



MAS2905

**Probability and Bayesian
Methods with R**

Part 2

Semester 2, 2020—2021

Lecturer: Dr. Lee Fawcett

School of Mathematics, Statistics & Physics

MAS2905: Probability and Bayesian Methods with R

Lecturer: Dr. Lee Fawcett

Office: Room 2.07, Herschel Building Phone: 208 7228

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

Classes

As you now know, things are very different this academic year! The aim was for you to have a blended learning experience 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. 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

Zoom details will be available on Canvas; make sure you attend the session you have been allocated to!

- Fortnightly *synchronous online* (SO) classes, alternating with the small group problems classes. These will take place on Fridays 12–1. Again, Zoom details will be advertised on Canvas.
- 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. This will mainly consist of directed reading (e.g. lecture notes), completing exercises/practice questions, summaries/discussions of learning objectives, assessment feedback and other fun activities (e.g. quizzes!). These activities must be completed in your own time, whenever you can fit them in, but it is vital that they are completed by their completion dates as advertised on Canvas.

Students can take part in other NSO activities, should they wish, including interacting with the Canvas discussion board; optional Zoom drop-in sessions; and one-to-one Zoom meetings. And of course, you will all be expected to prepare and complete homeworks in your own time, and revise for the end-of-year exam!

Schedule

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

Assessment

The full 20 credit MAS2905 module will be assessed via a formal written examination in summer 2021 (likely 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)

Exam questions in section A, relating to the material from semester 1 of the module, will focus mainly on the theory and probability/statistics developed throughout the semester (rather than programming in R).

The semester 2 homeworks will likely take the form of traditional written assignments, uploaded to Canvas. Some questions within these assignments might require you to produce plots in R, and so it might be a good idea to type your solutions and upload a single PDF file for each assignment (which includes any necessary plots/computing work). The assignment questions will be taken from the bank of questions in Chapter 5 of these course notes.

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

At various points in the module you will be directed to additional reading, should you want extra examples or just another source of information relating to the material presented in the notes. This will primarily be text books (many of which are available to view online), but perhaps other sources (e.g. journal publications, magazine articles etc.).

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?

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.

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 referred to as a *point estimation* problem.
- What is a plausible range of values?
This is referred to as 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.

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! However, there are some connections between them and their Bayesian alternatives which will be explored later in the course. For example we'll see that the idea of the *likelihood function* is very important in Bayesian statistics.

Motivating example

Air France Flight 447 disappeared over the Atlantic Ocean on June 1st 2009, during an overnight flight from Rio de Janeiro to Paris. A search and rescue mission was launched the following morning. After five days some floating wreckage was found, but there was no sign of the “black box” recorders which would reveal what had happened to the flight, and the search stalled.

The black boxes could be anywhere within an area of the South Atlantic the size of Switzerland (6,500 square miles). The “mid-Atlantic ridge” on the ocean floor lies between two tectonic plates and is just as mountainous as Switzerland! It's also so remote that scientists have not yet charted the sea-bed. The search method used was sonar-detectors, emitting sound waves which would bounce back once they hit something. However, after two years of meticulously searching an area north of the planes trajectory (after analysing debris drift), nothing had been found.

Metron, Inc., of Retson, Virginia had been performing Bayesian analyses to help the search effort. Included in their analysis were:

1. Data from 9 previous airline crashes involving loss of pilot control - reduced search area to 1,600 square miles.
2. Expert opinions on the credibility of the flight data.
3. Expert opinions about whether or not the black box ‘pingers’ might have become damaged on impact.
4. Positions/recovery times of bodies found drifting - expert opinions assigned to the reliability of this data because of the turbulent equatorial waters.
5. Expert information from oceanographers: Sea state, visibility, underwater geography,...

All were combined through Bayes Theorem to give a probability map of the most likely locations to search. In April 2011 they decided to try assuming that the ‘pingers’ in item 3 were probably damaged. One week later the black boxes were found!¹

¹For more details see <https://www.technologyreview.com/s/527506/how-statisticians-found-air-france-flight-447-two-years-after-it-crashed-into-atlantic/>.

The frequentist/Bayesian debate

This example illustrates some key features of the Bayesian approach:

- It incorporates expert opinion.
- It outputs a probability distribution over the possible choices.
- The calculations required are all based on Bayes Theorem.

In contrast in frequentist statistics it is difficult (but not impossible) to incorporate opinions, and the output is not in the form of a probability distribution. For example a 95% frequentist confidence interval is a range which will contain the true value 95% of the time if the analysis is repeated a large number of times for different data obtained under the same process². In contrast a Bayesian 95% interval estimate (see Chapter 4) is a range with a 95% probability of containing the true value. This Bayesian version is arguably much easier to interpret.

There has been a decades-long argument about whether Bayesian or frequentist methods are better. Both sides have some good points. For example, there is the argument we've just made about Bayesian interval estimates being easy to interpret. On the other hand one of the main arguments against the Bayesian approach is that it is *subjective*. This is because its results are based on expert opinions, and if we got a different opinion from another expert our results would change even though the data are the same! Bayesians have counter-arguments. For example many would argue that subjectivity is in fact a desirable feature, or that frequentist methods also have hidden subjective elements e.g. using 5% as a significance threshold rather than some other value. The philosophical debate seems unlikely to end any time soon!

In practice many problems are better suited to be tackled with methods from one approach or the other. For example Bayesian statistics was well suited to the Air France example as incorporating expert knowledge was particularly important in the solution. Therefore it's good to be familiar with both Bayesian and frequentist methods.

The history of Bayesian statistics

The reason why we use the word “Bayesian” is because Bayes’ Theorem is crucial for statistical analysis if we adopt the Bayesian approach. Thomas Bayes (1702–1761; see picture below) was a Presbyterian minister in Tunbridge Wells, Kent. Bayes’ solution to a problem in probability theory was presented in the *Essay towards Solving a Problem in the Doctrine of Chances*, published posthumously by his friend Richard Price in the *Philosophical Transactions of the Royal Society of London*. This paper gave us Bayes’ Theorem. Its modern form is due to Pierre-Simon Laplace who generalised Bayes’ work

²Here 95% is the *long-run frequency* that the interval contains the correct value. This kind of property is where the name “frequentist” comes from.



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.*

Dear Sir,

Read Dec. 23, 1763. **I** Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many in it as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circum-

Figure 1: Portrait believed to be of Thomas Bayes, and letter from Richard Price to the Philosophical Transactions of the Royal Society of London

and produced the first theory of Bayesian inference, which was referred to as “inverse probability”. Laplace’s theory had weaknesses³ and in the early 20th century it was replaced by the more rigorous frequentist approach. R. A. Fisher was one of the leading creators of this approach. He was also the first to use the term “Bayesian statistics”.

Nonetheless there was still some interest in Bayesian methods. For example, Alan Turing’s group invented some new Bayesian methods to help break the Enigma code in World War II. Before and, increasingly, after the war a few researchers worked on developing a rigorous theory of Bayesian statistics.

Bruno de Finetti (1906–1985) was an Italian probabilist who developed his ideas on *subjective probability*⁴ from the 1920s onwards, drawing upon ideas from H. Jeffreys, I.J. Good (one of the Enigma codebreakers) and B.O. Koopman. His classic book on the topic is *The Theory of Probability* (1974).

Dennis Lindley (1923–2013) was a leading British Bayesian. In his early career, he worked to find a mathematical basis for the subject of statistics. In 1954, Lindley met Leonard Savage and both found a deeper justification for statistics in Bayesian theory, turning into critics of the classical statistical inference they had hoped to justify.

Many of these theoretical advances are covered in this course. However, Bayesian statistics was still little used in practice. The difficulty was that applying Bayes theorem often

³For example it relied mainly on *flat priors*, which will be discussed and criticised in Chapter 3.

⁴Covered in Chapter 1.



Figure 2: Pierre-Simon Laplace (left) and R. A. Fisher (right)

became very mathematically challenging, resulting in problems that were too difficult or impossible to solve by hand.

Recent developments

In the 1990s a solution to these mathematical problems was developed, making use of the emergence of more powerful computers and “Markov chain Monte Carlo” (MCMC) numerical algorithms. If you take 3rd and 4th year Bayesian courses you will be introduced to these techniques. MCMC, and subsequent developments, have revolutionised the use of Bayesian Statistics, to the extent that today Bayesian data analyses are as popular as frequentist approaches and are routinely used in fields as diverse as artificial intelligence, biology, astrophysics and sociology.

Newcastle’s contribution

Since 1980, the number of academic staff in Mathematics & Statistics at Newcastle publishing advanced research using Bayesian methods has increased dramatically. In the 1980s, there was only one Bayesian at Newcastle. Now there are at least 12. Our Bayesian research topics include environmental extremes, genetics, and modelling biological systems at a molecular level.

Chapter 1

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.

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. $\theta = 1/2$. We could then test the null hypothesis $H_0 : \theta = 1/2$, giving

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

From this p -value, we would conclude that we had very strong evidence against the null hypothesis (that the choices were made randomly), and 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. 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.

The traditional approach to Statistics, sometimes called *Frequentist Statistics* or *Classical Statistics*, may try to take this prior information into account by modifying the pure significance testing approach described above to an assessment of an “appropriate” hypothesis test. 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 by constructing a probability distribution which describes how likely we believe different values are to occur. 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.

1.1 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. Most of your courses will have used the frequentist interpretation: repeated sampling ideas are fundamental to the techniques described.

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

win $\pounds c$ if A occurs and $\pounds 0$ if A^c occurs.

How much would you be prepared to pay (stake) for placing such a bet? In terms of expected winnings, you should be prepared to stake $\pounds cp$ if you believe that $\Pr_H(A) = p$. Why?



One problem with this approach is that, in general, p will depend on c : a person who is willing to bet $\pounds 1$ on the spin of a coin to win $\pounds 2$ if it lands heads may refuse to bet if the stakes are raised to $\pounds 1000$ – most people are *risk-averse*. Therefore, we shall restrict our attention to the $c = 1$ case: pay $\pounds p$ for the bet

win $\pounds 1$ if A occurs and $\pounds 0$ if A^c occurs.

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 $\pounds 0.50$. However, if you weren’t honest you might try to buy the bet for any amount less than $\pounds 0.50$, say $\pounds 0.20$. If you were hosting the bet, you would take the bet for any amount more than $\pounds p$, say for $\pounds 0.80$. These conflicting interests can be offset if, when choosing p , it is equally likely that you are hosting the bet or placing the bet. In such circumstances it is in your own interests to give the value of p that you believe to be “correct”.

Example 1.1

1. The probability $\Pr(\text{Newcastle Utd win the Championship this season})$ could only be determined using a subjective assessment.
2. The probability $\Pr(\text{M\&S student chosen at random was born in January})$ could be determined using a frequency or classical interpretation (with a list of all M\&S students and their birth dates) or a subjective interpretation.
3. The probability $\Pr(\text{England win the toss at a given Test Match})$ could be determined using either the classical or subjective interpretation.

There are potential drawbacks with each of these ways of understanding probability:

Frequency interpretation

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 Section 2.4 of the preface to these lecture notes).

Classical interpretation

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

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

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

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.

1.2 Bayes' Theorem

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

Definition 1.1

Consider two events E and F , where $\Pr(F) > 0$. The *conditional probability* of E given that F has occurred is

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

Definition 1.2

The events $E_1, E_2, \dots, E_n$ form a *partition* of the sample space $\mathcal{S}$ if they are disjoint events ($E_i \cap E_j = \emptyset$, $i \neq j$) with $\Pr(E_i) > 0$, $i = 1, 2, \dots, n$, and $\cup_{i=1}^n E_i = \mathcal{S}$. Figure 1.1 gives a diagram of a typical partition with an additional event F .

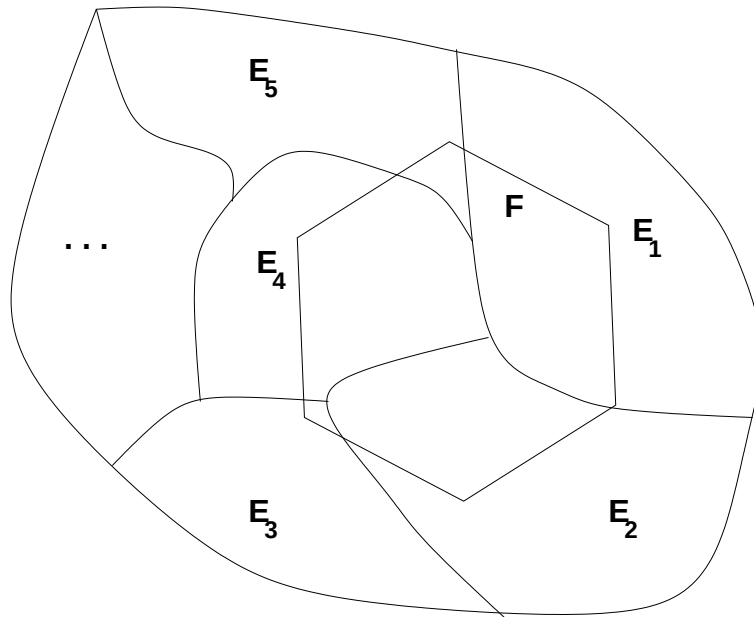


Figure 1.1: Diagram of a partition $E_1, E_2, \dots, E_n$ and an event F

Law of Total Probability

If $E_1, E_2, \dots, E_n$ are a *partition* of $\mathcal{S}$ and F is any event then

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

Proof

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

$$\begin{aligned}\Pr(F) &= \sum_{i=1}^n \Pr(F \cap E_i) \\ &= \sum_{i=1}^n \Pr(F|E_i)\Pr(E_i).\end{aligned}$$

Bayes' Theorem

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

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

Proof**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...*

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

x_1 : overheating only,
 x_2 : irregular traction only,
 x_3 : both symptoms.

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

	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

Table 1.1: Likelihood of symptoms for both faults.



...Solution to Example 1.3...

 ...Solution to Example 1.3 continued...

	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

Table 1.2: Posterior probabilities of the faults for various symptoms.

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

$$\begin{aligned} \text{Prior odds : } \frac{\Pr(\theta_1)}{\Pr(\theta_2)} &= \frac{0.4}{0.6} = \frac{2}{3} \quad \text{or 3:2 in favour of } \theta_2 \\ \text{Posterior odds : } \frac{\Pr(\theta_1|x_3)}{\Pr(\theta_2|x_3)} &= \frac{0.625}{0.375} = \frac{5}{3} \quad \text{or 5:3 in favour of } \theta_1. \end{aligned}$$

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 best (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)

Table 1.3: Diagnostic rule for faults.



...Solution to Example 1.3 continued...

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...*

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

The main problem with this solution is that, in order to use this table we must know the exact value of θ : we have actually found an expression for $\Pr(K|C, \theta)$ and not for $\Pr(K|C)$. If we know θ fairly accurately – say it was between 0.49 and 0.51 – then, in practice, we can conclude that $\Pr(K|C)$ is around 0.83. However, if we are less certain about a correct value for θ we might be able to express our uncertainty through a probability distribution for θ . We will see more about this in the next Chapter.

$\Pr(K)$ $= \theta$	$\Pr(K C)$ $= 5\theta/(1 + 4\theta)$
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

Table 1.4: Values of $\Pr(K|C)$ for various values of $\Pr(K)$.

1.3 Likelihood

Suppose that an experiment results in data $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ and we decide to model the data using a probability (density) function $f(\mathbf{x}|\theta)$. This p(d)f describes how likely different data $\mathbf{x}$ are to occur given a value of the (unknown) parameter θ . However, once we have observed the data, $f(\mathbf{x}|\theta)$ 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 $L(\theta|\mathbf{x})$ or $L(\theta)$ but, whatever the notation used for the likelihood function, it is simply the joint probability (density) function of the data, $f(\mathbf{x}|\theta)$, regarded as a function of θ rather than of $\mathbf{x}$.

The likelihood function can be simplified if we have further structure in the data. For example, we may have independent observations, in which case

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

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

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n f_X(x_i|\theta). \quad (1.2)$$

In this course, we will not consider models with correlated observations. Moreover, we will concentrate on how to make inferences from random samples using prior information. This will require extensive use of the (1.2) form of the likelihood function.

Example 1.5

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ of radioactive particle counts. A typical model for such data would be $X_i|\theta \sim \text{Poisson}(\theta)$, usually abbreviated $X_i|\theta \sim \text{Po}(\theta)$, (independent). Determine the likelihood function for θ .



...Solution to Example 1.5...

Example 1.6

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ of times between radioactive particle emissions. If the emissions occur randomly in time then a plausible model for such data would be $X_i|\theta \sim \text{Exponential}(\theta)$, usually abbreviated $X_i|\theta \sim \text{Exp}(\theta)$, (independent). Determine the likelihood function for θ .



...Solution to Example 1.6...

Example 1.7

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ from a normal distribution: $X_i | \mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \dots, n$ (independent). Determine the likelihood function for (μ, σ) .



...Solution to Example 1.7...

1.4 Sufficiency

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

$$\begin{aligned} f(\mathbf{x}|\theta) &= \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} \\ &= \left(\prod_{i=1}^n x_i! \right)^{-1} \times \theta^{n\bar{x}} e^{-n\theta}. \end{aligned}$$

Notice that this likelihood function depends on the data only through $\prod_{i=1}^n (x_i!)^{-1}$ and $\bar{x}$. Further, in $f(\mathbf{x}|\theta)$, θ only “interacts” with $\bar{x}$ — the other term simply scales $f(\mathbf{x}|\theta)$ — so that, for example, the point at which $f(\mathbf{x}|\theta)$ is maximized is determined only by $\bar{x}$. 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 .

Definition 1.3

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

Definition 1.4

The statistic $T(\mathbf{X})$ is *sufficient* for θ if $f(\mathbf{x}|T(\mathbf{X}) = t)$ 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...

 ...*Solution to Example 1.8 continued...*

Definition 1.5

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

Factorisation Theorem

Under certain regularity conditions

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

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...



...*Solution to Example 1.9...*

Example 1.10

Suppose we have a random sample from an exponential distribution: $X_i|\theta \sim \text{Exp}(\theta)$, $i = 1, 2, \dots, n$ (independent). Determine a sufficient statistic for θ .

 ...Solution to Example 1.10...

Example 1.11

Suppose we have a random sample from a normal distribution: $X_i | \mu, \sigma \sim N(\mu, \sigma^2)$, $i = 1, 2, \dots, n$ (independent). Determine sufficient statistics for (μ, σ) .



...Solution to Example 1.11...

Comment

If a parameter has a sufficient statistic then, in fact, it has an infinite number of sufficient statistics. For example, in Example 1.4, where we had a random sample from an exponential distribution, we found that $T = \sum X_i$ was sufficient for θ . However, (obviously) the whole data $\mathbf{X}$ is sufficient for θ since

$$\begin{aligned} f(\mathbf{x}|\theta) &= 1 \times f(\mathbf{x}|\theta) \\ &= h(\mathbf{x}) g(\mathbf{x}, \theta), \end{aligned}$$

where $h(\mathbf{x}) = 1$, $g(t, \theta) = f(t|\theta)$ and $T = \mathbf{X}$. Also, any bijective (1-1) function of T is also sufficient for θ . For example, since $\bar{x} > 0$

$$\begin{aligned} f(\mathbf{x}|\theta) &= 1 \times \theta^n \exp(-n\theta\bar{x}) \\ &= h(\mathbf{x}) g_1(\bar{x}, \theta), \text{ or} \\ &= h(\mathbf{x}) g_2((\bar{x})^2, \theta), \text{ or} \\ &= h(\mathbf{x}) g_3((\bar{x})^3, \theta), \text{ or} \\ &= h(\mathbf{x}) g_4((\bar{x})^4, \theta), \text{ or} \\ &= \dots \end{aligned}$$

where $h(\mathbf{x}) = 1$ and $g_j(t, \theta) = \theta^n \exp(-\theta t^{1/j})$, $j = 1, 2, \dots$. Therefore, $\sum X_i$ is sufficient for θ , but so are any of $\mathbf{X}$, $\bar{X}$, $(\bar{X})^2$, $(\bar{X})^3$, $(\bar{X})^4$, $\dots$, $\log(\bar{X})$, $\exp(\bar{X})$, $\dots$.

Further, if $\mathbf{T}$ is sufficient for θ and $\mathbf{S}$ is any other statistic then $(\mathbf{T}, \mathbf{S})$ is also sufficient for θ . For example,

$$\begin{aligned} f(\mathbf{x}|\theta) &= 1 \times \theta^n \exp(-n\theta\bar{x}) \\ &= h(\mathbf{x}) g_1(\bar{x}, x_1, \theta), \text{ or} \\ &= h(\mathbf{x}) g_2(\bar{x}, x_2 \sin x_6, \theta), \text{ or} \\ &= \dots \end{aligned}$$

where $h(\mathbf{x}) = 1$ and $g_j(t, \mathbf{s}, \theta) = \theta^n \exp(-\theta t)$.

Minimal sufficiency

The main role of a sufficient statistic is to summarise information in the data about the parameters — this should be expressed as concisely as possible. Therefore, we want as few sufficient statistics as possible.

Definition 1.6

A statistic T is minimal sufficient if it is a function of every other sufficient statistic.

This definition does not uniquely define a minimally sufficient statistic, since any bijective function of a minimally sufficient statistic is also minimal sufficient. However, it does achieve the greatest reduction of the data without losing any information about the parameters.

