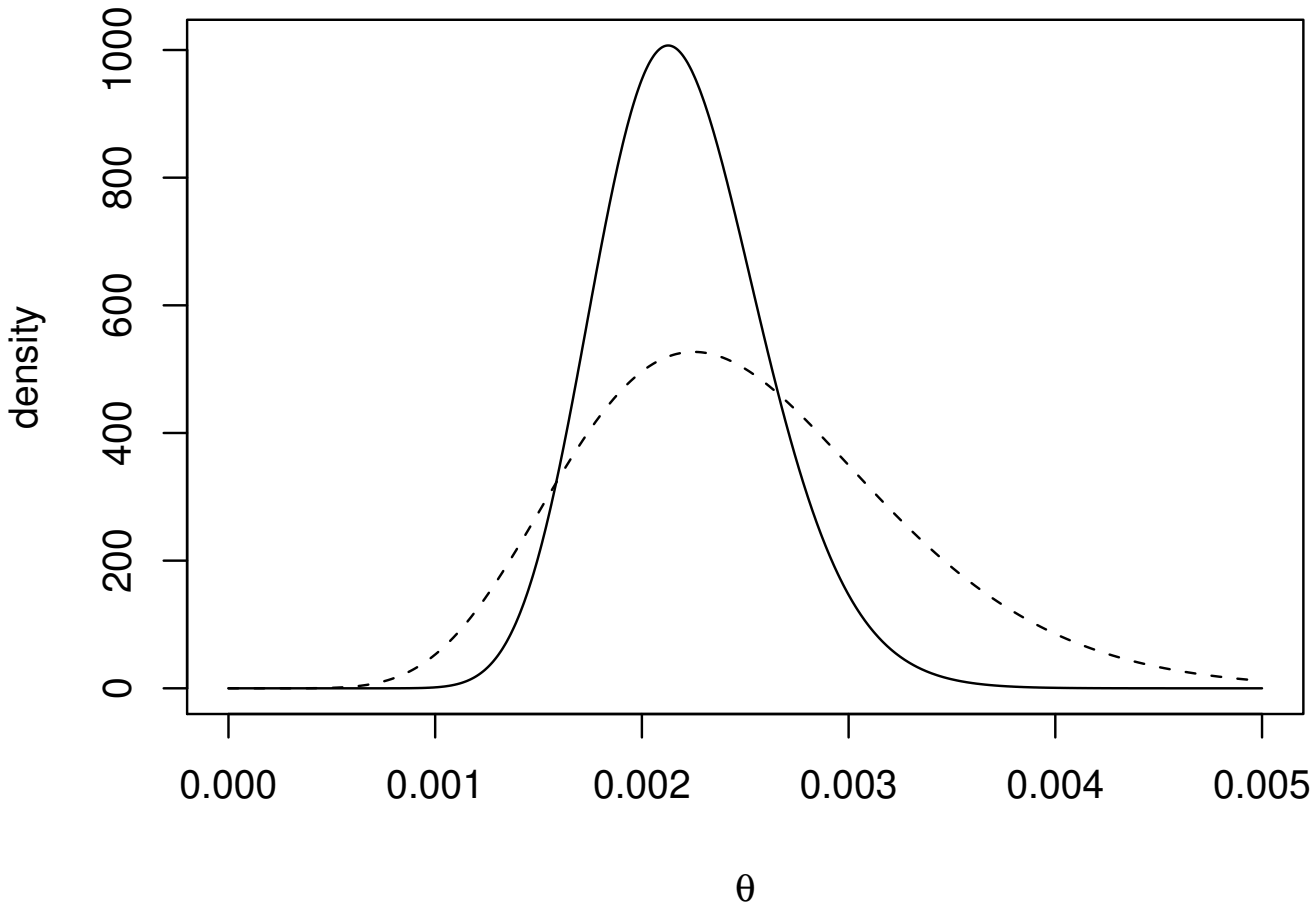
Therefore, the posterior distribution must be $\theta|\boldsymbol{x} \sim Ga(30, 13633)$.

Example 2.4: Summary

The data have updated our beliefs about θ from a $Ga(10, 4000)$ distribution to a $Ga(30, 13633)$ distribution.

Plots of these distributions are given in Figure 2.9, and Table 2.4 gives a summary of the main changes induced by incorporating the data.

Example 2.4: Summary



Example 2.4: Summary

	Prior (2.13)	Likelihood (2.14)	Posterior (2.15)
$Mode(\theta)$	0.00225	0.00208	0.00213
$E(\theta)$	0.00250	–	0.00220
$SD(\theta)$	0.00079	–	0.00040

Notice that, as the mode of the likelihood function is close to that of the prior distribution, the information in the data is consistent with that in the prior distribution.

This results in a reduction in variability from the prior to the posterior distributions.

The similarity between the prior beliefs and the data has reduced the uncertainty we have about the likely earthquake rate θ.

Example 2.5

We now consider the general case of the problem discussed in Example 2.4.

Suppose $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent) and our prior beliefs about θ are summarised by a $Ga(g, h)$ distribution (with g and h known), with density

$$\pi(\theta) = \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)}, \quad \theta > 0. \tag{2.16}$$

Determine the posterior distribution for θ.

Solution to Example 2.5 (1/2)

The likelihood function for θ is

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}, \qquad \theta > 0$$

$$= \theta^n e^{-\theta x_1 - \theta x_2 - \ldots - \theta x_n}$$

$$= \theta^n e^{-\theta \sum x}$$

$$= \theta^n e^{-n\bar{x}\theta}, \qquad \theta > 0. \tag{2.17}$$

Solution to Example 2.5 (2/2)

We now apply Bayes Theorem. The posterior density function is

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

$$\propto \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \theta^n e^{-n\bar{x}\theta}, \qquad \theta > 0.$$

$$\pi(\theta|\boldsymbol{x}) = k\theta^{g+n-1} e^{-(h+n\bar{x})\theta}, \qquad \theta > 0, \tag{2.18}$$

where k is a constant that does not depend on θ.

Therefore, the posterior distribution takes the form $k\theta^{g-1}e^{-h\theta}$, $\theta > 0$ and so must be a gamma distribution.

Thus we have $\theta|\boldsymbol{x} \sim Ga(g+n, h+n\bar{x})$.

Example 2.5: Summary

If

- $X_i \sim Exp(\theta)$ and
- $\theta \sim Ga(g, h)$, then
- $\theta|\boldsymbol{x} \sim Ga(g + n, h + n\bar{x})$

The changes in our beliefs about θ are summarised in Table 2.5, taking $g \geq 1$.

	Prior (2.16)	Likelihood (2.17)	Posterior (2.18)
$Mode(\theta)$	$(g-1)/h$	$1/\bar{x}$	$(g+n-1)/(h+n\bar{x})$
$E(\theta)$	g/h	–	$(g+n)/(h+n\bar{x})$
$SD(\theta)$	$\sqrt{g}/h$	–	$\sqrt{g+n}/(h+n\bar{x})$

Chapter 2 so far...

Model	**Prior**	**Posterior**
Binomial(k, θ)	$\theta \sim U(0, 1)$	Beta
	$\theta \sim Beta(a, b)$	Beta
Exponential(θ)	$\theta \sim Ga(g, h)$	Gamma
	$\theta \sim exp(g)$	??
Normal(μ, σ^2)	$\mu \sim N(b, 1/d)$	Normal

Conjugacy

Definition (Conjugacy)

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 $\mathfrak{F}$.

Example 2.6

Suppose we have a random sample from a Normal distribution.

In Bayesian statistics, when dealing with the Normal distribution, the mathematics is more straightforward if we work with the **precision** ($= 1/\text{variance}$) of the distribution rather than the variance itself.

So we will assume that this population has unknown mean μ but known precision τ: $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent), where τ is known.

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

$$\pi(\mu) = \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \tag{2.19}$$

Determine the posterior distribution for μ.

Solution to Example 2.6 (1/5)

The likelihood function for μ is

$$
\begin{aligned}
f(\boldsymbol{x}|\mu) &= \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x_i-\mu)^2\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}([x_i-\bar{x}]+[\bar{x}-\mu])^2\right\}
\end{aligned}
$$

Solution to Example 2.6 (2/5)

Giving

$$
\begin{aligned}
f(\boldsymbol{x}|\mu) &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right.\right. \\
&\qquad \left.\left. + 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) + \sum_{i=1}^{n}(\bar{x}-\mu)^2\right]\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]\right\},
\end{aligned}
$$

Let

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2;$$

Then

$$f(\boldsymbol{x}|\mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\}. \qquad (2.20)$$

Solution to Example 2.6 (3/5)

Applying Bayes Theorem, the posterior density function is

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

$$\propto \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \times \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2}\left[s^2 + (\bar{x} - \mu)^2\right]\right\}$$

$$= k_1 \exp\left\{-\frac{d}{2}(\mu - b)^2 - \frac{n\tau}{2}\left[s^2 + (\bar{x} - \mu)^2\right]\right\}$$

$$= k_1 \exp\left\{-\frac{1}{2}\left[d(\mu - b)^2 + n\tau(\bar{x} - \mu)^2\right]\right\}$$

where k_1 is a constant that does not depend on μ.

Solution to Example 2.6 (4/5)

We can also simplify the exponent. We have

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

where c does not depend on μ.

Solution to Example 2.6 (5/5)

Let

$$B = \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad D = d + n\tau. \tag{2.21}$$

Then

$$\pi(\mu|\boldsymbol{x}) = k_1 \exp\left\{-\frac{D}{2}(\mu - B)^2 - \frac{c}{2}\right\}$$

$$= k \exp\left\{-\frac{D}{2}(\mu - B)^2\right\}, \tag{2.22}$$

where k is a constant that does not depend on μ.

Therefore, the posterior distribution takes the form $k \exp\{-D(\mu - B)^2/2\}$, $-\infty < \mu < \infty$ and so must be a normal distribution: we have $\mu|\boldsymbol{x} \sim N(B, 1/D)$.

Example 2.6: Summary

If we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known) and our prior beliefs about μ follow a $N(b, 1/d)$ distribution then, after incorporating the data, our (posterior) beliefs about μ follow a $N(B, 1/D)$ distribution.

Example 2.6: Summary

The changes in our beliefs about μ are summarised in Table 2.6.

	Prior (2.19)	Likelihood (2.20)	Posterior (2.21)
$Mode(\mu)$	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	–	$d + n\tau$

Example 2.6: Summary

Notice that the way prior information and observed data combine is through the parameters of the normal distribution:

$$b \to \frac{db + n\tau\bar{x}}{d + n\tau} \qquad \text{and} \qquad d \to d + n\tau.$$

Example 2.6: Summary

Notice also that the posterior variance (and precision) does not depend on the data, and the posterior mean is a convex combination of the prior and sample means, that is,

$$
\begin{aligned}
B &= \frac{db + n\tau\bar{x}}{d + n\tau} \\
&= \frac{db}{d + n\tau} + \frac{n\tau\bar{x}}{d + n\tau} \\
&= \frac{d}{d + n\tau} b + \frac{n\tau}{d + n\tau}\bar{x} \\
&= \left(\frac{d}{d + n\tau}\right) b + \left(\frac{d + n\tau - d}{d + n\tau}\right)\bar{x} \\
&= \alpha b + (1 - \alpha)\bar{x}, \qquad \alpha = \frac{d}{d + n\tau}.
\end{aligned}
$$

Example 2.6: Summary

This equation for the posterior mean, which can be rewritten as

$$E(\mu|\boldsymbol{x}) = \alpha E(\mu) + (1 - \alpha)\bar{x},$$

arises in other models and is known as the **Bayes linear rule** (see Problems Sheet 2, question 3).

Example 2.6: Summary

Notice that the 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

$$E(\mu|\boldsymbol{x}) > E(\mu) \quad \Longleftrightarrow \quad Mode[f(\boldsymbol{x}|\mu)] > E(\mu).$$

Also, the standard deviation of the posterior distribution is smaller than that of the prior distribution.

Example 2.7

The ages of *Ennerdale granophyre* rocks can be determined using the relative proportions of rubidium–87 and strontium–87 in the rock.

An expert in the field suggests that the ages of such rocks (in millions of years) $X|\mu \sim N(\mu, 8^2)$ and that a prior distribution $\mu \sim N(370, 20^2)$ is appropriate.

A rock is found whose chemical analysis yields $x = 421$. What is the posterior distribution for μ and what is the probability that the rock will be older than 400 million years?

Solution to Example 2.7 (1/2)

We have $n = 1$, $\bar{x} = x = 421$, $\tau = 1/64$, $b = 370$ and $d = 1/400$. Therefore, using the results in Example 2.6,

$$\begin{aligned} B &= \frac{db + n\tau\bar{x}}{d + n\tau} \\ &= \frac{370/400 + 421/64}{1/400 + 1/64} = 414.0 \qquad \text{and} \end{aligned}$$

$$\begin{aligned} D &= d + n\tau \\ &= 1/400 + 1/64 = 1/7.43^2. \end{aligned}$$

Solution to Example 2.7 (2/2)

Thus, the posterior distribution is $\mu|x = 421 \sim N(414.0, 7.43^2)$.

The (posterior) probability that the rock will be older than 400 million years is

$$\Pr(\mu > 400|x = 421) = 0.9702$$

calculated using the R commands `1-pnorm(400,414,7.43)` or `1-pnorm(-1.884)` (or indeed by hand using tables).

Without the chemical analysis, the only basis for determining the age of the rock is via the prior distribution.

The (prior) probability that the rock will be older than 400 million years is $\Pr(\mu > 400) = 0.0668$ calculated using the R command `1-pnorm(400,370,20)`.

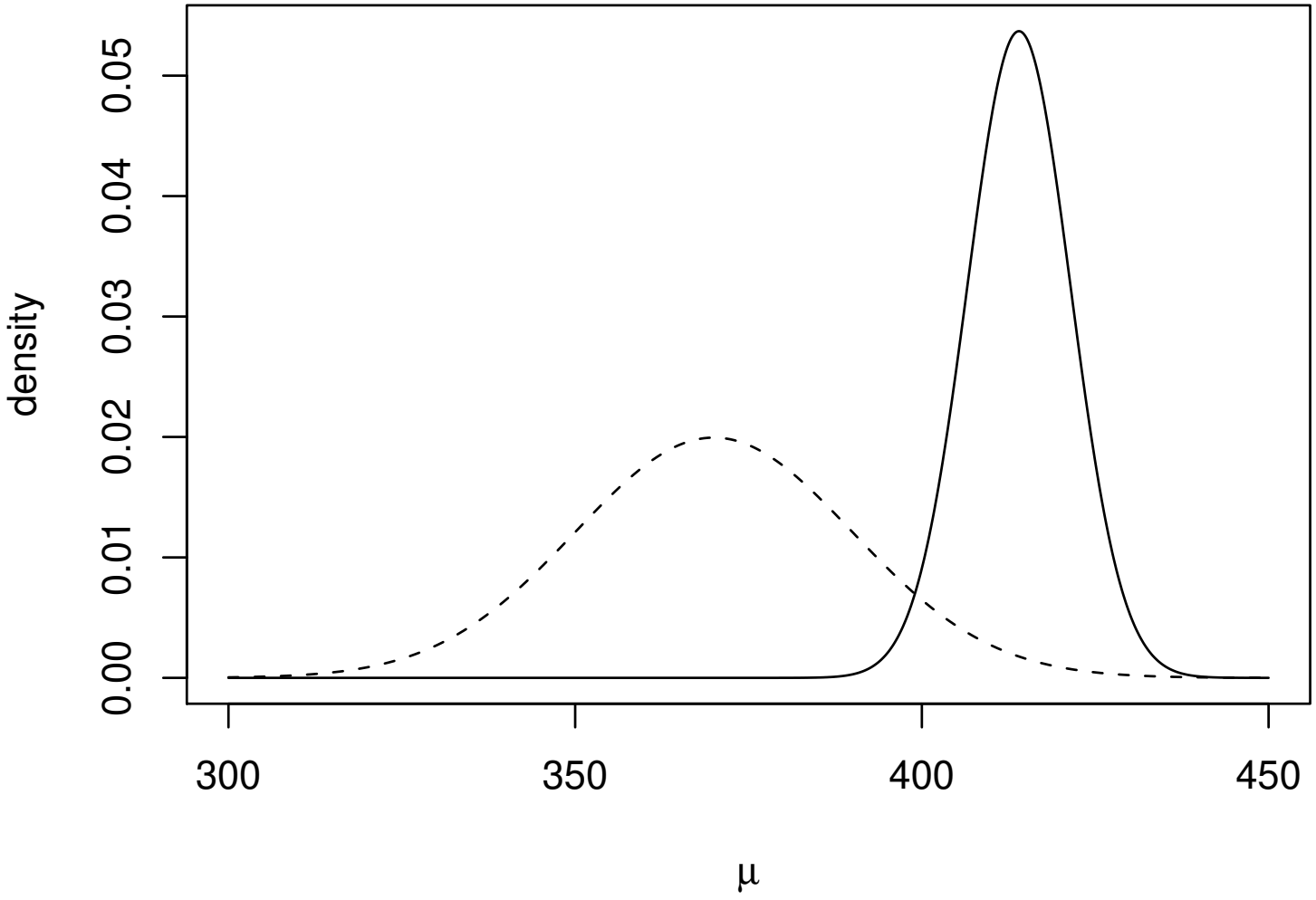
Example 2.7: Summary

This highlights the benefit of taking the chemical measurements.

Note that the large difference between these probabilities is not necessarily due to the expert's prior distribution being inaccurate, *per se*.

It is probably due to the large prior uncertainty about rock ages, as shown in Figure 2.10.

Example 2.7: Summary



Posterior distributions and sufficient statistics

We have already met the concept of sufficient statistics. Not surprisingly they also play a role in Bayesian Inference.

Suppose that we have data $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ available and we want to make inferences about the parameter θ in the statistical model $f(\boldsymbol{x}|\theta)$.

If T is a sufficient statistic then by the Factorisation Theorem

$$f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\, g(t, \theta).$$

Posterior distributions and sufficient statistics

Therefore, using Bayes Theorem,

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\ &\propto \pi(\theta)\, h(\boldsymbol{x})\, g(t,\theta) \\ &\propto \pi(\theta)\, g(t,\theta).\end{aligned}$$

It can be shown that, up to a constant not depending on θ, $g(t,\theta)$ is equal to the probability (density) function of T; that is

$$g(t,\theta) \propto f_T(t|\theta).$$

Hence

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta)\, f_T(t|\theta).$$

Posterior distributions and sufficient statistics

However, applying Bayes Theorem to the data t gives

$$\pi(\theta|t) \quad \propto \quad \pi(\theta)\, f_T(t|\theta)$$

and so, since

$$\pi(\theta|\boldsymbol{x}) \quad \propto \quad \pi(\theta|t)$$

and both are probability (density) functions, we have

$$\pi(\theta|\boldsymbol{x}) \quad = \quad \pi(\theta|t).$$

Posterior distributions and sufficient statistics

Therefore, our (posterior) beliefs about θ having observed the full data $\boldsymbol{x}$ are the same as if we had observed only the sufficient statistic T.

This is what we would expect if all the information about θ in the data were contained in the sufficient statistic.

Example 2.8

Suppose we have a random sample from an exponential distribution with a gamma prior distribution, that is, $X_i|\theta \sim Exp(\theta)$, $i = 1, 2, \ldots, n$ (independent) and $\theta \sim Ga(g, h)$.

Determine a sufficient statistic T for θ and verify that $\pi(\theta|\boldsymbol{x}) = \pi(\theta|t)$.

Solution to Example 2.8 (1/2)

The density of the data is

$$
\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \theta e^{-\theta x_i} \\
&= \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \\
&= 1 \times \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) \\
&= h(\boldsymbol{x})\, g(\Sigma x_i, \theta)
\end{aligned}
$$

and therefore, by the Factorisation Theorem, $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Solution to Example 2.8 (2/2)

Now $T|\theta \sim Ga(n, \theta)$ and so

$$f_T(t|\theta) = \frac{\theta^n t^{n-1} e^{-\theta t}}{\Gamma(n)}, \quad \theta > 0.$$

Also

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

Therefore, by Bayes Theorem,

$$\begin{aligned} \pi(\theta|t) &\propto \pi(\theta) f(t|\theta) \\ &\propto \frac{h^g \, \theta^{g-1} e^{-h\theta}}{\Gamma(g)} \times \frac{\theta^n t^{n-1} e^{-\theta t}}{\Gamma(n)}, \qquad \theta > 0 \\ &\propto \theta^{g+n-1} e^{-(h+t)\theta}, \qquad \theta > 0 \end{aligned}$$

and so the posterior distribution is $\theta|t \sim Ga(g+n, h+t)$. This is the same as the result we obtained previously for $\theta|\boldsymbol{x}$.

Example 2.9

Suppose we have a random sample from a normal distribution with known variance and a normal prior distribution for the mean parameter, that is, $X_i|\mu \sim N(\mu, 1/\tau)$, $i = 1, 2, \ldots, n$ (independent) and $\mu \sim N(b, 1/d)$.

Determine a sufficient statistic T for μ and verify that $\pi(\mu|\boldsymbol{x}) = \pi(\mu|t)$.

Solution to Example 2.9 (1/3)

Recall from Equation (2.20) that

$$
\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}|\mu) &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau}{2}\left[s^2 + (\bar{x}-\mu)^2\right]\right\} \\
&= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{n\tau s^2}{2}\right\} \times \exp\left\{-\frac{n\tau}{2}(\bar{x}-\mu)^2\right\} \\
&= h(\boldsymbol{x})\, g(\bar{x}, \mu)
\end{aligned}
$$

and therefore, by the Factorisation Theorem, $T = \bar{X}$ is sufficient for μ.

Solution to Example 2.9 (2/3)

Now $T|\mu \sim N(\mu, 1/(n\tau))$ (by the Central Limit Theorem) and so

$$f_T(t|\mu) = \left(\frac{n\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{n\tau}{2}(t-\mu)^2\right\}.$$

Also

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

Solution to Example 2.9 (3/3)

Therefore, by Bayes' Theorem,

$$
\begin{aligned}
\pi(\mu|t) &\propto \pi(\mu)f(t|\mu) \\
&\propto \left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu-b)^2\right\} \\
&\qquad\qquad \times \left(\frac{n\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{n\tau}{2}(t-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{d}{2}(\mu-b)^2 - \frac{n\tau}{2}(t-\mu)^2\right\} \\
&\quad \vdots \\
&\propto \exp\left\{-\frac{D}{2}(\mu-B)^2\right\}
\end{aligned}
$$

where B and D are as in Equation (2.21), with t replacing $\bar{x}$; that is, $\mu|t \sim N(B, 1/D)$, the same distribution as $\mu|\boldsymbol{x}$.

Chapter 3

Prior elicitation

Introduction

In this chapter we will think about how to **construct** a suitable **prior distribution** $\pi(\theta)$ for our parameter of interest θ.

For example:

- Why did we use a $Be(77, 5)$ distribution for θ in the music expert example?
- Why did we use a $Be(2.5, 12)$ distribution for θ in the video game pirate example?
- Why did we assume a $Ga(10, 4000)$ distribution for θ in the earthquake example?

Introduction

Prior elicitation – the process by which we attempt to construct the most suitable prior distribution for θ – is a huge area of research in Bayesian Statistics.

The aim in this course is to give a brief (and relatively simple) overview.

Introduction

We will

- have another case study lecture (before Easter)
- have an interactive music session (using the TURNINGPOINT voting system, after Easter)

to demonstrate the techniques.

Introduction

We will consider the cases of:

- **Substantial prior knowledge**
  - turning expert opinion into a probability distribution for θ
  - re–visit the examples about the music expert, the video game pirate and earthquakes in Chapter 2.
- **Vague prior knowledge**
  - No expert available
  - Choose a prior which "makes sense" and keeps the maths simple!
- **Prior ignorance**
  - Assume all values of θ are equally likely

Substantial prior knowledge

We will consider various methods for constructing prior distributions when we have **substantial prior knowledge**:

- Use of **suggested prior summaries**
- The **trial roulette method**
- The **bisection method**

We will make use of an online elicitation tool called `MATCH`, courtesy of Professor Tony O'Hagan and Dr Jeremy Oakley (Sheffield University).

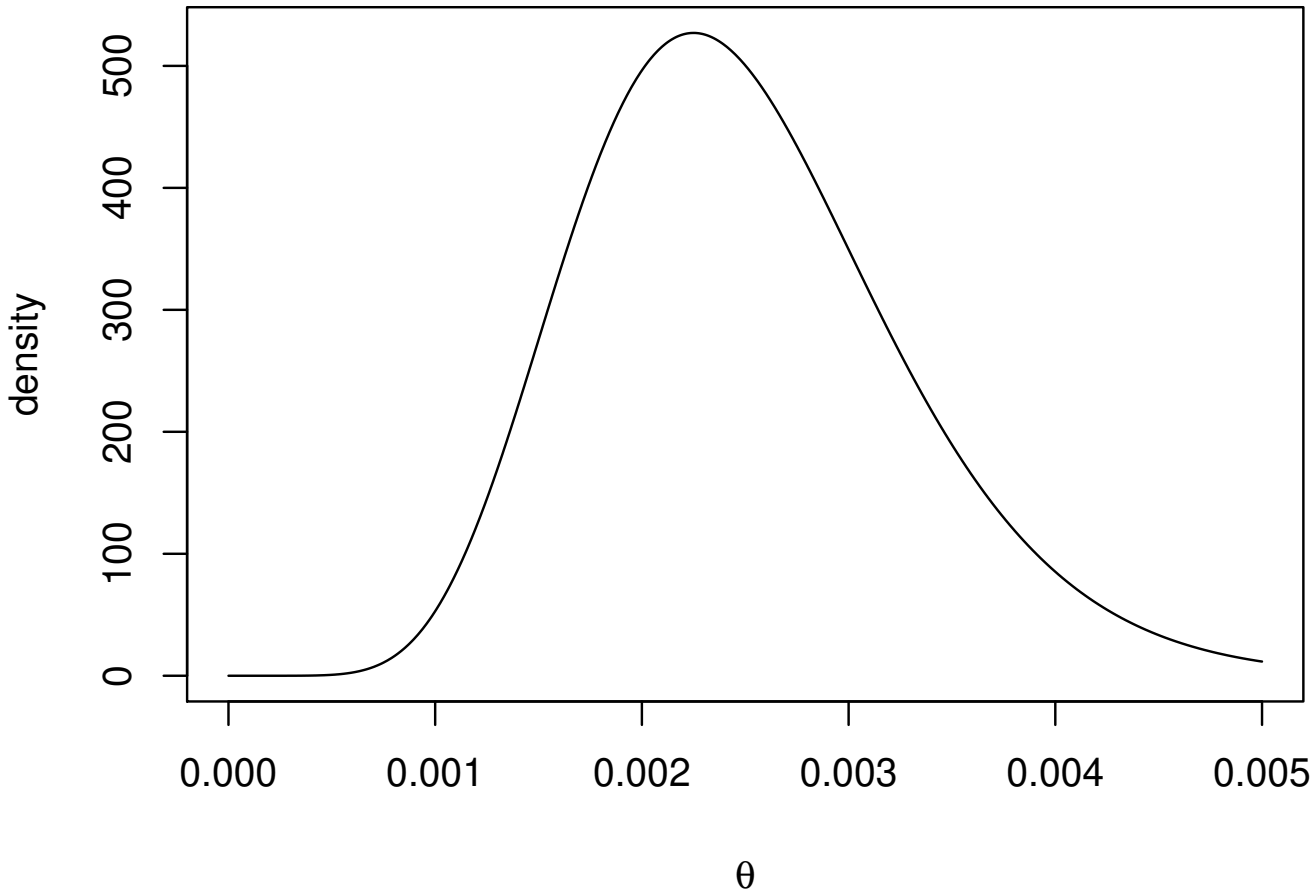
Example 3.1: Using suggested prior summaries

Let us return to **Example 2.4** of Chapter 2.

Recall that we were given some data on the "waiting times", in days, between 21 earthquakes, and we discussed why an exponenetial distribution $Exp(\theta)$ might be appropriate to model the waiting times.

Further, we were told that an **expert on earthquakes** has prior beliefs about the rate θ, described by a $Ga(10, 4000)$ distribution.

Example 3.1: Using suggested prior summaries



How did we get from the expert's beliefs to a $Ga(10, 4000)$?

Example 3.1: Using suggested prior summaries

Suppose the expert tells us that earthquakes in the region we are interested in usually occur **less than once per year**.

In fact, they occur on average **once every 400 days**.

This gives us a rate of occurrence of about 1/400 = 0.0025 per day.

Further, he is fairly certain about this and specifies a **very small variance** of 6.25×10^{-7}.

Example 3.1: Using suggested prior summaries

A $Ga(a, b)$ distribution seems sensible, since we can't observe a negative daily earthquake rate and the Gamma distribution is specified over positive values only.

Using the information provided by the expert, verify our use of $a = 10$ and $b = 4000$.

Solution to Example 3.1 (1/1)

We know that, if $\theta \sim Ga(a, b)$, then $E(\theta) = a/b$ and $Var(\theta) = a/b^2$. Thus

$$\frac{a}{b} = 0.0025 \implies a = 0.0025b.$$

Substituting into $a/b^2 = 0.000000625$ gives

$$\frac{0.0025b}{b^2} = 0.000000625, \qquad \text{giving}$$

$$b = 4000.$$

Thus $a = 0.0025 \times 4000 = 10$.

Example 3.2: Using suggested prior summaries

Now let us return to **Example 2.2** of Chapter 2.

We considered an experiment to determine how good a music expert is at distinguishing between pages from Haydn and Mozart scores.

When presented with a score from each composer, the expert makes the correct choice with probability θ.

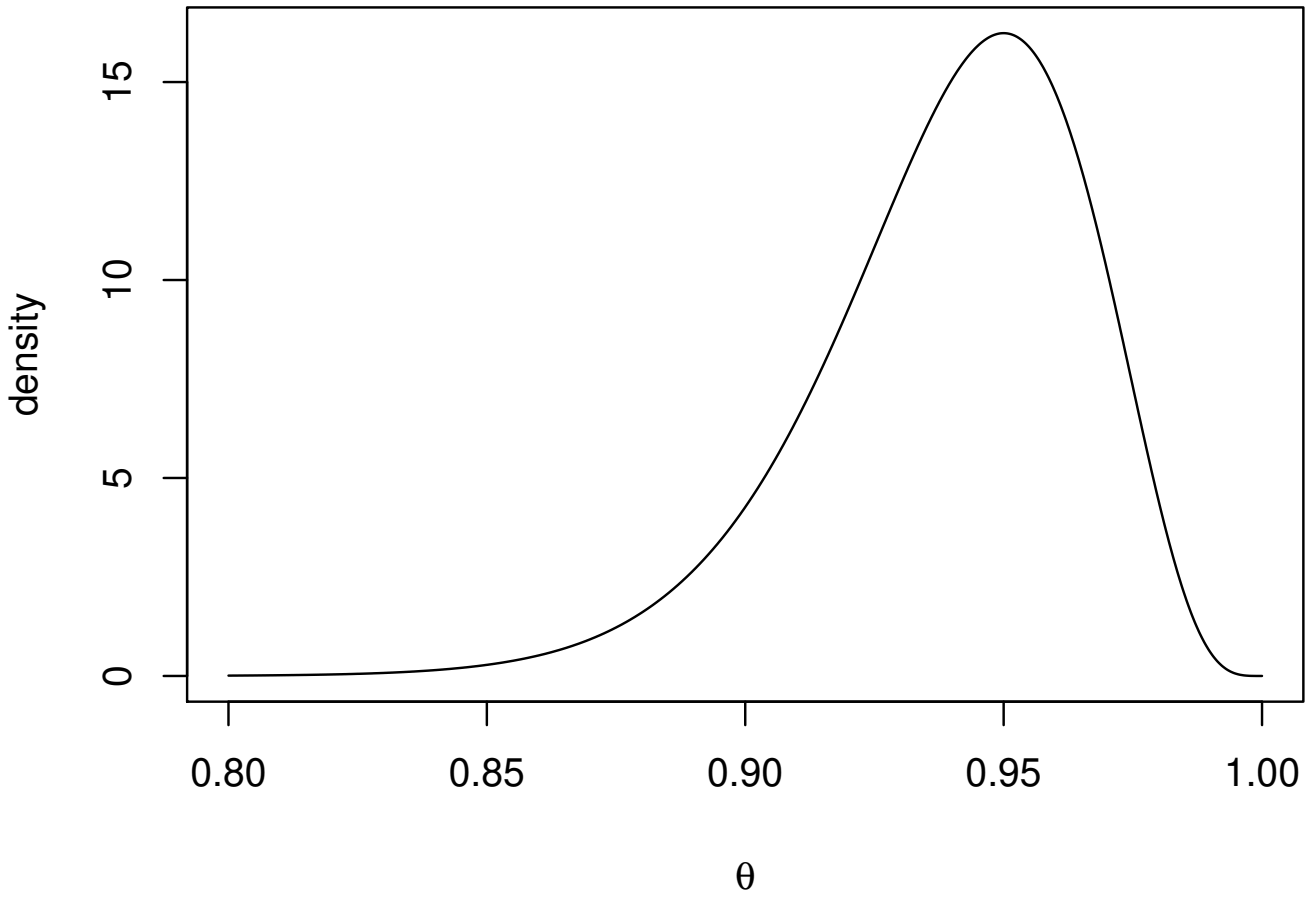
Example 3.2: Using suggested prior summaries

Before conducting the experiment, we were told that the expert is very competent. In fact, we were told that

- θ should have a prior distribution **peaking at around 0.95**
- $\Pr(\theta < 0.8)$ should be **very small**

To achieve this, we assumed that $\theta \sim Be(77, 5)$, with density given in Figure 2.4.

Example 3.2: Using suggested prior summaries



How did we know a $Be(77, 5)$ would work?

Example 3.2: Using suggested prior summaries

We are told that the mode of the distribution should be around 0.95. Using the formulae on **page 23**, we can write

$$\begin{aligned}
\frac{a-1}{a+b-2} &= 0.95 \\
\Rightarrow a-1 &= 0.95(a+b-2) \\
\Rightarrow a-0.95a &= 0.95b-1.9+1 \\
\Rightarrow 0.05a &= 0.95b-0.9 \\
\Rightarrow a &= 19b-18.
\end{aligned}$$

Example 3.2: Using suggested prior summaries

We are also told that $\Pr(\theta < 0.8)$ must be small.

In fact, suppose we are told that $\theta < 0.8$ might occur with probability 0.0001.

This means that if we integrate the probability density function for our beta distribution between 0 and 0.8, we would get 0.0001; from Equation (2.1) on **page 23**, we can write this as

$$\int_0^{0.8} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} d\theta = 0.0001, \qquad \text{i.e.}$$

$$\int_0^{0.8} \frac{\theta^{(19b-18)-1}(1-\theta)^{b-1}}{B(19b-18,b)} d\theta = 0.0001.$$

Example 3.2: Using suggested prior summaries

In other words, we set the cumulative distribution function for a $Be(19b - 18, b)$, evaluated at 0.8, equal to 0.0001 and solve for b.

Although this would be tricky to do by hand, we can do this quite easily in R.

Recall that the R command:

- `dbeta(x,a,b)` evaluates the **density** of the $Be(a, b)$ distribution at the point `x`
- `pbeta(x,a,b)` evaluates the **cumulative distribution function** at `x`

Example 3.2: Using suggested prior summaries

First of all, we re–write (3.3) to set it equal to zero:

$$\int_0^{0.8} \frac{\theta^{(19b-18)-1}(1-\theta)^{b-1}}{B(19b-18, b)} d\theta - 0.0001 = 0. \qquad (3.4)$$

We then write a function in R which computes the left–hand–side of Equation (3.4):

```
f=function(b){
     answer=pbeta(0.8,((19*b)-18),b)-0.0001
     return(answer)}
```

Example 3.2: Using suggested prior summaries

The trick now is to use a **numerical procedure** to find the root of `answer` in our R function.

In other words, find the value `b` which equates `answer` to zero.

The R function `uniroot(f, lower=, upper=)`

- uses a numerical search algorithm to find the root of the expression provided by the function `f`
- requires the user to provide a `lower` and `upper` bound to search within

We know from the formulae on **page 23** that $a, b > 1$ when using expression (3.1) for the mode $\rightarrow$ so we search for a root over some specified domain > 1.

Example 3.2: Using suggested prior summaries

For example, we might use `lower=1` and `upper=100`, giving:

```
> uniroot(f,lower=1,upper=100)
$root
[1] 5.06513

$f.root
[1] 6.008134e-09

$iter
[1] 14

$estim.prec
[1] 6.103516e-05
```

Example 3.2: Using suggested prior summaries

Thus, the solution to Equation (3.3) is $b = 5.06513$.

For simplicity, rounding down to $b = 5$ and then substituting into (3.2) gives

$$a = 19 \times 5 - 18 = 77,$$

hence the use of $\theta \sim Be(77, 5)$ in Example 2.2 in Chapter 2.

Example 3.3: Trial roulette method

We now return to **Example 2.3** in Chapter 2.

Recall that Max is a video game pirate, and he is trying to identify the proportion θ of potential customers who might be interested in buying *Call of Duty: Elite* next month.

Why did we use $\theta \sim Be(2.5, 12)$?

Example 3.3: Trial roulette method

For each month over the last two years Max knows the proportion of his customers who have bought similar games; these proportions are shown below in Table 3.1.

0.32	0.25	0.28	0.15	0.33	0.12	0.14	0.18	0.12	0.05	0.25	0.08
0.07	0.16	0.24	0.38	0.18	0.15	0.22	0.05	0.01	0.19	0.08	0.15

Example 3.3: Trial roulette method

The **trial roulette method** of elicitation works in the following way:

- Divide the sample space for θ into m "bins"
- Ask the expert/person "in the know" to distribute n "chips" amongst the bins
- The proportion of chips in a particular bin represents the probability that θ lies in that bin
- Done graphically, we can see the shape of the distribution forming as the expert allocates the chips
- We then find a model that closely matches the distribution of chips

Example 3.3: Trial roulette method

We will implement the trial roulette method using the **MATCH Uncertainty Elicitation Tool**.

This was developed by **Dr. Jeremy Oakley** and **Professor Tony O'Hagan** at Sheffield University.

This can be accessed via any web browser:

http://optics.eee.nottingham.ac.uk/match/uncertainty.php

Example 3.4: Bisection Method

Over the past 15 years there has been considerable scientific interest in the rate of retreat, θ (feet per year), of glaciers in Greenland (as discussed in the recent *Frozen Planet* series shown on the BBC).

Indeed, this has often been used as an indicator of **global warming**.

We are interested in eliciting a suitable prior distribution for θ for the *Zachariae Isstrøm* glacier in Greenland.

Example 3.4: Bisection Method

Records from an expert glaciologist show that glaciers in Greenland have been retreating at a rate of between 0 and 70 feet per year since 1995.

We will use these values as the lower and upper limits for θ, respectively. We now attempt to elicit the **median** and **quartiles** for θ from the glaciologist.

Example 3.4: Bisection Method

Step 1: Eliciting the median

Ask the expert to provide a value m (in the range of permissable values for θ), such that

$$\Pr(\text{minimum} < \theta < m) = \Pr(m < \theta < \text{maximum}) = \frac{1}{2}.$$

- The value m **bisects** the range for θ into two halves of equal probability
- If the expert is "statistically aware", it might be possible to ask them for their **median** for θ
- Otherwise, m might be the value that the expert believes θ is most likely to take.

Example 3.4: Bisection Method

Step 1: Eliciting the median from the glaciologist

- Glaciers in Greenland have been retreating at a rate of between 0 and 70 feet per year since 1995, depending on how far north the glacier is. Thus, we will say that $\theta \in (0, 70)$.
- The *Zachariae Isstrøm* glacier lies in quite a northerly location, so is not quite so prone to rapid retreat.
- The glaciologist specifies that $m = 24$ might be suitable for bisecting the range for θ – notice how m is closer to the lower bound than the upper.

Example 3.4: Bisection Method

Step 2: Eliciting the lower quartile

Ask the expert to provide a value ℓ such that

$$\Pr(\text{minimum} < \theta < \ell) = \Pr(\ell < \theta < m),$$

i.e. ℓ **bisects** the lower half of the range for θ.

- This can be more tricky for the expert to do – it's not quite so intuitive a task.
- If the expert struggles, help him/her a bit:
  - Split the lower half into two, and ask them in which part θ is most likely to occur
  - Then ℓ should probably lie in the part which is more likely to occur
- Note that the more certain the expert is, the closer ℓ will be to m

Example 3.4: Bisection Method

Step 2: Eliciting the lower quartile from the glaciologist

The glaciologist found this task a bit more difficult...

- We ask the expert whether $[0, 12]$ or $[12, 24]$ is more likely
- The expert is fairly sure about $m = 24$, so says $[12, 24]$ is more likely for θ than $[0, 12]$
  - Areas further north than the *Zachariae Isstrøm* glacier have much slower rates of retreat
  - Only the most northerly glaciers have zero retreat
- Focussing on $[12, 24]$, the glaciologist settles on $\ell = 19$.

Example 3.4: Bisection Method

Step 3: Eliciting the upper quartile

Same sort of process for u as for ℓ.

Step 3: Eliciting the lower quartile from the glaciologist

Using the same process as for ℓ, the glaciologist settles on $u = 30$.

Example 3.4: Bisection Method

Step 4: Reflection

Based on the elicited values for ℓ, m and u, the expert should be asked to **reflect**, i.e., does the following seem plausible:

$$\Pr(\min < \theta < \ell) = \Pr(\ell < \theta < m) = \Pr(m < \theta < u) = \Pr(u < \theta < \max)?$$

Step 4: Let the glaciologist reflect

The glaciologist seems fine with this!

Example 3.4: Bisection Method

Step 5: Fit a parametric distribution to these judgements

We can use the *MATCH* software for this.

Step 5: Fitting a parametric distribution to the glaciologist's judgements

Doing this in *MATCH* gives $\theta \sim Ga(9, 0.36)$.

Example 3.4: Bisection Method

Step 6: Feedback and refinement

- From the fitted parametric distribution, provide the expert with some summaries: for example, tail probabilities.
- See if these tail probabilities correspond closely to the expert's intuition!
- If not, perhaps ask the expert to refine their choices of ℓ or m or u, or perhaps all three!

Example 3.4: Bisection Method

Step 6: Feedback to the glaciologist, and possible refinement

- We show the glaciologist the plot of the $Ga(9, 0.36)$ density. Does this look OK? Yes!
- Now feedback some specific properties:
  - The 1%–ile and 99%–iles are about 10 feet and 48 feet, respectively. This means that

    $$\Pr(\theta < 10) = \Pr(\theta > 48) = 0.01, \text{or once in a hundred years.}$$

  - Does this seem OK?
  - The glaciologist thinks this is "imaginable"...
- No refinement needed here!

Example 3.5

Let Y be the retreat, in feet, of the *Zachariae Isstrøm* glacier. A *Pareto* distribution with rate θ is often used to model such geophysical activity, with probability density function

$$f(y|\kappa, \theta) = \theta\kappa^{\theta} y^{-(\theta+1)}, \qquad \theta, \kappa > 0 \text{and } y > \kappa.$$

(a) Obtain the likelihood function for θ given the parameter κ and some observed data $y_1, y_2, \ldots, y_n$ (independent).

(b) Suppose we observe a retreat of 20 feet at the *Zachariae Isstrøm* glacier in 2012. Write down the likelihood function for θ.

(c) Using the elicted prior for the rate of retreat we obtained from the expert glaciologist in Example 3.4, and assuming κ is known to be 12, obtain the posterior distribution $\pi(\theta|y_1 = 20)$.

Solution to Example 3.5(a) (1/1)

We have

$$f(\boldsymbol{y}|\theta,\kappa) = \theta\kappa^{\theta} y_1^{-(\theta+1)} \times \cdots \times \theta\kappa^{\theta} y_n^{-(\theta+1)}$$

$$= \theta^n \kappa^{n\theta} \prod_{i=1}^{n} y_i^{-(\theta+1)}. \tag{3.5}$$

Solution to Example 3.5(b) (1/1)

We simply substitute $n = 1$ and $y_1 = 20$ into Equation (3.5), giving

$$f(y_1 = 20|\theta, \kappa) = \theta\kappa^{\theta}20^{-(\theta+1)}.$$

Solution to Example 3.5(c) (1/3)

Using Bayes' Theorem, and following the examples in Chapter 2, we know that

$$\pi(\theta|y_1 = 20) \propto \pi(\theta) \times f(y_1 = 20|\theta, \kappa).$$

Recall from **Example 3.4** that our elicited prior for θ is $Ga(9, 0.36)$, which has density

$$\pi(\theta) = \frac{0.36^9 \theta^8 e^{-0.36\theta}}{\Gamma(9)}.$$

Solution to Example 3.5(c) (2/3)

Combining this with the likelihood above (and using $\kappa = 12$) gives

$$\pi(\theta|y_1 = 20) = \frac{0.36^9\theta^8 e^{-0.36\theta}}{\Gamma(9)} \times \theta 12^\theta 20^{-(\theta+1)}$$

$$\propto \theta^9 e^{-0.36\theta} 12^\theta 20^{-(\theta+1)}$$

$$\propto \theta^9 e^{-0.36\theta} 12^\theta 20^{-\theta}. \tag{3.6}$$

Solution to Example 3.5(c) (3/3)

Now consider the term $12^{\theta}20^{-\theta}$. Taking logs, we get

$$\theta \ln 12 - \theta \ln 20 = (\ln 12 - \ln 20)\theta;$$

exponentiating to 're–balance', you should see that

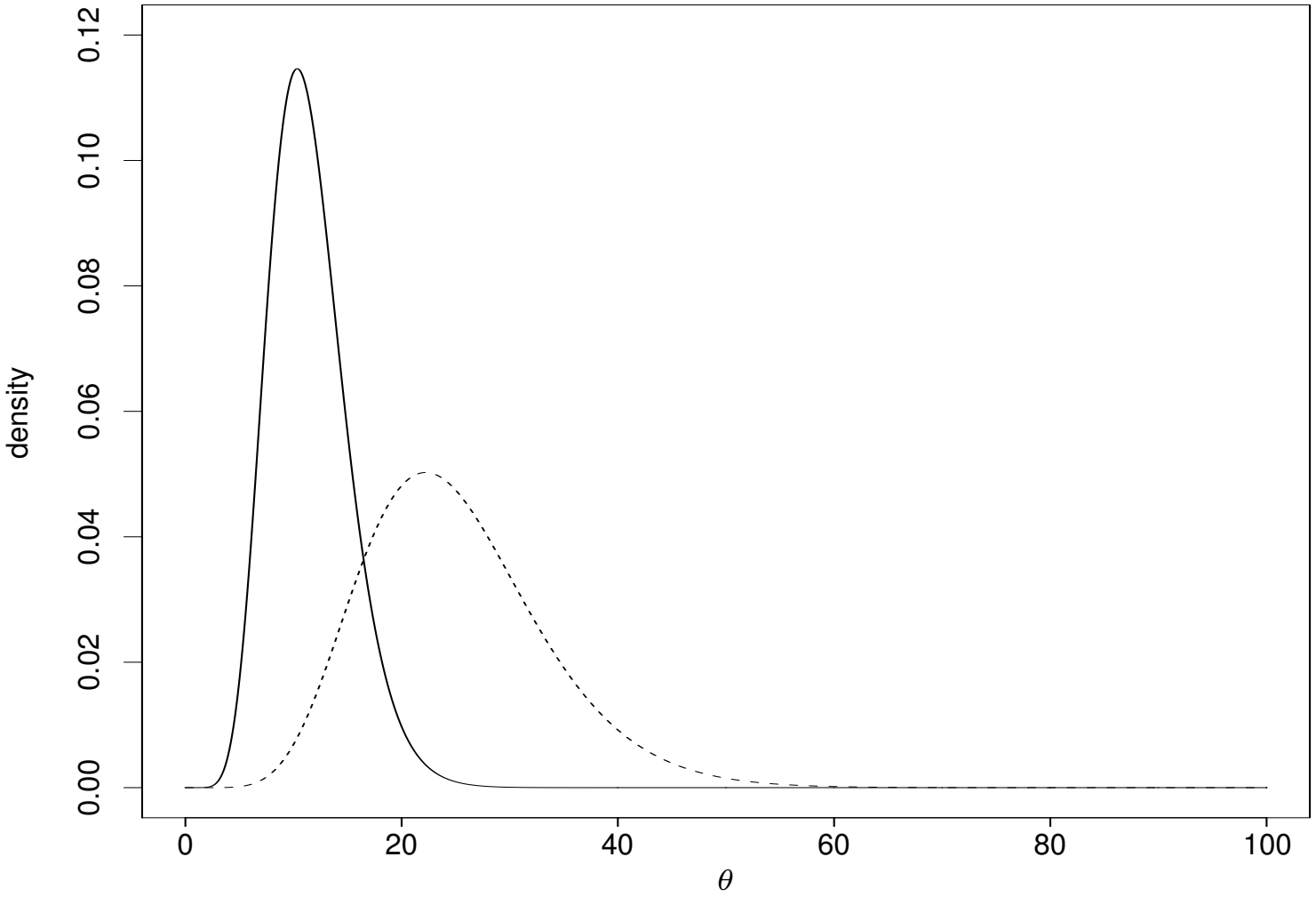
$$12^{\theta}20^{-\theta} = e^{(\ln 12 - \ln 20)\theta}.$$

Substituting back into (3.6) gives

$$\begin{aligned} \pi(\theta|y_1 = 20) &\propto \theta^9 e^{-0.36\theta} e^{(\ln 12 - \ln 20)\theta} \qquad \text{i.e.} \\ &\propto \theta^9 e^{-0.36\theta + (\ln 12 - \ln 20)\theta} \\ &\propto \theta^9 e^{-0.87\theta}, \end{aligned}$$

i.e. $\theta|y_1 = 20 \sim Ga(10, 0.87)$.

Example 3.5



Feedback on your feedback: The good

- **Lecture materials**
  - Booklet clearly structured
  - Good amount of examples
  - Like the amount of writing we have to do in notes
  - Good cross-referencing
  - Like the chapter summaries

Feedback on your feedback: The good

- **Lee**
  - Quite engaging (at times)
  - Jordy not an issue for me, Warm Jordy vocals, Soothing Jordy tones
  - Good pace
  - Occasionally enthusiastic
  - Oscar-worthy lectures
  - At last, you're free
- **Chris**
  - Prefer you to Chris anyway
  - Lee $>>>$ Chris

Feedback on your feedback: The bad

- **Lecture materials**
  - Remove the boxes in notes!
  - Mathematical font - difficult to tell difference between X and x
  - Do written stuff on visualiser, NOT SLIDES
  - Less Geography please
  - More time to write down from slides
  - Need more examples. Use NUMBAS perhaps?

    https://mas-shiny.ncl.ac.uk/2903Questions

  - Too many formulas. Need formula sheet

Feedback on your feedback: The bad

- **Lee**
  - Speak. More. S l o w l y
  - You over-explain the simple things
  - Always late, sort it out mate
  - Owe us 20 minutes from MAS2602
  - Go faster plz
- **Chris**
  - Where's Chris? Like him
  - Any more of Chris please?
  - Still love Chris, my climbing boi

Feedback on your feedback: The ugly

- Was really disappointed to learn that Lee was teaching us again
- Please shut up while we're copying down
- Engage us more and give us more breaks
- Less R code, it makes me angry that it's in here
- Bayesian sux frequentist 4 lyf
- You're much better at this, you were rubbish at R
- Too much irrelevant talking
- Not very engaging,... he just talks at us,...

Taking stock

- **Posterior $\propto$ Prior $\times$ Likelihood**
  - Always look for a gamma or a beta distribution first
  - Notation: $\pi(\theta)$ is just notation to represent our PDF for θ; we usually use $f(x)$ to represent the PDF of the data
  - Prior distribution should not have x's in it – the random variable is the parameter!
  - Notation: $E[\theta] =$ prior mean, $E[\theta|\boldsymbol{x}] =$ posterior mean
  - Notation: $f(x|\theta)$ is just our likelihood – form the product over the PDF for each observation, <u>if</u> you have multiple observations (careful with the Binomial!)
  - Intepretation: Compare prior/posterior means and variances; is the posterior closer to the data, or the prior?

Taking stock

- **Constructing priors**
  - For a gamma or a beta prior: two parameters, so two bits of information needed (e.g. mean/variance? mode/probability?)
  - Reference data: linear regression for the prior mean?
  - Historical records: Trial roulette method?
  - Bisection method (today)
- **Sufficiency**
  - Posterior using the likelihood for t is identical to that using the full dataset $\boldsymbol{x}$
  - Much more efficient for “Bayesian updating" – e.g. $T = \sum X_i$

Substantial prior information

Definition (Substantial prior information)

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

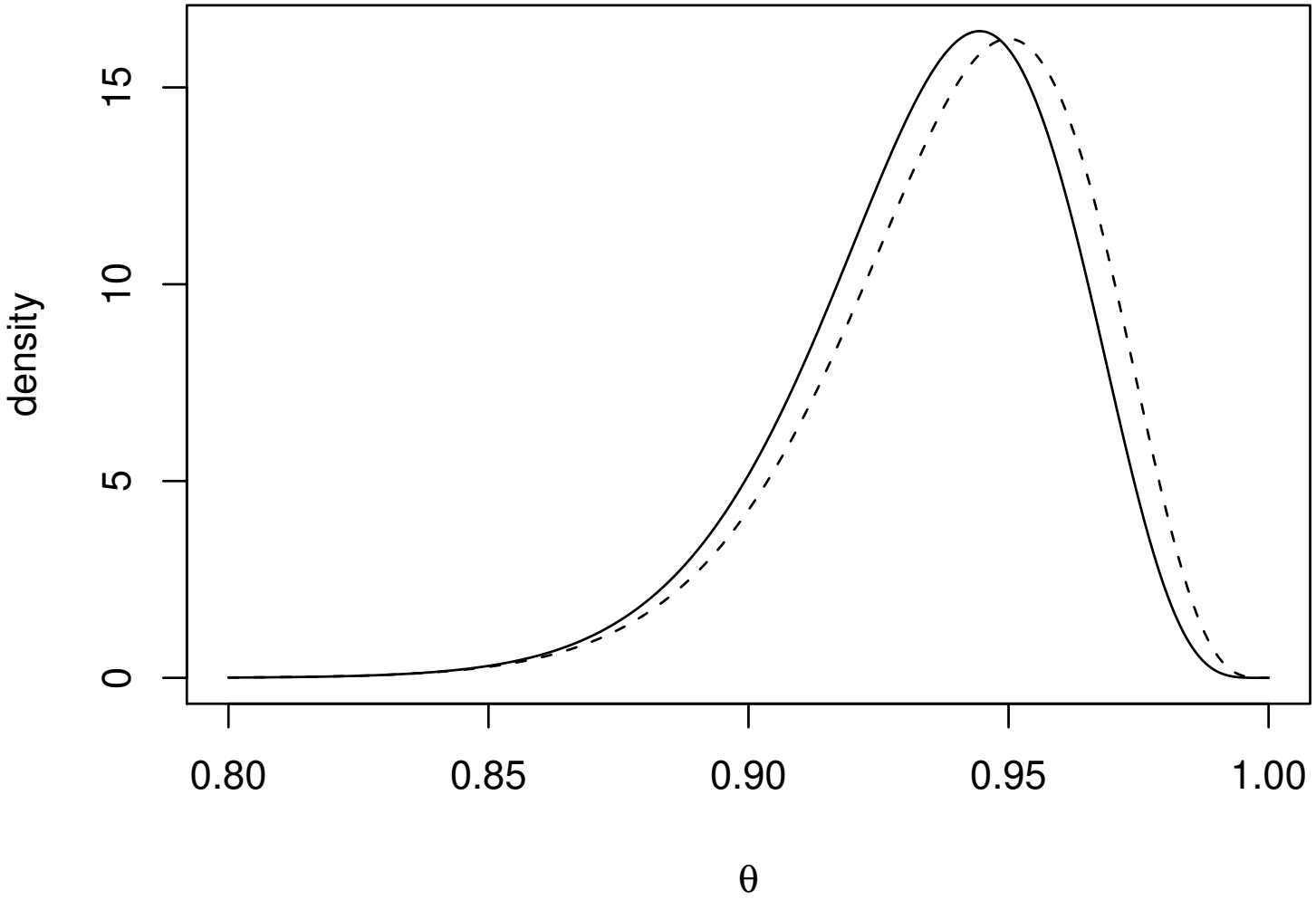
Substantial prior information

An example of substantial prior knowledge was given in **Example 2.2** where a music expert was trying to distinguish between pages from Mozart and Haydn scores.

Figure 3.9 shows the prior and posterior distributions for θ, the probability that the expert makes the correct choice.

Notice the similarity between the prior and posterior distributions. Observing the data has not altered our beliefs about θ very much.

Substantial prior information



Substantial prior information

When we have substantial prior information there can be some difficulties:

1. the **intractability of the mathematics** in deriving the posterior distribution — though with modern computing facilities this is less of a problem,
2. the **practical formulation of the prior distribution** — coherently specifying prior beliefs in the form of a probability distribution is far from straightforward although, as we have seen, this can be attempted using computer software.

Parameter Constraints

[We will come back to this soon... For now, turn to page 73!]

Vague Prior Knowledge/Prior Ignorance

If we have very little or no prior information about the model parameter θ, we must still choose a prior distribution in order to operate Bayes Theorem.

Obviously, it would be sensible to choose a prior distribution which is not concentrated about any particular value, that is, one with a very large variance.

In particular, 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 Knowledge/Prior Ignorance

An example of vague prior knowledge was given in **Example 2.1** where a possibly biased coin was assessed.

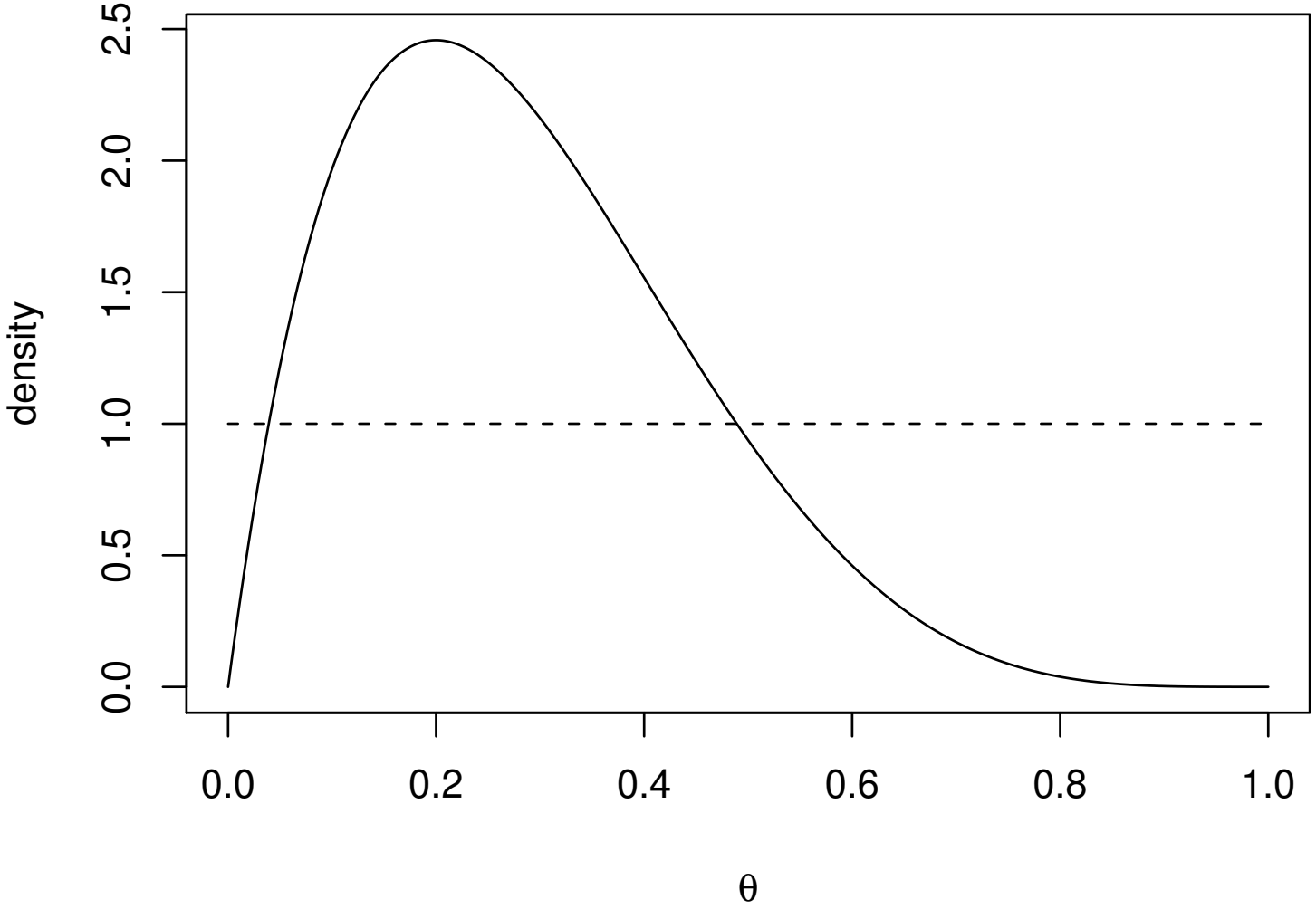
Figure 3.13 shows the prior and posterior distributions for $\theta = \Pr(\text{Head})$.

Notice that the prior and posterior distributions look very different.

In fact, in this example, the posterior distribution is simply a scaled version of the likelihood function – likelihood functions are not usually proper probability (density) functions and so scaling is required to ensure that it integrates to one.

Most of our beliefs about θ have come from observing the data.

Vague Prior Knowledge/Prior Ignorance



Vague prior knowledge

We represent **vague prior knowledge** by using a prior distribution which is conjugate to the model for $\boldsymbol{x}$ and which has "infinite" variance.

Example 3.9

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 to Example 3.9 (1/1)

Conjugate prior: Normal distribution. From **Example 2.6**, if $\mu \sim N(b, 1/d)$ then $\mu|\boldsymbol{x} \sim N(B, 1/D)$ where

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

If we now make our prior knowledge vague about μ by letting the prior variance tend to infinity ($d \to 0$), we obtain

$$B \to \bar{x} \qquad \text{and} \qquad D \to n\tau.$$

giving $\mu|\boldsymbol{x} \sim N(\bar{x}, 1/(n\tau))$ posterior distribution. Notice that the posterior mean is the sample mean (the likelihood mode) and that the posterior variance $1/D \to 0$ as $n \to \infty$.

Example 3.10

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 posterior distribution assuming a vague prior for θ.

Solution to Example 3.10 (1/1)

Conjugate prior: Gamma distribution. Recall that a $Ga(g, h)$ distribution has mean $m = g/h$ and variance $v = g/h^2$.

Rearranging these formulae we obtain

$$g = \frac{m^2}{v} \qquad \text{and} \qquad h = \frac{m}{v}.$$

Clearly $g \to 0$ and $h \to 0$ as $v \to \infty$ (for fixed m).

We have seen how taking a $Ga(g, h)$ prior distribution results in a $Ga(g + n, h + n\bar{x})$ posterior distribution (**Example 2.5**).

Therefore, taking a vague prior distribution will give a $Ga(n, n\bar{x})$ posterior distribution.

Note that the posterior mean is $1/\bar{x}$ (the likelihood mode) and that the posterior variance $1/(n\bar{x}^2) \to 0$ and $n \to \infty$.

Taking stock

Preface

- Setting the scene
- Real applications of Bayesian statistics
- Bayesian versus classical

Chapter 1

- Review/preparation
- Different ways of finding probabilities
- Conditional probability, LTT and Bayes Theorem
- Likelihood: A review and a look at sufficient statistics

Taking stock

Chapter 2

- **"Posterior is proportional to prior times likelihood"**
- Loads of examples, where the prior is given to us – some numerical, some algebraic
- The role of sufficiency in Bayesian statistics

Chapter 3

- Constructing priors: Substantial prior knowledge, vague prior knowledge and prior ignorance
- ~~Section 3.3: Parameter constraints~~
- Asymptotic posterior distribution (~~proof~~)

Taking stock

- **Posterior $\propto$ Prior $\times$ Likelihood**
  - Always look for a gamma or a beta distribution first
  - Notation: $\pi(\theta)$ is just notation to represent our PDF for θ; we usually use $f(x)$ to represent the PDF of the data
  - Prior distribution should not have x's in it – the random variable is the parameter!
  - Notation: $E[\theta] =$ prior mean, $E[\theta|\boldsymbol{x}] =$ posterior mean
  - Notation: $f(x|\theta)$ is just our likelihood – form the product over the PDF for each observation, <u>if</u> you have multiple observations (careful with the Binomial!)
  - Intepretation: Compare prior/posterior means and variances; is the posterior closer to the data, or the prior?

Reminders

- Last assignment: due **4pm, Thursday 29th April**
  - Questions 14, 30, 40 and 41 from Chapter 5
  - Would really help if you watched the Case Study video first
  - Can work in pairs again, if you want
- Learning outcomes for chaprters 1–3 now available in Canvas
- Mock exam will be available soon, with solutions
- Last written assignment: marks now in, will be released soon (with video feedback)

Prior ignorance

We could represent ignorance by the concept "all values of θ are equally likely".

If θ were **discrete** with m possible values then we could assign each value the same probability $1/m$.

However, if θ is **continuous**, we need some limiting argument (from the discrete case).

Prior ignorance

Suppose that θ can take values between a and b, where $-\infty < a < b < \infty$.

Letting all (permitted) values of θ be equally likely results in taking a uniform $U(a, b)$ distribution as our prior distribution for θ.

However, if the parameter space is not finite then we cannot do this: there is no such thing as a $U(-\infty, \infty)$ distribution.

Prior ignorance

Convention suggests that we should use the "**improper**" uniform prior distribution

$$\pi(\theta) = \text{constant}.$$

This distribution is improper because

$$\int_{-\infty}^{\infty} \pi(\theta)\, d\theta$$

is not a convergent integral, let alone equal to one.

Prior ignorance

We have a similar problem if θ takes positive values — we cannot use a $U(0, \infty)$ prior distribution.

Now if $\theta \in (0, \infty)$ then $\phi = \log \theta \in (-\infty, \infty)$, and so we could use an "improper" uniform prior for ϕ: $\pi(\phi) = \textit{constant}$.

In turn, this induces a distribution on θ. Recall the result from **Distribution Theory**:

Fact (Distribution of a transformation)

Suppose that X is a random variable with probability density function $f_X(x)$. If g is a bijective (1–1) function then the random variable $Y = g(X)$ has probability density function

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|. \tag{3.7}$$

Prior ignorance

Applying this result to $\theta = e^{\phi}$ gives

$$\begin{aligned}
\pi_\theta(\theta) &= \pi_\phi(\log\theta)\left|\frac{d}{d\theta}\log\theta\right|, \qquad \theta > 0 \\
&= \textit{constant} \times \left|\frac{1}{\theta}\right|, \qquad \theta > 0 \\
&\propto \frac{1}{\theta}, \qquad \theta > 0.
\end{aligned}$$

This too is an improper distribution.

Prior ignorance

There is a drawback of using uniform or improper priors to represent prior ignorance: if we are "ignorant" about θ then we are also "ignorant" about any function of θ, for example, about $\phi_1 = \theta^3$, $\phi_2 = e^\theta$, $\phi_3 = 1/\theta$, ...

Is it **possible** to choose a distribution where we are ignorant about all these functions of θ?

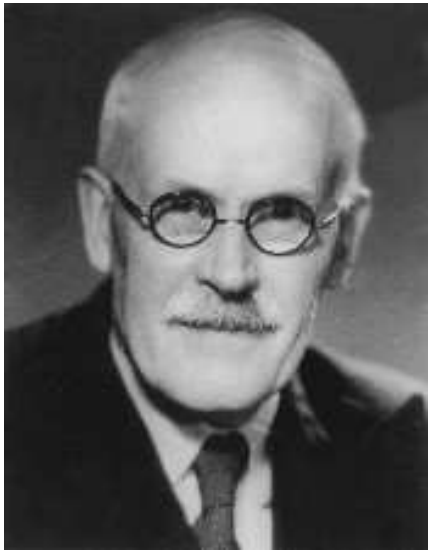
If not, on which function of θ should we place the uniform/improper prior distribution?

Prior ignorance

- There is no distribution which can represent ignorance for all functions of θ
- Assigning an ignorance prior to ϕ means we do not have an ignorance prior for e^{ϕ}

A solution to problems of this type was suggested by **Sir Harold Jeffreys**.

Sir Harold Jeffreys, FRS



- April 1891 – March 1989
- Mathematician, **Statistician**, Geophysicist, Astronomer
- Studied at **Armstrong College**, **Durham**, now Newcastle University (see plaque)
- Seminal book – **Theory of Probability** – revived the Bayesian view of probability
- Married a Physicist – Bertha Swirles – together they wrote **Methods of Mathematical Physics**

Jeffrey's prior

Jeffreys' suggestion was specified in terms of **Fisher's Information**

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right]. \tag{3.8}$$

He recommended that we represent prior ignorance by the prior distribution

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{3.9}$$

Such a prior distribution is known as a **Jeffreys prior distribution**.

Jeffrey's prior

Advantages of using a Jeffrey's prior

- satisfies the **local uniformity property**: it does not change much in the region over which the likelihood is significant $\Rightarrow$ represents ignornace
- It is **invariant** with respect to one–to–one transformations

Disadvantages

- Often **improper**, and can lead to improper posteriors
- Can be cumbersome to use in high dimensions

Example 3.11

Suppose we have a random sample from a distribution with probability density function

$$f(x|\theta) = \frac{2x\, e^{-x^2/\theta}}{\theta}, \qquad x > 0,\ \theta > 0.$$

Determine the Jeffreys prior for this model.

Solution to Example 3.11 (1/5)

The likelihood function is

$$
\begin{aligned}
f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{2x_i\, e^{-x_i^2/\theta}}{\theta} \\
&= \frac{2^n}{\theta^n} \left(\prod_{i=1}^{n} x_i \right) \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} x_i^2 \right\}.
\end{aligned}
$$

Solution to Example 3.11 (2/5)

Therefore

$$\begin{aligned}
\log f(\boldsymbol{x}|\theta) &= n\log 2 - n\log\theta + \sum_{i=1}^{n}\log x_i - \frac{1}{\theta}\sum_{i=1}^{n} x_i^2 \\
\Rightarrow \quad \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta) &= -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^2 \\
\Rightarrow \quad \frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{x}|\theta) &= \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i^2
\end{aligned}$$

Solution to Example 3.11 (3/5)

$$
\begin{aligned}
\Rightarrow \quad I(\theta) &= E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right] \\
&= -\frac{n}{\theta^2} + \frac{2}{\theta^3}E_{\boldsymbol{X}|\theta}\left[\sum_{i=1}^{n} X_i^2\right] \\
&= -\frac{n}{\theta^2} + \frac{2}{\theta^3}\left(E_{X|\theta}\left[X_1^2\right] + \ldots + E_{X|\theta}\left[X_n^2\right]\right) \\
&= -\frac{n}{\theta^2} + \frac{2n}{\theta^3}E_{X|\theta}\left[X^2\right]
\end{aligned}
$$

since the X_i are identically distributed.

Solution to Example 3.11 (4/5)

Now

$$E_{X|\theta}\left[X^2\right] = \int_0^\infty x^2 \times \left(\frac{2x\, e^{-x^2/\theta}}{\theta}\right) dx.$$

If we let $y = \dfrac{x^2}{\theta}$, then

$$\frac{dy}{dx} = \frac{2x}{\theta} \Rightarrow dx = \frac{\theta}{2x} dy.$$

Substituting into the integral above gives

$$\begin{aligned}\int_0^\infty y\, 2xe^{-y} \frac{\theta}{2x} dy &= \theta \int_0^\infty ye^{-y} dy \\ &= \theta \times 1 = \theta,\end{aligned}$$

since the remaining integral is the mean of a unit exponential.

Solution to Example 3.11 (5/5)

Therefore

$$I(\theta) = -\frac{n}{\theta^2} + \left(\frac{2n}{\theta^3} \times \theta\right) = \frac{n}{\theta^2}.$$

Hence, the Jeffreys prior for this model is

$$\begin{aligned} \pi(\theta) &\propto I(\theta)^{1/2} \\ &\propto \frac{\sqrt{n}}{\theta}, \qquad \theta > 0 \\ &\propto \frac{1}{\theta}, \qquad \theta > 0. \end{aligned}$$

Example 3.11

Notice that this distribution is **improper** since $\int_0^\infty d\theta/\theta$ is a divergent integral, and so we cannot find a constant which ensures that the density function integrates to one.

Example 3.12

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 to Example 3.12 (1/2)

Recall that

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \theta^n e^{-n\bar{x}\theta},$$

and therefore

$$\log f(\boldsymbol{x}|\theta) = n \log \theta - n\bar{x}\theta$$

$$\Rightarrow \quad \frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta) = \frac{n}{\theta} - n\bar{x}$$

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

$$\Rightarrow \quad I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{X}|\theta)\right] = \frac{n}{\theta^2}.$$

Solution to Example 3.12 (2/2)

Hence, the Jeffreys prior for this model is

$$\begin{aligned}\pi(\theta) &\propto I(\theta)^{1/2} \\ &\propto \frac{\sqrt{n}}{\theta}, \qquad \theta > 0 \\ &\propto \frac{1}{\theta}, \qquad \theta > 0.\end{aligned}$$

Example 3.12

Notice that this distribution is **improper** since $\int_0^\infty d\theta/\theta$ is a divergent integral, and so we cannot find a constant which ensures that the density function integrates to one.

Notice also that this density is, in fact, a limiting form of a $Ga(g, h)$ density (ignoring the integration constant) since

$$\frac{h^g\,\theta^{g-1}e^{-h\theta}}{\Gamma(g)} \propto \theta^{g-1}e^{-h\theta} \to \frac{1}{\theta}, \qquad \text{as } g \to 0,\ h \to 0.$$

Therefore, we obtain the same posterior distribution whether we adopt the Jeffreys prior or vague prior knowledge.

Example 3.13

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known).

Determine the Jeffreys prior for this model.

Solution to Example 3.13 (1/2)

Recall from Example 2.6 that

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

and therefore

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

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

Solution to Example 3.13 (2/2)

Also

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

$$\Rightarrow \quad I(\mu) = E_{\boldsymbol{X}|\mu}\left[-\frac{\partial^2}{\partial \mu^2} \log f(\boldsymbol{X}|\mu)\right] = n\tau.$$

Hence, the Jeffreys prior for this model is

$$\begin{aligned} \pi(\mu) &\propto I(\mu)^{1/2} \\ &\propto \sqrt{n\tau}, \qquad -\infty < \mu < \infty \\ &= \textit{constant}, \qquad -\infty < \mu < \infty. \end{aligned}$$

Example 3.13

Notice that this distribution is **improper** since $\int_{-\infty}^{\infty} d\mu$ is a divergent integral, and so we cannot find a constant which ensures that the density function integrates to one.

Also it is a limiting form of a $N(b, 1/d)$ density (ignoring the integration constant) since

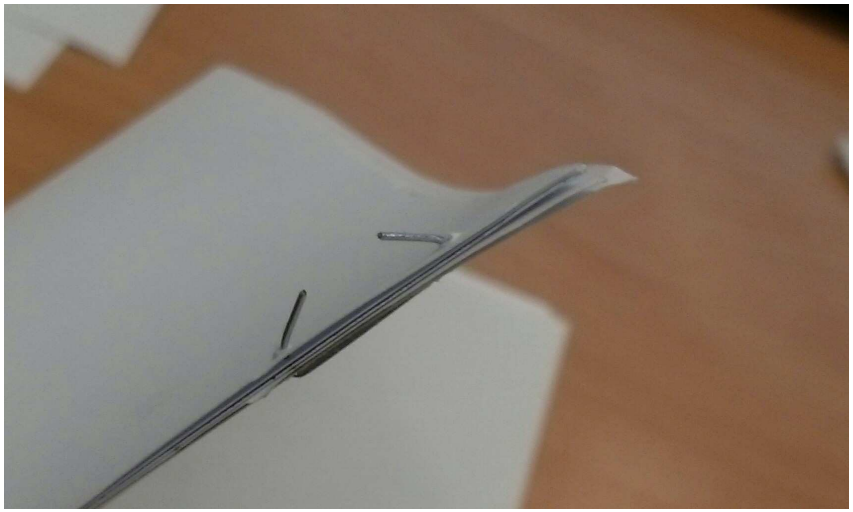
$$\left(\frac{d}{2\pi}\right)^{1/2} \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \propto \exp\left\{-\frac{d}{2}(\mu - b)^2\right\} \to 1,$$

as $d \to 0$.

Therefore, we obtain the same posterior distribution whether we adopt the Jeffreys prior or vague prior knowledge.

Dangerous stapling

Several assignments were poorly stapled



Staples like this are dangerous and can cause pages to go missing. Many of your assignments had to be re-stapled.
It is your responsibility to make sure your work is held together securely and safely, by pushing the stapler down firmly.
Assignments with unsafe staples will not be marked!

Asymptotic posterior distribution

There are many limiting results in Statistics.

The one you will probably remember is the **Central Limit Theorem**.

This concerns the distribution of $\bar{X}_n$, the mean of n independent and identically distributed random variables (each with known mean μ and known variance σ^2), as the sample size $n \to \infty$.

It is easy to show that $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$, and so if we define

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma},$$

then we know that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} N(0,1) \qquad \text{as } n \to \infty.$$

The following theorem gives a similar result for the posterior distribution.

Asymptotic posterior distribution

Theorem (Asymptotic posterior distribution)

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

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

Outline proof (1/7)

Using Bayes Theorem, the posterior distribution for θ is

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

Let $\psi = \sqrt{n}(\theta - \hat{\theta})$ and

$$\ell_n(\theta) = \frac{1}{n} \log f(\boldsymbol{x}|\theta)$$

be the average log–likelihood per observation, in which case,

$$f(\boldsymbol{x}|\theta) = e^{n\ell_n(\theta)}.$$

Outline proof (2/7)

Recall Equation (3.7), which tells us about the distribution of a random variable $Y = g(X)$:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}\,g^{-1}(y)\right|.$$

We want to know the distribution of $\psi = g(\theta)$, where

$$g(\theta) = \sqrt{n}(\theta - \hat{\theta}).$$

Now

$$g^{-1}(\psi) = \hat{\theta} + \frac{\psi}{\sqrt{n}} \qquad \text{and} \qquad \frac{d}{d\psi}g^{-1}(\psi) = \frac{1}{\sqrt{n}},$$

giving

$$\pi_\psi(\psi) = \pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\middle|\,\boldsymbol{x}\right) \times \frac{1}{\sqrt{n}}.$$

Outline proof (3/7)

Thus

$$\begin{aligned}\pi_\psi(\psi) &= \pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\middle|\boldsymbol{x}\right) \times \frac{1}{\sqrt{n}} \\ &\propto \pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right) \exp\left\{n\ell_n\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right)\right\}.\end{aligned}$$

Taking Taylor series expansions about $\psi = 0$ gives

$$\begin{aligned}\pi_\theta\left(\hat{\theta} + \frac{\psi}{\sqrt{n}}\right) &= \pi_\theta(\hat{\theta}) + \pi'_\theta(\hat{\theta})\frac{\psi}{\sqrt{n}} + \frac{1}{2!}\pi''_\theta(\hat{\theta})\left[\frac{\psi}{\sqrt{n}}\right]^2 \\ &\qquad + \frac{1}{3!}\pi'''_\theta(\hat{\theta})\left[\frac{\psi}{\sqrt{n}}\right]^3 + \ldots \\ &\approx \pi_\theta(\hat{\theta}).\end{aligned}$$

Outline proof (4/7)

Similarly,

$$
\begin{aligned}
n\ell_n\left(\hat{\theta}+\frac{\psi}{\sqrt{n}}\right) &= n\left\{\ell_n(\hat{\theta})+\ell_n'(\hat{\theta})\frac{\psi}{\sqrt{n}}+\frac{1}{2!}\ell_n''(\hat{\theta})\left[\frac{\psi}{\sqrt{n}}\right]^2\right. \\
&\qquad\qquad \left.+\frac{1}{3!}\ell_n'''(\hat{\theta})\left[\frac{\psi}{\sqrt{n}}\right]^3+\dots\right\} \\
&\approx n\ell_n(\hat{\theta})+n\ell_n'(\hat{\theta})\frac{\psi}{\sqrt{n}}+n\frac{1}{2!}\ell_n''(\hat{\theta})\left[\frac{\psi}{\sqrt{n}}\right]^2 \\
&\approx n\ell_n(\hat{\theta})+\frac{1}{2}\ell_n''(\hat{\theta})\psi^2.
\end{aligned}
$$

Outline proof (5/7)

This gives

$$\begin{aligned}
\pi_\psi(\psi|\boldsymbol{x}) &\propto \pi_\theta\left(\hat{\theta}+\frac{\psi}{\sqrt{n}}\right)\exp\left\{n\ell_n\left(\hat{\theta}+\frac{\psi}{\sqrt{n}}\right)\right\} \\
&\approx \pi_\theta(\hat{\theta})\exp\left\{n\ell_n(\hat{\theta})+\frac{1}{2}\ell_n''(\hat{\theta})\psi^2\right\} \\
&\propto \exp\left\{\frac{1}{2}\ell_n''(\hat{\theta})\psi^2\right\} \\
&\propto \exp\left\{-\frac{\left[-\ell_n''(\hat{\theta})\right]}{2}(\psi-0)^2\right\},
\end{aligned}$$

showing that

$$\psi|\boldsymbol{x} \sim N\left(0, [-\ell_n''(\hat{\theta})]^{-1}\right) \qquad \text{as } n \to \infty.$$

Outline proof (6/7)

The result on the last slide gives

$$Var\left\{\sqrt{n}(\theta-\hat{\theta})\right\} = \left[-\ell_n''(\hat{\theta})\right]^{-1}.$$

Multiplying the term inside $\{\}$ by $\sqrt{-\ell_n''(\hat{\theta})}$ gives

$$\begin{aligned}
Var\left\{\sqrt{-\ell_n''(\hat{\theta})}\times\sqrt{n}(\theta-\hat{\theta})\right\} &= -\ell_n''(\hat{\theta})\times Var\left\{\sqrt{n}(\theta-\hat{\theta})\right\}, \text{ i.e.} \\
Var\left\{\sqrt{-n\ell_n''(\hat{\theta})}(\theta-\hat{\theta})\right\} &= -\ell_n''(\hat{\theta})\times\left[-\ell_n''(\hat{\theta})\right]^{-1}, \qquad \text{i.e.} \\
Var\left\{\sqrt{J(\hat{\theta})}(\theta-\hat{\theta})\right\} &= 1.
\end{aligned}$$

Outline proof (7/7)

Thus, we have the equivalent result

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

Dividing by $\sqrt{J(\hat{\theta})}$ and adding $\hat{\theta}$ also gives

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

MEMORISE!

Example 3.14

Suppose we have a random sample from a distribution with probability density function

$$f(x|\theta) = \frac{2x\, e^{-x^2/\theta}}{\theta}, \qquad x > 0,\ \theta > 0.$$

Determine the asymptotic posterior distribution for θ. Note that from Example 3.11 we have

$$\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^2,$$

$$J(\theta) = -\frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{x}|\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3}\sum_{i=1}^{n} x_i^2 = \frac{n}{\theta^3}\left(-\theta + \frac{2}{n}\sum_{i=1}^{n} x_i^2\right).$$

Solution to Example 3.14 (1/2)

The asymptotic posterior distribution is given by

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

First, let's find $\hat{\theta}$. Now

$$\frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^2;$$

Setting equal to zero and solving for $\theta = \hat{\theta}$ gives

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \overline{x^2}.$$

Solution to Example 3.14 (2/2)

Also,

$$\begin{aligned} J(\hat{\theta}) &= \frac{n}{\hat{\theta}^3}\left(-\hat{\theta} + \frac{2}{n}\sum_{i=1}^{n} x_i^2\right) \\ &= \frac{n}{\left(\overline{x^2}\right)^3}\left(-\overline{x^2} + 2\overline{x^2}\right) \\ &= \frac{n}{\left(\overline{x^2}\right)^3}\overline{x^2} = \frac{n}{\left(\overline{x^2}\right)^2}. \end{aligned}$$

Therefore, for large n, the (approximate) posterior distribution for θ is

$$\theta|\boldsymbol{x} \sim N\left(\overline{x^2}, \frac{1}{n}\left(\overline{x^2}\right)^2\right).$$

Example 3.15

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 asymptotic posterior distribution for θ.

Note that from Example 3.12 we have

$$\frac{\partial}{\partial\theta} \log f(\boldsymbol{x}|\theta) = \frac{n}{\theta} - n\bar{x},$$

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

Solution to Example 3.15 (1/1)

We have

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta) = 0 \qquad &\Longrightarrow \qquad \hat{\theta} = \frac{1}{\bar{x}} \\
&\Longrightarrow \qquad J(\hat{\theta}) = \frac{n}{\left(\frac{1}{\bar{x}}\right)^2} = n\bar{x}^2 \\
&\Longrightarrow \qquad J(\hat{\theta})^{-1} = \frac{1}{n\bar{x}^2}.
\end{aligned}$$

Therefore, for large n, the (approximate) posterior distribution for θ is

$$\theta|\boldsymbol{x} \sim N\left(\frac{1}{\bar{x}}, \frac{1}{n\bar{x}^2}\right).$$

Example 3.15

Recall that, assuming a vague prior distribution, the posterior distribution is a $Ga(n, n\bar{x})$ distribution, with mean $1/\bar{x}$ and variance $1/(n\bar{x}^2)$.

The Central Limit Theorem tells us that, for large n, the gamma distribution tends to a normal distribution, matched, of course, for mean and variance.

Therefore, we have shown that, for large n, the asymptotic posterior distribution is the same as the posterior distribution under vague prior knowledge. Not a surprising result!

Example 3.16

Suppose we have a random sample from a $N(\mu, 1/\tau)$ distribution (with τ known). Determine the asymptotic posterior distribution for μ. Note that from Example 3.13 we have

$$\frac{\partial}{\partial \mu} \log f(\boldsymbol{x}|\mu) = n\tau(\bar{x} - \mu),$$

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

Solution to Example 3.16 (1/1)

We have

$$\begin{aligned}
\frac{\partial}{\partial \mu} \log f(\boldsymbol{x}|\mu) = 0 \qquad &\Longrightarrow \qquad \hat{\mu} = \bar{x} \\
&\Longrightarrow \qquad J(\hat{\mu}) = n\tau \\
&\Longrightarrow \qquad J(\hat{\mu})^{-1} = \frac{1}{n\tau}.
\end{aligned}$$

Therefore, for large n, the (approximate) posterior distribution for μ is

$$\mu|\boldsymbol{x} \sim N\left(\bar{x}, \frac{1}{n\tau}\right).$$

Example 3.16

Again, we have shown that the asymptotic posterior distribution is the same as the posterior distribution under vague prior knowledge.

Chapter 5: Question 29

Using a random sample from a $Bin(k, \theta)$ (with k known), determine the posterior distribution for θ assuming

(i) vague prior knowledge;

(ii) the Jeffreys prior distribution;

(iii) a very large sample.

Solution

The **conjugate prior distribution** is a $Beta(g, h)$ distribution. Using this prior distribution, the posterior density is

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\
&\propto \theta^{g-1}(1-\theta)^{h-1} \times \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{k-x_i}, \qquad 0 < \theta < 1 \\
&\propto \theta^{g+n\bar{x}-1}(1-\theta)^{h+nk-n\bar{x}-1}, \qquad 0 < \theta < 1
\end{aligned}$$

$$i.e.\ \theta|\boldsymbol{x} \sim Beta(G = g + n\bar{x}, H = h + nk - n\bar{x}).$$

Solution

We represent **vague prior information** by taking a conjugate prior distribution with large variance.

As the beta distribution restricts values to the range (0,1), there is a finite upper limit to the variance.

Intuitively, the maximum variance is achieved when the probability density is pushed to the extremes of the range, that is, equal mass at $\theta = 0$ and $\theta = 1$ – this distribution is obtained in the limit $g \to 0$ and $h \to 0$.

Thus we will take this limit to represent vague prior information.Hence the posterior distribution under vague prior information is

$$\theta | \boldsymbol{x} \sim Beta(n\bar{x}, nk - n\bar{x}).$$

Solution

The **Jeffreys prior distribution** is

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

Now

$$\begin{aligned} f(\boldsymbol{x}|\theta) &\propto \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{k-x_i} \\ &\propto \theta^{n\bar{x}}(1-\theta)^{kn-n\bar{x}}. \end{aligned}$$

Solution

Therefore

$$
\begin{aligned}
\log f(\boldsymbol{x}|\theta) &= \mathit{constant} + n\bar{x}\log\theta + n(k-\bar{x})\log(1-\theta) \\
&\Rightarrow \quad \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta) = \frac{n\bar{x}}{\theta} - \frac{n(k-\bar{x})}{1-\theta} \\
&\Rightarrow \quad \frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{x}|\theta) = -\frac{n\bar{x}}{\theta^2} - \frac{n(k-\bar{x})}{(1-\theta)^2} \\
&\Rightarrow \quad I(\theta) = E_{\boldsymbol{X}|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right] \\
&\qquad = \frac{nE_{\boldsymbol{X}|\theta}(\bar{X})}{\theta^2} + \frac{n[k - E_{\boldsymbol{X}|\theta}(\bar{X})]}{(1-\theta)^2}.
\end{aligned}
$$

Solution

Now $E_{\boldsymbol{X}|\theta}(\bar{X}) = E_{X|\theta}(X) = k\theta$. Therefore

$$I(\theta) = \frac{nk}{\theta} + \frac{nk}{1-\theta} = \frac{nk}{\theta(1-\theta)}.$$

Hence the Jeffreys prior for this model is

$$\begin{aligned}
\pi(\theta) &\propto \sqrt{I(\theta)} \\
&\propto \sqrt{\frac{nk}{\theta(1-\theta)}}, \qquad 0 < \theta < 1 \\
&\propto \theta^{-1/2}(1-\theta)^{-1/2}, \qquad 0 < \theta < 1.
\end{aligned}$$

Solution

This is a $Beta(1/2, 1/2)$ prior distribution and so the resulting posterior distribution is

$$\theta|\boldsymbol{x} \sim Beta(1/2 + n\bar{x}, 1/2 + nk - n\bar{x}).$$

Solution

The **asymptotic posterior distribution** (as $n \to \infty$) is

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

where

$$J(\theta) = -\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{x}|\theta) = \frac{n\bar{x}}{\theta^2} + \frac{n(k-\bar{x})}{(1-\theta)^2}.$$

Solution

Now

$$\frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta) = 0 \qquad \Longrightarrow \qquad \frac{n\bar{x}}{\hat{\theta}} - \frac{n(k-\bar{x})}{1-\hat{\theta}} = 0$$

$$\Longrightarrow \qquad \hat{\theta} = \frac{\bar{x}}{k}$$

$$\Longrightarrow \qquad J(\hat{\theta}) = \frac{nk^3}{\bar{x}(k-\bar{x})}$$

$$\Longrightarrow \qquad J(\hat{\theta})^{-1} = \frac{\bar{x}(k-\bar{x})}{nk^3}.$$

Solution

Therefore, for large n, the posterior distribution for θ is

$$\theta|\boldsymbol{x} \sim N\left(\frac{\bar{x}}{k}, \frac{\bar{x}(k-\bar{x})}{nk^3}\right) \qquad \text{approximately.}$$

Welcome back!

Welcome back!

- Whatever that means...!
- I hope you're all safe and well
- We're certainly living in strange times. Impact upon this module:
  - Mid-semester test was cancelled (16th March)
  - Both assignments still stand – assignment 2 due in this coming Friday (3pm) via the Assessment link in BB
  - Exam: A weird "take home" exam that you will have 24 hours to complete
  - Not sure of the new weightings of the assignments and exam just yet
- Over the next two weeks I intend to finish the course – complete Chapter 4
- That will leave us with one week of revision, including going through a specimen paper in the style of your take-home exam

Welcome back!

At first glance, your MAS2905 exam will look just like a typical MAS2905 exam paper. However:

- It will be open-book
- You will have 24 hours to complete the paper, and you will upload your solutions to BB
- There will be hardly any ‘bookwork’ questions - e.g. no proofs or definitions directly out of the notes
- Expect more questions that require you to ‘interpret’

I’ve no idea how these new ReCap lectures will go:

- Can’t do doodles on the visualiser/whiteboard
- Will likely go too fast – just pause when you need to!
- Looking after my little boy – he will likely join us

Chapter 4

Bayesian Inference

Estimation: Point estimates

There are many useful summaries for a typical value of a random variable with a particular distribution.

For example,

- the mean;
- the mode;
- the median.

The **mode** is used more often as a summary than is the case in frequentist statistics.

Estimation: Interval estimates

A more useful summary of the posterior distribution is one which also reflects its **variation**.

For example, 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 a continuous quantity with posterior probability density function $\pi(\theta|\boldsymbol{x})$ then

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

Estimation: Interval estimates

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.

Mozart, Earthquakes, speed cameras, glaciers

Think back to the examples in Chapters 2 and 3: the music expert, the geologist, the hydrologist, the glaciologist...

In each example, we used **Bayes' Theorem for distributions** to combine the beliefs of the expert, with our data, to form the **posterior distribution**.

How should we feedback our findings to the experts?

Mozart, Earthquakes, speed cameras, glaciers

We often used a single posterior summary – usually the mean (often in the form of the **Bayes Linear rule**), or the mode.

But we really need more than this!

We *could* supply the expert with the full posterior.

Often, we use **Bayesian confidence intervals**.

Example: Flood defence in New Orleans

We will think about this in more detail in **Case Study 2**.

Hurricane Katrina destroyed centuries–old sea walls (levees) in New Orleans in August 2005.

Example: Flood defence in New Orleans

A Bayesian analysis of sea–surge in the Gulf of Mexico could be used to help design the new sea walls. This would:

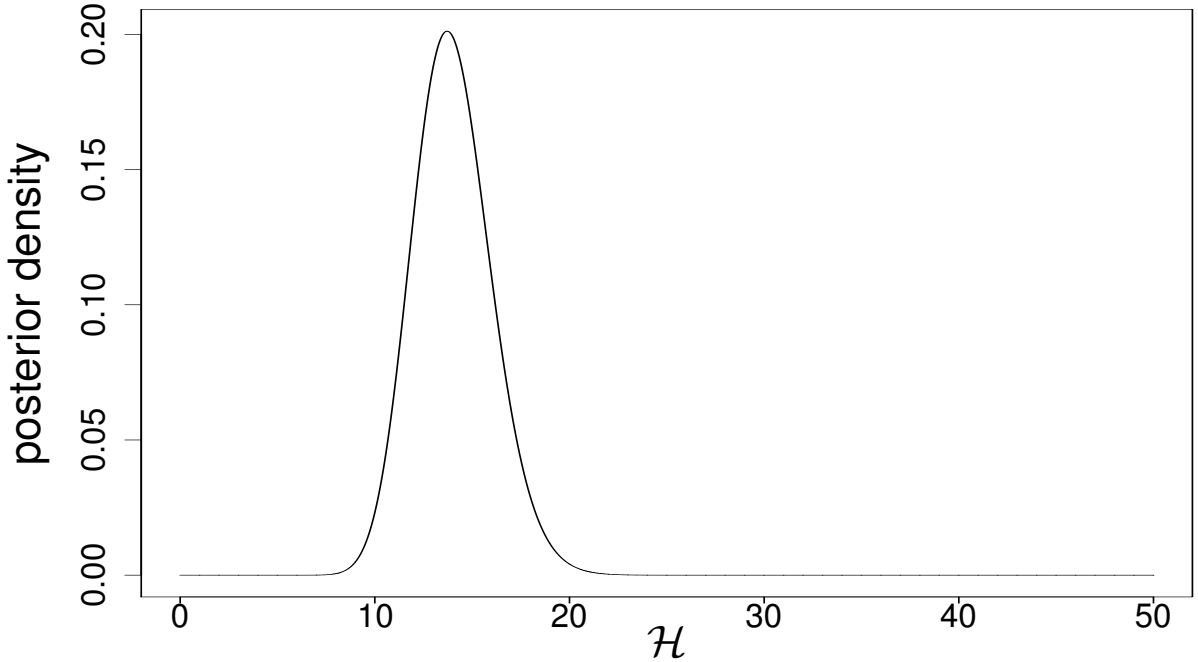
- Use data on the height of the sea (notably sea–surge)
- Supplement the analysis with the knowledge of an expert – perhaps an **oceanographer** – through a suitable prior distribution for the parameter of interest

Example: Flood defence in New Orleans

A recent Bayesian analysis by researchers at Tulane University, New Orleans (**JRSS, 2012**), found that

$$\mathcal{H}|\boldsymbol{x} \sim Ga(49, 3.5),$$

where $\mathcal{H}$ is the height (in feet) of the new sea wall, estimated to protect New Orleans against the **once in a hundred year** storm surge:



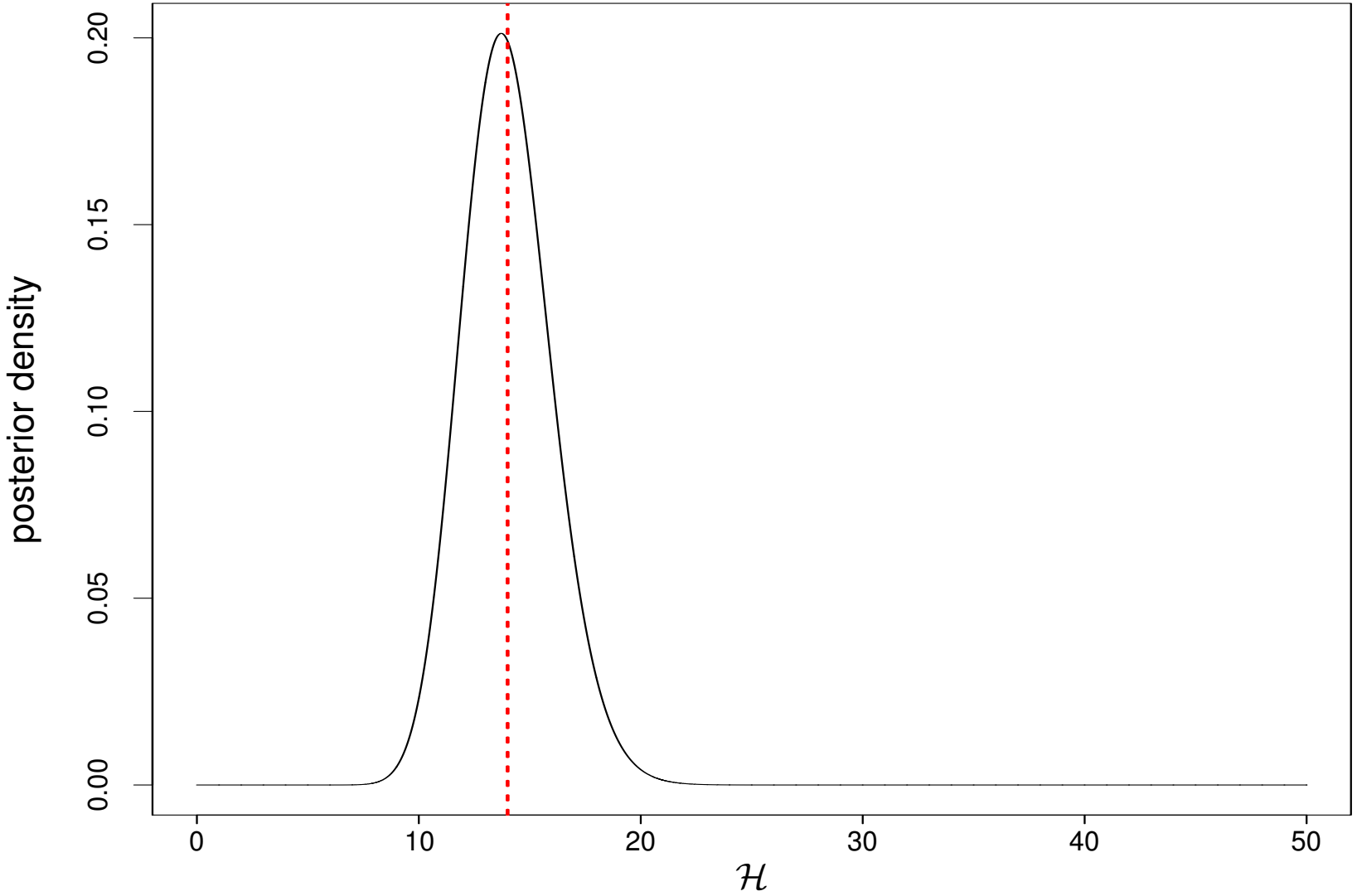
Example: Flood defence in New Orleans

What do we tell the engineers?

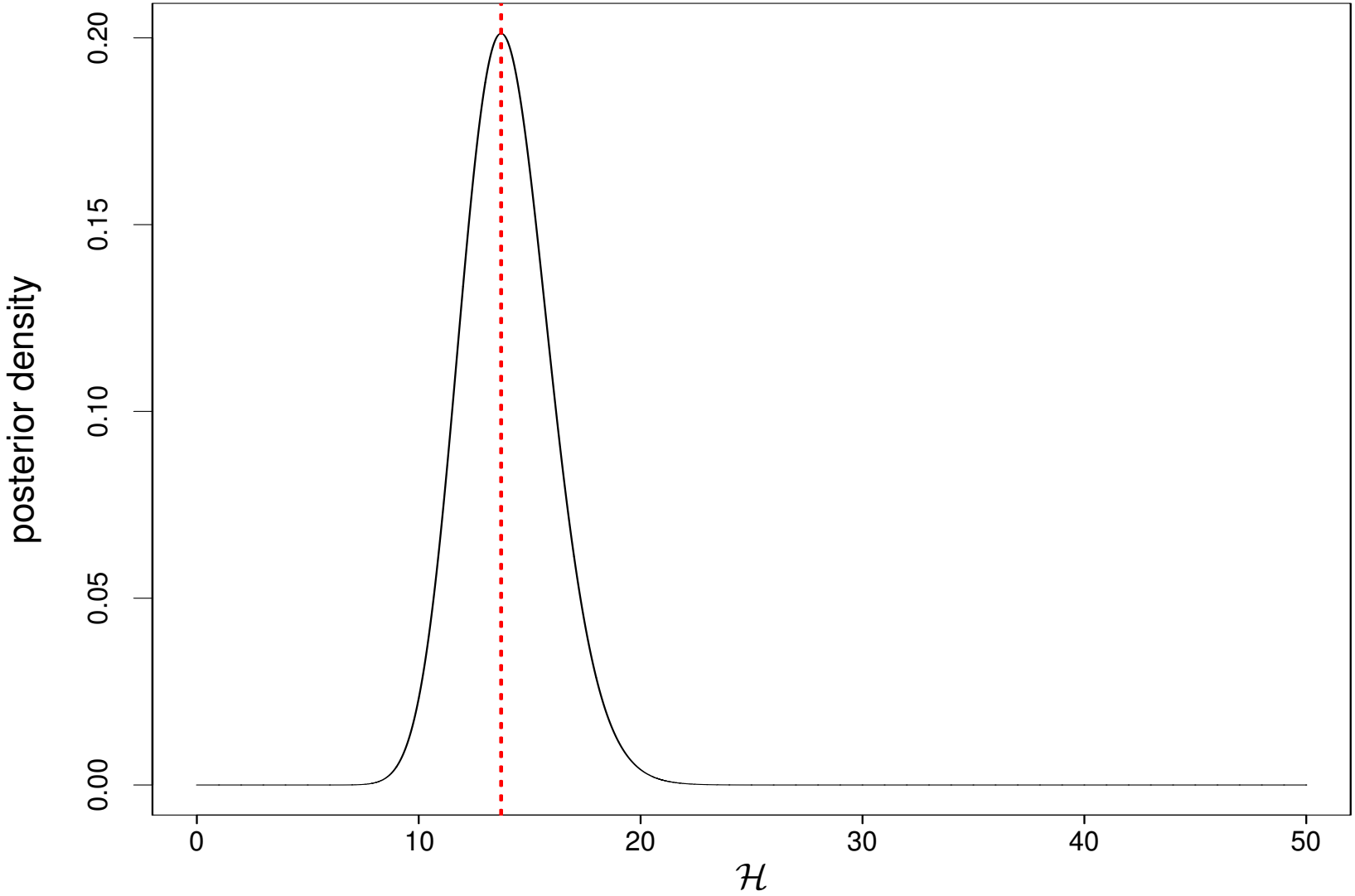
The gamma distribution for $\mathcal{H}|\boldsymbol{x}$ covers the entire positive real line...

- "Your sea wall needs to be be anywhere between 0 feet and ∞ feet tall"?
- "Your sea wall needs to be
$$E(\mathcal{H}|\boldsymbol{x}) = \frac{49}{3.5} = 14 \text{ feet tall"?}$$
- "Your sea wall needs to be
$$Mode(\mathcal{H}|\boldsymbol{x}) = \frac{49-1}{3.5} = 13.7 \text{ feet tall"?}$$
- $SD(\mathcal{H}|\boldsymbol{x}) = \sqrt{\frac{49}{3.5^2}} = 2$ feet

Example: Flood defence in New Orleans



Example: Flood defence in New Orleans



Estimation: Interval estimates

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 $\Pr(\theta \in C_\alpha|\boldsymbol{x}) = 1-\alpha$.

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.

For example, the HDI could take the form $C_\alpha = (a,b) \cup (c,d) \cup (e,f)$, where $a < b < c < d < e < f$.

Interpretation of confidence intervals

Suppose C_B is a 95% **Bayesian** confidence interval for θ and C_F is a 95% **frequentist** confidence interval for θ.

These intervals do not have the same interpretation:

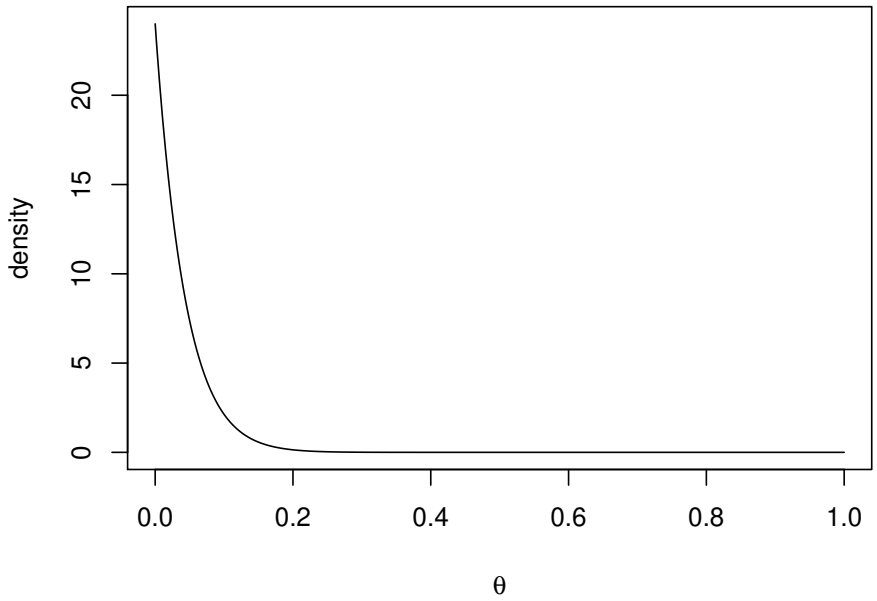
- the probability that C_B contains θ is 0.95;
- the probability that C_F contains θ is either 0 or 1 — since θ does not have a (non-degenerate) probability distribution;
- the interval C_F covers the true value θ on 95% of occasions — in repeated applications of the formula.

Example 4.1

Suppose that the posterior distribution for θ is a *Beta*$(1, 24)$ distribution, with probability density function

$$\pi(\theta|\boldsymbol{x}) = 24\,(1-\theta)^{23}, \quad 0 < \theta < 1.$$

A plot of this distribution is given in Figure 4.1 (See also the gravestones question in Chapter 5):



Determine the $100(1-\alpha)\%$ HDI for θ.

Solution to Example 4.1 (1/2)

The HDI must include those values of θ with highest posterior density and so must take the form $C_\alpha = (0, b)$.

The end–point b must satisfy

$$\int_0^b 24\,(1-\theta)^{23}\,d\theta = 1 - \alpha.$$

Now

$$\int_0^b 24\,(1-\theta)^{23}\,d\theta = \Big[-(1-\theta)^{24}\Big]_0^b = 1 - (1-b)^{24}.$$

Solution to Example 4.1 (2/2)

Hence

$$1-(1-b)^{24} = 1-\alpha \quad \Longrightarrow \quad 1-b = \alpha^{1/24} \quad \Longrightarrow \quad b = 1-\alpha^{1/24}.$$

Therefore, a $100(1-\alpha)\%$ HDI for θ is $(0, 1-\alpha^{1/24})$.

For example, the 95% HDI for θ is

$$(0, 1-(0.05^{1/24})) \longrightarrow (0, 0.117).$$

Example 4.2

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 to Example 4.2 (1/1)

This distribution has a symmetric bell shape and so the HDI takes the form $C_\alpha = (a, b)$ with end–points

$$a = \bar{x} - \frac{z_{\alpha/2}}{\sqrt{n\tau}} \qquad \text{and} \qquad b = \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n\tau}},$$

where z_α is the upper α-quantile of the $N(0, 1)$ distribution.

Therefore, the 95% HDI for μ is

$$\left(\bar{x} - \frac{1.96}{\sqrt{n\tau}}, \bar{x} + \frac{1.96}{\sqrt{n\tau}}\right).$$

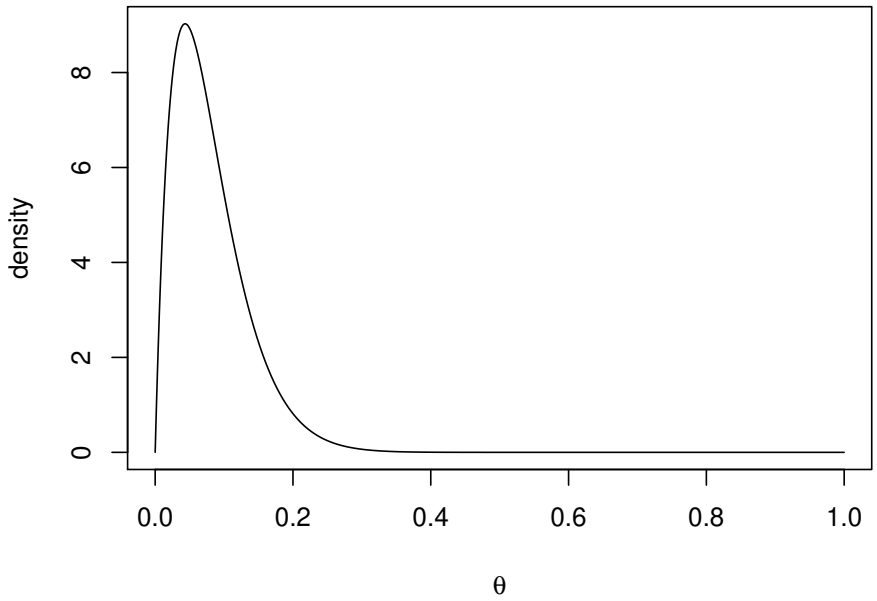
Note that this interval is numerically identical to the 95% frequentist confidence interval for the (population) mean of a normal random sample with known variance. However, the interpretation is very different.

Example 4.3

Suppose that the posterior distribution for θ is a *Beta*$(2, 23)$ distribution, with probability density function

$$\pi(\theta|\boldsymbol{x}) = 552\,\theta(1-\theta)^{22}, \quad 0 < \theta < 1.$$

A plot of this distribution is given in Figure 4.2.



Determine the $100(1-\alpha)\%$ HDI for θ.

Solution to Example 4.3

The HDI must take the form $C_\alpha = (a, b)$ if it is to include the values of θ with the highest probability density.

Therefore, the end–points a and b must satisfy

$$\int_a^b 552\,\theta(1-\theta)^{22}\,d\theta = 1-\alpha \qquad \text{and} \qquad 552a(1-a)^{22} = 552b(1-b)^{22}$$

Unfortunately, there is no simple analytical solution to these equations and so numerical methods have to be employed to determine a and b.

Computation of HDIs for unimodal distributions

Suppose that we require the HDI (a, b) for a unimodal distribution with density $f(\cdot)$ and distribution function $F(\cdot)$.

We have seen that if one of the end–points is known (because of the shape of the distribution) or the distribution is symmetric then the solution is in terms of the distribution's percentage points.

Computation of HDIs for unimodal distributions

When this is not the case, the problem requires a numerical scheme to find a and b satisfying

$$F(b) - F(a) = 1 - \alpha \qquad \text{and} \qquad f(a) = f(b).$$

Thus,

$$F(b) - F(a) - (1 - \alpha) = 0 \qquad \text{and} \qquad f(b) - f(a) = 0.$$

Any numerical solver, such as `optim` in R, will search through different a and b until the above conditions are met.

We can help the computer by squaring both (ensuring both differences will always be positive); we can also combine both squares:

$$\{F(b) - F(a) - (1 - \alpha)\}^2 + \{f(b) - f(a)\}^2 ,$$

and then ask R to find a and b such that this sum is minimsed.

Computation of HDIs for unimodal distributions

In fact, in order to use `optim` in R we also need to introduce a 'tuning parameter' k, and hence minimise:

$$g(a, b) = \left\{F(b) - F(a) - (1 - \alpha)\right\}^2 + k\left\{f(b) - f(a)\right\}^2,$$

where $k > 0$ tries to ensure that both terms are zeroed.

Computation of HDIs for unimodal distributions

In R:

```
g=function(x)
     {
      a=x[1]
      b=x[2]
       (pbeta(b,2,23)-pbeta(a,2,23)-0.95)**2
       +0.0001*(dbeta(b,2,23)-dbeta(a,2,23))**2
     }
```

Then:

```
initiala=qbeta(0.025,2,23)
initialb=qbeta(0.975,2,23)
res=optim(c(initiala,initialb),g,
     method="L-BFGS-B",lower=0,upper=1)
a=res$par[1]
b=res$par[2]
```

Computation of HDIs for unimodal distributions

This gives

- $a = 0.002211733$
- $b = 0.1840109$
- $F(b) - F(a) = 0.9500041$
- $f(b) - f(a) = -0.004484121$.

Thus the 95% HDI is $(0.002211733, 0.1840109)$.

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.

Prediction

Clearly there will be uncertainty about the future outcome of an experiment.

Suppose this future outcome Y is described by a probability (density) function $f(y|\theta)$.

There are several ways we could make inferences about what values of Y are likely.

Prediction

For example, if we have an estimate $\hat{\theta}$ of θ we might base our inferences on $f(y|\theta = \hat{\theta})$ (i.e. a **frequentist** approach).

Obviously this is not the best we can do, as such inferences ignore the fact that it is very unlikely that $\theta = \hat{\theta}$.

Prediction

Implicit in the Bayesian framework is the concept of the **predictive distribution**.

This distribution describes how likely are different outcomes of a future experiment.

The predictive probability (density) function is calculated as

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

when θ is a continuous quantity.

From this equation, we can see that the predictive distribution is formed by weighting the possible values of θ in the future experiment $f(y|\theta)$ by how likely we believe they are to occur $\pi(\theta|\boldsymbol{x})$.

Prediction

If the true value of θ were known, say θ_0, then any prediction can do no better than one based on $f(y|\theta = \theta_0)$.

However, as (generally) θ is unknown, the predictive distribution is used as the next best alternative.

Prediction

We can use the predictive distribution to provide a useful range of plausible values for the outcome of a future experiment.

This **prediction interval** is 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 4.4

Suppose that X is the number of expensive goods sold in a shop over 24 days.

If θ is the expected number of sales per day then it may be plausible that $X|\theta \sim Po(24\,\theta)$.

Also, suppose our prior distribution for θ is as given in Table 4.1.

	θ	$\pi(\theta)$
"great"	1/2	0.2
"good"	1/4	0.5
"poor"	1/8	0.3

Example 4.4

Clearly, we believe that the most likely value of θ is 1/4, indicating that we would expect around 6 expensive goods to be sold in any 24 day period.

Suppose now that we observe that $x = 10$ expensive goods were sold in the last 24 days.

This will impact our beliefs about θ.

We can calculate the posterior distribution for θ as follows.

Example 4.4

The likelihood term is

$$\Pr(X = 10|\theta) = \frac{(24\theta)^{10}e^{-24\theta}}{10!} = \begin{cases} 0.1048 & \text{if } \theta = 1/2 \\ 0.0413 & \text{if } \theta = 1/4 \\ 0.0008 & \text{if } \theta = 1/8 \end{cases}$$

and so, using Bayes' Theorem,

$$\begin{aligned}\pi(\theta = 1/2|x = 10) = \frac{\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2)}{[\Pr(X = 10|\theta = 1/2)\pi(\theta = 1/2)} \\ + \Pr(X = 10|\theta = 1/4)\pi(\theta = 1/4) \\ + \Pr(X = 10|\theta = 1/8)\pi(\theta = 1/8)]\end{aligned}$$

$$= \frac{0.1048 \times 0.2}{0.1048 \times 0.2 + 0.0413 \times 0.5 + 0.0008 \times 0.3}$$

$$= 0.501$$

Example 4.4

Similarly,

$$\begin{aligned} \pi(\theta = 1/4|x = 10) &= \cdots = 0.493 \qquad \text{and} \\ \pi(\theta = 1/8|x = 10) &= \cdots = 0.006. \end{aligned}$$

Thus, the posterior distribution for θ is as shown in Table 4.2, with the most likely value of θ now being 1/2, and standard deviation $SD(\theta|x = 10) = 0.126$.

	θ	$\pi(\theta\|x = 10)$
"great"	1/2	0.501
"good"	1/4	0.493
"poor"	1/8	0.006

Example 4.4

Suppose now we want to predict the number of sales Y in the next 24 days.

If there have been no changes in the sales process (no special advertising campaigns etc) then we can take $Y|\theta \sim Po(24\,\theta)$.

Determine the **predictive probability function** for Y.

Solution to Example 4.4 (1/2)

As θ is discrete, the predictive probability function for Y is

$$
\begin{aligned}
f(y|x=10) &= \sum_{\theta=\frac{1}{2},\frac{1}{4},\frac{1}{8}} f(y|\theta)\,\pi(\theta|x=10) \\
&= \left(\frac{12^y e^{-12}}{y!} \times 0.501\right) + \left(\frac{6^y e^{-6}}{y!} \times 0.493\right) \\
&\qquad + \left(\frac{3^y e^{-3}}{y!} \times 0.006\right) \\
&= \frac{0.501(12^y e^{-12}) + 0.493(6^y e^{-6}) + 0.006(3^y e^{-3})}{y!}
\end{aligned}
$$

Solution to Example 4.4 (2/2)

For example:

$$\begin{aligned}\Pr(Y = 10|x = 10) &= \frac{0.501(12^{10}e^{-12}) + 0.493(6^{10}e^{-6}) + 0.006(3^{10}e^{-3})}{10!} \\ &= 0.073.\end{aligned}$$

Example 4.4

This probability can be compared with a more **naive** predictive probability calculated assuming that $\theta = \hat{\theta}$, the likelihood mode.

Here, $\hat{\theta} = 1/2$ and so $Y|\theta = \hat{\theta} \sim Po(24\hat{\theta}) \equiv Po(12)$, whence

$$\Pr(Y = 10|\theta = \hat{\theta}) = \frac{12^{10}e^{-12}}{10!} = 0.1048.$$

In the same manner, we can calculate the **entire predictive distribution** and **entire naive predictive distribution**; see Table 4.3.

Example 4.4

y	correct $f(y\|x = 10)$	naive $f(y\|\theta = \hat{\theta})$
0	0.002	0.000
1	0.008	0.000
2	0.024	0.000
3	0.046	0.002
4	0.070	0.005
5	0.086	0.013
6	0.092	0.025
7	0.090	0.044
8	0.084	0.066
9	0.078	0.087
10	0.073	0.105
11	0.068	0.114
12	0.063	0.114
13	0.055	0.106
14	0.046	0.090
15	0.037	0.072
16	0.027	0.054
17	0.019	0.038
18	0.013	0.026
19	0.008	0.016
20	0.005	0.010
.	.	.

Example 4.4

Notice that the correct predictive probability distribution has more probability out in the tails of its distribution – that is, the probabilities of 0, 1, 2, . . . are larger than their “naive” equivalents.

- Common occurrence
- Due to ignoring the uncertainty about the parameter estimate

Example 4.4

Essentially, the naive predictive distribution is a predictive distribution which, instead of using the correct posterior distribution, uses the degenerate posterior distribution

$$\pi^*(\theta|x = 10) = \begin{cases} 1 & \text{if } \theta = 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

a distribution with standard deviation $SD_{\pi^*}(\theta|x = 10) = 0$.

The *correct* posterior standard deviation of θ is $SD_{\pi}(\theta|x = 10) = 0.126$.

Example 4.4

Therefore, the predictive distribution using the naive posterior π^* is, loosely speaking, **too confident** that it "knows" the value of θ and so produces a predictive distribution with too small a standard deviation:

$$SD(Y|x = 10) = \begin{cases} 4.26 & \text{using the correct } \pi(\theta|x = 10) \\ 3.46 & \text{using the naive } \pi^*(\theta|x = 10). \end{cases}$$

Example 4.4

We can also use the above table of predictive probabilities to determine a **prediction interval** for Y.

Using the correct predictive distribution, and recalling the highest density feature of prediction intervals, we obtain

$$\Pr(2 \leq Y \leq 17 | x = 10) = 0.959.$$

The corresponding naive calculation is

$$\Pr(6 \leq Y \leq 19 | \theta = \hat{\theta}) = 0.958.$$

Hence the correct (approximate) 96% prediction interval for Y is $\{2, 3, \ldots, 17\}$.

The naive version, and hence narrower interval, is $\{6, 7, \ldots, 19\}$.

Definition (Beta–binomial distribution)

The random variable Y follows a Beta–binomial $BetaBin(n, a, b)$ distribution (n positive integer, $a > 0$, $b > 0$) if it has probability function

$$f(y|n, a, b) = \binom{n}{y} \frac{B(y + a, b + n - y)}{B(a, b)}, \qquad y = 0, 1, \ldots, n,$$

where $B(a, b)$ is the beta function defined earlier. It can be shown that

$$E(Y) = \frac{na}{a + b} \qquad \text{and} \qquad Var(Y) = \frac{nab(a + b + n)}{(a + b)^2(a + b + 1)}.$$

Example 4.5

Suppose that X is the number of defective items in a sample of size 5.

If the items are defective independently of one another and they each have the same probability θ of being defective then $X|\theta \sim Bin(5, \theta)$.

Suppose we believe that defective items are quite unlikely and so take a $Beta(1, 19)$ prior distribution with mean and standard deviation

$$E(\theta) = 0.05 \qquad \text{and} \qquad SD(\theta) = 0.048.$$

Example 4.5

Suppose we take a sample of size 5 and observe $x = 1$ defective item.

In this case, the likelihood mode is $\hat{\theta} = 1/5 = 0.2$, higher than the prior mean.

We can combine the prior with our data via Bayes' Theorem, to show that the posterior distribution is a $Beta(2, 23)$ distribution, with mean and standard deviation

$$E(\theta|x = 1) = 0.08 \qquad \text{and} \qquad SD(\theta|x = 1) = 0.053.$$

The posterior mean is larger than the prior mean and the standard deviation has also increased (slightly).

Example 4.5

If we observe another sample of 5 items, what is the predictive probability distribution of the number found to be defective?

Solution to Example 4.5

Suppose the number of defective items in this future sample is Y, with $Y|\theta \sim Bin(5, \theta)$.

The predictive probability function of Y is, for $y = 0, 1, 2, \ldots, 5$,

$$\begin{aligned} f(y|x=1) &= \int_{\Theta} f(y|\theta)\,\pi(\theta|x=1)\,d\theta \\ &= \int_0^1 \binom{5}{y} \theta^y (1-\theta)^{5-y} \times \frac{\theta(1-\theta)^{22}}{B(2,23)}\,d\theta \\ &= \binom{5}{y} \frac{1}{B(2,23)} \int_0^1 \theta^{y+1}(1-\theta)^{27-y}\,d\theta \\ &= \binom{5}{y} \frac{B(y+2, 28-y)}{B(2,23)}, \end{aligned}$$

that is, $Y|x = 1 \sim BetaBin(5, 2, 23)$.

Example 4.5

We can compare this predictive distribution with a naive predictive distribution based on an estimate of θ, for example, the likelihood mode or the posterior mode.

Here we shall base our naive predictive distribution on the posterior mode $\hat{\theta} = 1/23$, that is, use the distribution

$$Y|\theta = \hat{\theta} \sim Bin(5, 1/23).$$

Thus, the naive predictive probability function is, for $y = 0, 1, \ldots, 5$,

$$f(y|\theta = \hat{\theta}) = \binom{5}{y} \hat{\theta}^y (1 - \hat{\theta})^{5-y} = \binom{5}{y} \frac{22^{5-y}}{23^5}.$$

Example 4.5

Numerical values for the predictive and naive predictive probability functions are given in Table 4.4:

y	correct $f(y\|x=1)$	naive $f(y\|\theta=\hat{\theta})$
0	0.680	0.801
1	0.252	0.182
2	0.058	0.017
3	0.009	0.001
4	0.001	0.000
5	0.000	0.000

Example 4.5

As before, the naive predictive distribution

- allows only one value for θ: $\theta = 1/23$
- has a standard deviation of zero: $SD(\theta|x = 1) = 0$

The **correct** posterior standard deviation is $SD(\theta|x = 1) = 0.053$.

This results in

$$SD(Y|x = 1) = \begin{cases} 0.652 & \text{using the correct predictive dist.} \\ 0.456 & \text{using the naive predictive dist.,} \end{cases}$$

these values being calculated from $BetaBin(5, 2, 23)$ and binomial $Bin(5, 1/23)$ distributions.

Example 4.5

Using the numerical table of predictive probabilities, we can see that $\{0, 1\}$ is a 93.2% prediction set/interval.

This is to be contrasted with the more "optimistic" calculation using the naive predictive distribution which shows that $\{0, 1\}$ is a 98.3% prediction set/interval.

Predictive distribution

In the previous example, a non–trivial integral had to be evaluated.

However, when the past data $\boldsymbol{x}$ and future data y are independent (given θ) and we use a conjugate prior distribution, another (easier) method can be used to determine the predictive distribution.

Predictive distribution

Using Bayes' Theorem, the posterior density for θ given $\boldsymbol{x}$ and y is

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

Predictive distribution

Rearranging, we obtain

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

The right–hand–side of this equation looks as if it depends on θ, but, in fact, any terms in θ will be cancelled between the numerator and denominator.

Example 4.5 revisited

We had

$$\theta \sim Beta(1, 19), \quad X|\theta \sim Bin(5, \theta), \quad Y|\theta \sim Bin(5, \theta),$$

from which we obtain

$$\theta|x = 1 \sim Beta(2, 23), \quad \theta|x = 1, y \sim Beta(y + 2, 28 - y).$$

Example 4.5 revisited

Therefore, for $y = 0, 1, 2, \ldots, 5$

$$
\begin{aligned}
f(y|x=1) &= \frac{f(y|\theta)\pi(\theta|x=1)}{\pi(\theta|x=1,y)} \\
&= \frac{\binom{5}{y}\theta^y(1-\theta)^{5-y} \times \frac{\theta(1-\theta)^{22}}{B(2,23)}}{\frac{\theta^{y+1}(1-\theta)^{27-y}}{B(y+2,28-y)}} \\
&= \binom{5}{y}\frac{\theta^{y+1}(1-\theta)^{27-y}B(y+2,28-y)}{\theta^{y+1}(1-\theta)^{27-y}B(2,23)} \\
&= \binom{5}{y}\frac{B(y+2,28-y)}{B(2,23)}.
\end{aligned}
$$

Definition (Inverse–Beta distribution)

The random variable Y follows a Inverse-Beta $InBe(a, b, c)$ distribution ($a > 0$, $b > 0$, $c > 0$) if it has probability density function

$$f(y|a, b, c) = \frac{c^b y^{a-1}}{B(a, b)(y + c)^{a+b}} \qquad y > 0,$$

where $B(a, b)$ is the beta function defined earlier. It can be shown that

$$E(Y) = \frac{ac}{b-1} \qquad \text{and} \qquad Var(Y) = \frac{ac^2(a+b-1)}{(b-1)^2(b-2)}.$$

The distribution gets its name because $Y/(Y + c) \sim Beta(a, b)$. Also note that if $Y \sim InBe(a, b, 1)$ then $cY \sim InBe(a, b, c)$.

Example 4.6

Suppose we have a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ from a $Ga(k, \theta)$ distribution, where k is known, and our prior beliefs are described by a $Ga(g, h)$ distribution.

The likelihood function is

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^k x_i^{k-1} e^{-\theta x_i}}{\Gamma(k)} \propto \theta^{nk} e^{-n\bar{x}\theta}.$$

Therefore, using Bayes Theorem, the posterior density is

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)\, f(\boldsymbol{x}|\theta) \\
&\propto \theta^{g-1} e^{-h\theta} \times \theta^{nk} e^{-n\bar{x}\theta}, \quad \theta > 0 \\
&\propto \theta^{g+nk-1} e^{-(h+n\bar{x})\theta}, \quad \theta > 0.
\end{aligned}$$

Example 4.6

Hence, the posterior distribution is a $Ga(G = g + nk, H = h + n\bar{x})$ distribution.

Notice that this implies that the gamma distribution is the conjugate prior distribution for the model "random sample from a $Ga(k, \theta)$ distribution, with k known".

Determine the predictive distribution for a future outcome Y.

Solution to Example 4.6 (1/4)

The predictive density function is, for $y > 0$,

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

Now

- $y|\theta \sim Ga(k, \theta)$
- $\theta|\boldsymbol{x} \sim Ga(G, H)$
- $\theta|\boldsymbol{x}, y \sim ??$

Solution to Example 4.6 (2/4)

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}, y) &\propto \pi(\theta|\boldsymbol{x}) \times f(y|\theta) \\
&\propto \theta^{G-1} e^{-H\theta} \times \theta^{k} y^{k-1} e^{-\theta y} \\
&\propto \theta^{G+k-1} e^{-(H+y)\theta}
\end{aligned}$$

giving $\theta|\boldsymbol{x}, y \sim Ga(G+k, H+y)$.

Solution to Example 4.6 (3/4)

Therefore, the predictive density function is, for $y > 0$,

$$\begin{aligned} f(y|\boldsymbol{x}) &= \frac{f(y|\theta)\pi(\theta|\boldsymbol{x})}{\pi(\theta|\boldsymbol{x}, y)} \\ &= \frac{\dfrac{\theta^k y^{k-1} e^{-\theta y}}{\Gamma(k)} \times \dfrac{H^G \theta^{G-1} e^{-H\theta}}{\Gamma(G)}}{\dfrac{(H+y)^{G+k} \theta^{G+k-1} e^{-(H+y)\theta}}{\Gamma(G+k)}} \end{aligned}$$

Solution to Example 4.6 (4/4)

Giving

$$
\begin{aligned}
f(y|\boldsymbol{x}) &= \frac{\theta^{G+k-1}e^{-(H+y)\theta}y^{k-1}H^G\Gamma(G+k)}{\theta^{G+k-1}e^{-(H+y)\theta}(H+y)^{G+k}\Gamma(k)\Gamma(g)} \\
f(y|\boldsymbol{x}) &= \frac{\Gamma(G+k)H^G y^{k-1}}{\Gamma(k)\Gamma(G)(H+y)^{G+k}} \\
&= \frac{H^G y^{k-1}}{B(k,G)(H+y)^{G+k}},
\end{aligned}
$$

that is, $Y|\boldsymbol{x} \sim InBe(k, G, H)$.

Example 4.6

Consider the case where the data follow an exponential distribution, that is, where $k = 1$.

Determine the predictive density function and the $100(1 - \alpha)\%$ prediction interval for Y.

Solution to Example 4.6 (1/3)

Now

$$B(1, G) = \frac{\Gamma(G)\Gamma(1)}{\Gamma(G+1)} = \frac{(G-1) \times (G-2) \times \cdots}{G \times (G-1) \times (G-2) \times \cdots} = \frac{1}{G}.$$

So the predictive density function is

$$f(y|\boldsymbol{x}) \quad = \quad \frac{GH^G}{(H+y)^{G+1}}, \quad y > 0.$$

The prediction interval includes the most likely values of Y and so, given the shape of the density, must take the form $(0, b)$. The end–point b must satisfy

$$\int_0^b \frac{GH^G}{(H+y)^{G+1}}\, dy = 1 - \alpha.$$

Solution to Example 4.6 (2/3)

Using the substitution $z = H + y$, we have

$$\begin{aligned}
\int_0^b \frac{GH^G}{(H+y)^{G+1}}\,dy &= \int_H^{b+H} \frac{GH^G}{z^{G+1}}\,dz \\
&= \left[-\frac{H^G}{z^G}\right]_H^{b+H} \\
&= 1 - \frac{H^G}{(b+H)^G}.
\end{aligned}$$

Solution to Example 4.6 (3/3)

Therefore, we require

$$1 - \frac{H^G}{(b+H)^G} = 1 - \alpha \qquad \Longrightarrow \frac{b+H}{H} = \alpha^{-1/G}$$

$$\Longrightarrow b = H\left\{\alpha^{-1/G} - 1\right\}.$$

Hence, a $100(1-\alpha)\%$ prediction interval for Y is

$$\left(0, H\left\{\alpha^{-1/G} - 1\right\}\right).$$

The End!!