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

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

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

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

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 £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 £1 on the spin of a coin to win £2 if it lands heads may refuse to 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

$$\text{win £1 if } A \text{ occurs and £0 if } A^{\rm c} \text{ occurs.}$$

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 £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. 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(Newcastle Utd win the Championship this season) could only be determined using a subjective assessment.
2. The probability Pr(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(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
$$\begin{aligned}\Pr(A_1|H) > \Pr(A_2|H) \quad \text{and} \quad \Pr(A_2|H) > \Pr(A_3|H)&\\ \implies \quad \Pr(A_1|H) > \Pr(A_3|H).&\end{aligned}$$

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 $E_1, E_2, \ldots, E_n$ 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.

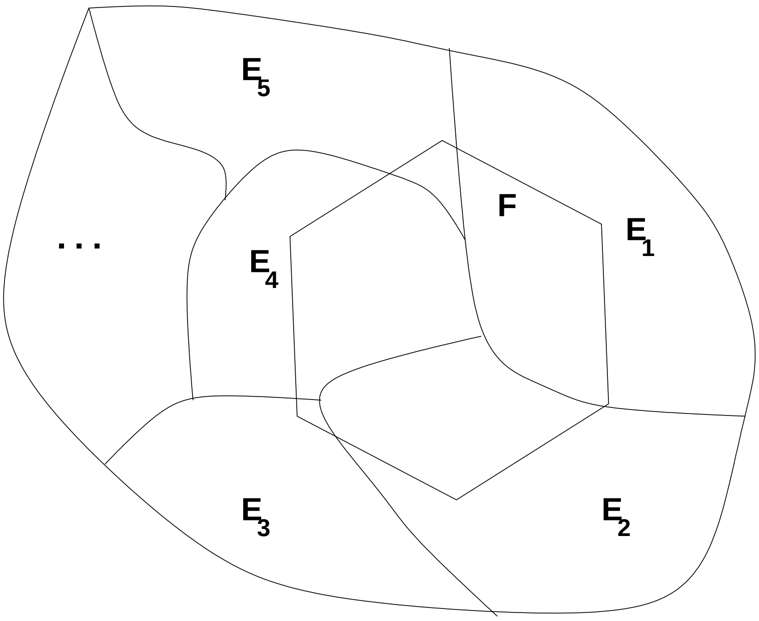

Figure 1.1: Diagram of a partition $E_1, E_2, \ldots, E_n$ and an event F

Law of Total Probability

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

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

Proof

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

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

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

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

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	I/T	Both
	x_1	x_2	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	I/T	Both
	x_1	x_2	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.

$\Pr(K)$	$\Pr(K\|C)$
$=\theta$	$=5\theta/(1+4\theta)$
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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ 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 $\boldsymbol{x}$.

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

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

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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ of radioactive particle counts. A typical model for such data would be $X_i|\theta \sim \mathit{Poisson}(\theta)$, usually abbreviated $X_i|\theta \sim \mathit{Po}(\theta)$, (independent). Determine the likelihood function for θ.

✎*...Solution to Example 1.5...*

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...*

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...*

1.4 Sufficiency

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

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

Notice that this likelihood function depends on the data only through $\prod_{i=1}^{n}(x_i!)^{-1}$ and $\bar{x}$. Further, in $f(\boldsymbol{x}|\theta)$, 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 $\bar{x}$. 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 $T(\boldsymbol{X})$ 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...*

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

Definition 1.5

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

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...*

✎*...Solution to Example 1.9...*

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...*

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 $T = \Sigma X_i$ was sufficient for θ. However, (obviously) the whole data $\boldsymbol{X}$ is sufficient for θ since

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

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

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

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

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

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

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

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.