# MAS2903: Introduction to Bayesian Methods

Dr. Lee Fawcett

**Semester 2, 2018—2019**

# 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`

# Timetable and Administrative arrangements

- **Lectures** are on Mondays at 11 (LT1) and Tuesdays at 3 (LT1). In the first two weeks there will be a third lecture on Thursdays at 2 (LT1).
- **Problems classes** are in ODD teaching weeks starting in teaching week 3. These take place on Thursdays at 2 in LT1.
- **Drop–in sessions** are at the same time and place in EVEN weeks starting in teaching week 4.
- **Computer practicals**: Fridays at 1 (Herschel full PC cluster). These happen in *some* EVEN weeks, and you will be notified of these well in advance.

# Timetable and Administrative arrangements

*So to summarise:*

| What? | When? | | Where? | How often? |
|---|---|---|---|---|
| Lecture | Monday | 11–12 | Herschel: LT1 | Every week |
| | Tuesday | 3–4 | Herschel: LT1 | Every week |
| Problems class/ Drop–in session | Thursday | 2–3 | Herschel: LT1 | Every week has one or the other |
| Computer practical | Friday | 1–2 | Herschel: PC cluster | Some even weeks |
| Office hours | Monday | 12–1 | My office | Every week |
| | Wednesday | 12–1 | | |

# Assessment

Assessment is by:

- End of semester exam in May/June (**85%**)
- In course assessment, including written solution to problems and computer practical work (**5%**)
- Mid-semester test (**10%**)
- No CBAs!

There will be two assignments, each one having a mixture of written and practical questions.

Solutions to the assignment questions should be submitted by the dates given in your notes.

We will work through the unassessed questions in problems classes and practicals.

# Recommended textbooks

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

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

# Other stuff

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

# MAS2602: Feedback on your feedback

- **Module overall**
  - Structure: 4.1/5
  - Intellectual stimulation: 4.1/5
  - Feedback: 4.5/5
  - Satisfaction: 4.0/5
- **Teacher**
  - Stimulated me to learn: 4.0/5 (Chris: 4.6/5)
  - Approachable: 4.4/5 (Chris: 4.6/5)

# MAS2602: Feedback on your feedback

- **Open-ended comments**
  - Both Lee and Chris were engaging and approachable. I enjoyed the module and I think that was more to do with the lecturers than the content :-)
  - Lee and Chris were brilliant, I hated computing going into this year but these two have really changed my opinion on it :-)
  - Enthusiasm of both lecturers was great :-)
  - Chris: Absolute lad, to bo honest
  - Chris: Top geez as always
  - Chris: Passionate, fun guy
  - Chris: Infectious enthusiasm for programming

# MAS2602: Feedback on your feedback

- **Open-ended comments**
  - Lee: Lectures felt rushed at times :-(
  - Lee: Speak more slowly and less Jordy ??
  - More Chris, less Lee :-(
  - Lee: More computing, less Stats
  - Lee: Polish your R-side so it matches Chris's side
  - Lee: Need more R prep/revision at the start
  - Lee: Wired guy
  - Chris > Lee, tbh

# 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!

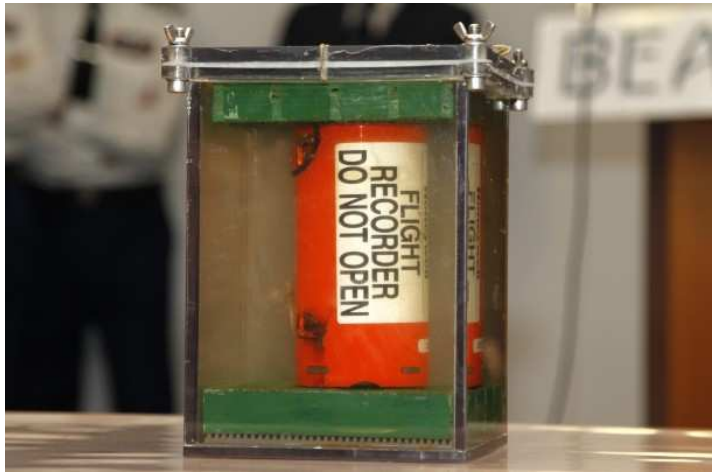
- The work of Thomas Bayes (see later) was published in 1764, 3 years after his death
- Over the course of the next 100–150 years, it received little attention
- In fact, some key figures in Statistics – e.g. **R.A. Fisher** – outrightly rejected the idea of Bayesian statistics
- By the start of WW2, 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

| 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.

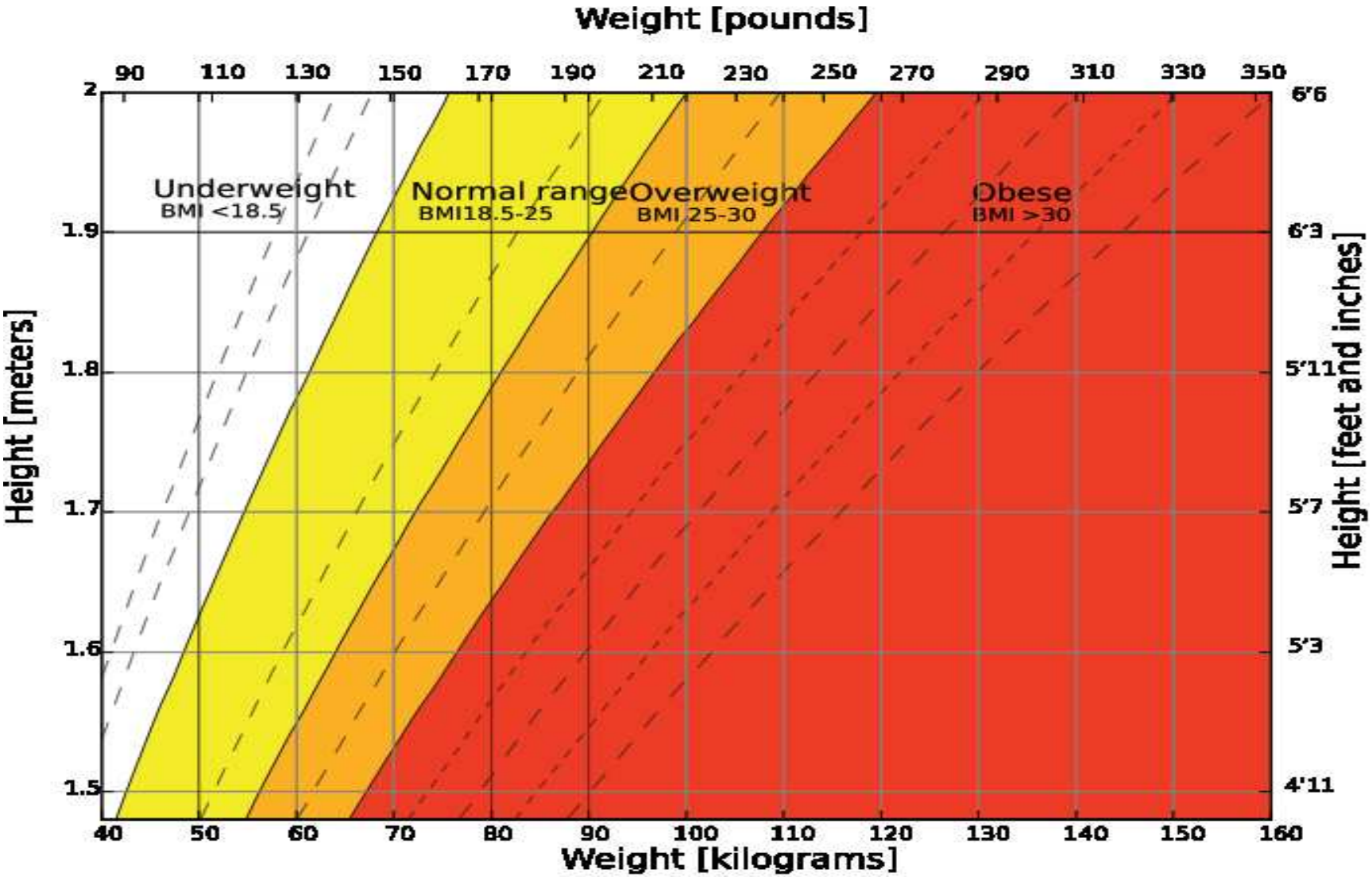
BMI is often used as a measure of ‘fatness’ or ‘thinness’: it is the ratio of a person’s weight (measured in kg) to their squared height (measured in metres).

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** $H_0$!

What are we to do? Convention tells us to work at the 5% level, but why? Some practitioners work at the 1% level!

# 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 $\mu_M$) are NOT random variables!

This means $\mu_M$ 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 MAS2903 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.*

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-

# 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 web–page for more details.

# Quick quiz

Suppose $A$, $\theta$ and $\mu_M$ 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. $\theta = 1/2$.

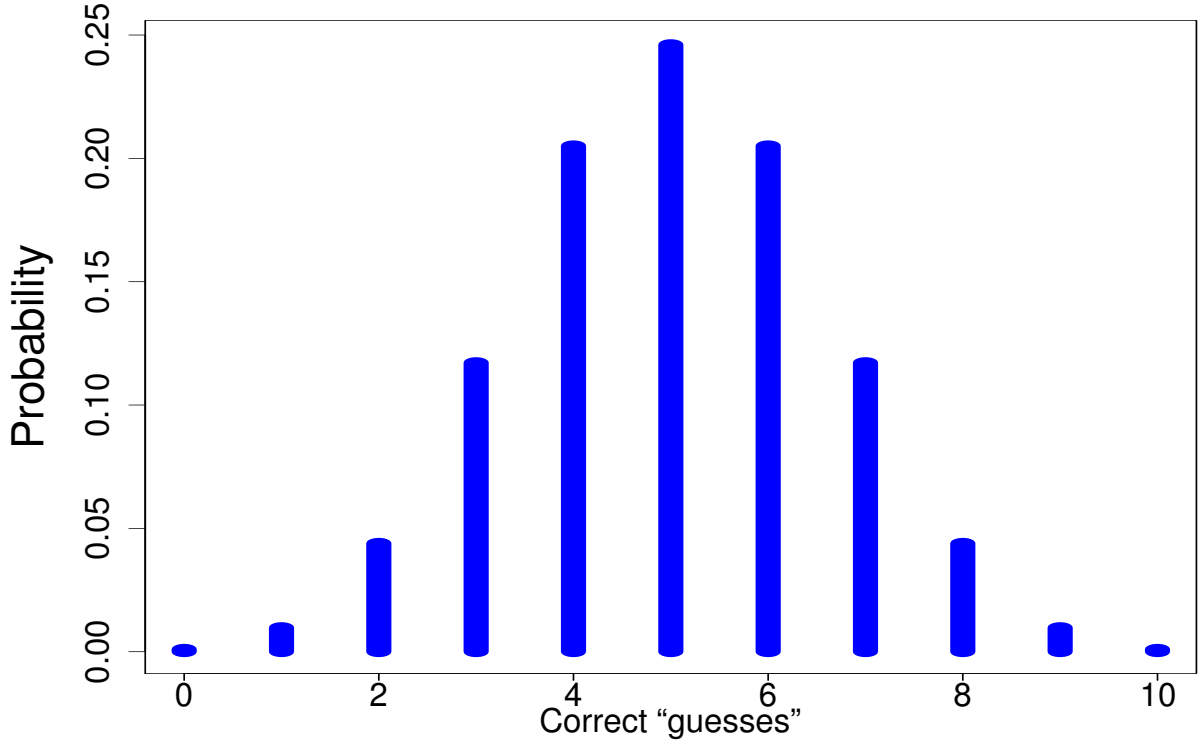
# Introduction

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:



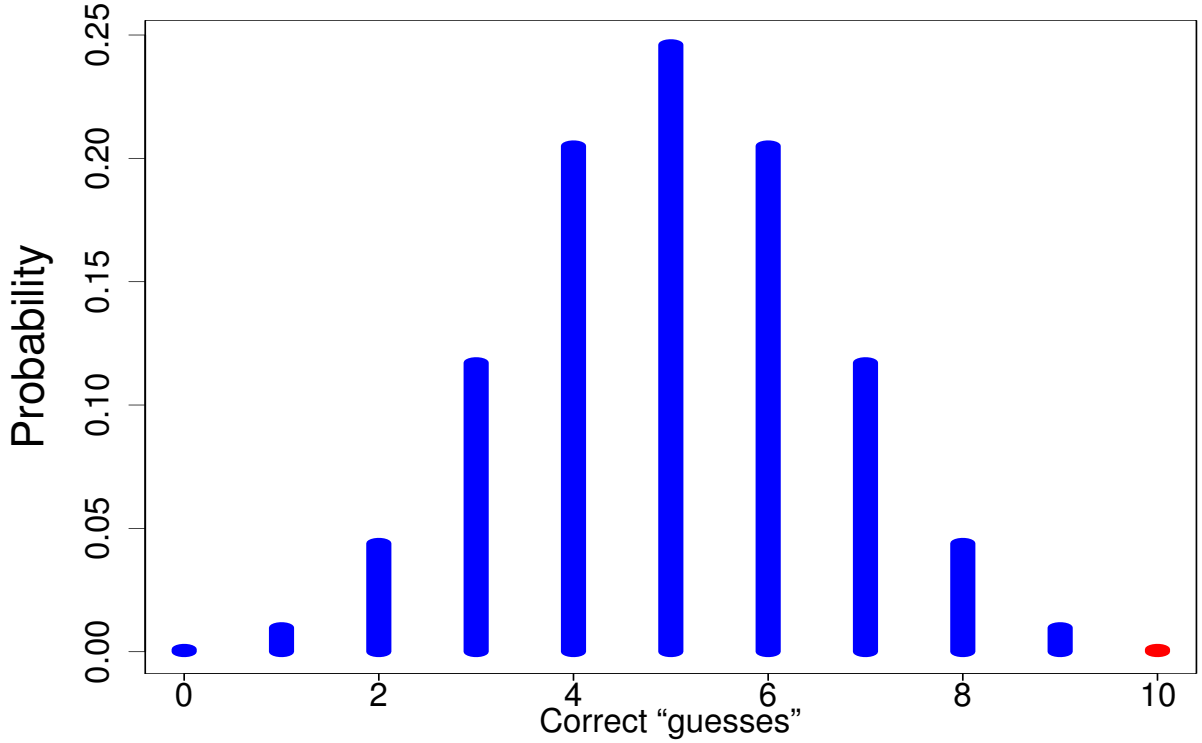
# Introduction

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 $\theta$ are **plausible** for each experiment?

# Introduction

Prior to each experiment, our beliefs about $\theta$ 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

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

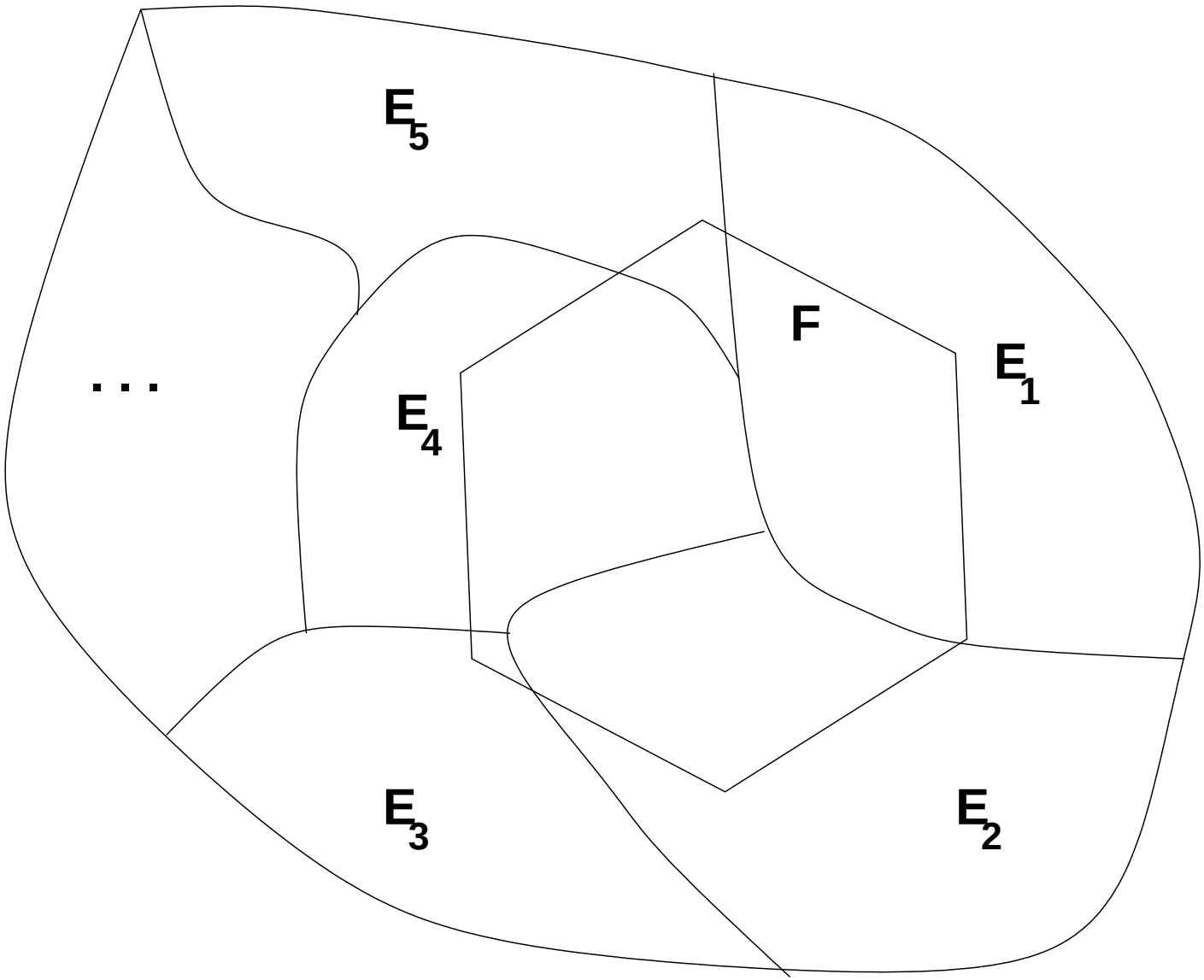
# Bayes' Theorem

## Definition (Parition)

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

**Fact (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 ($\theta_1$) and the other due to a fault in the gearbox ($\theta_2$).

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 $x_3$ |
|---|---|---|---|
| $\theta_1$: fault in engine | 0.1 | 0.4 | 0.5 |
| $\theta_2$: 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 $\theta_1$ and $\theta_2$ 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 $x_3$ |
|---|---|---|---|
| $\theta_1$: fault in engine | 0.118 | 0.471 | 0.625 |
| $\theta_2$: 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 ($\theta_1$) 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 ($\theta$).

Consider first that we observe overheating only ($x_1$). The posterior probabilities are in favour of declaring the fault as in the gearbox ($\theta_2$) 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 ($\theta_2$) |
| irregular traction only ($x_2$) | fault in gearbox ($\theta_2$) |
| both symptoms ($x_3$) | fault in engine ($\theta_1$) |

# 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 $\theta$) 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

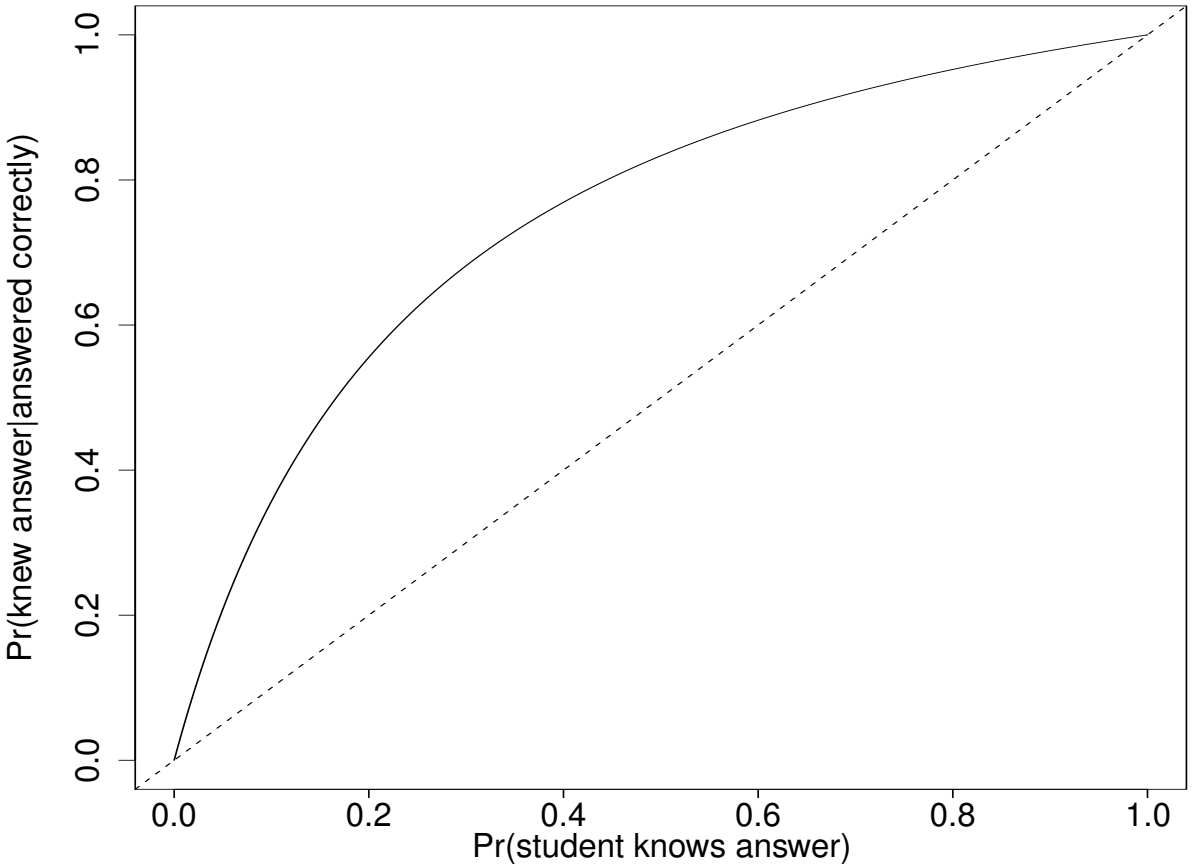
$$\Pr(K|C) = \frac{5\theta}{1 + 4\theta}.$$

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 $\theta$ – 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 $\theta$.

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

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 $\theta$ 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 $\theta$.

# 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 $\theta$.

# 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 $(\mu, \sigma)$.

# 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

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

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

Further, in $f(\boldsymbol{x}|\theta)$, $\theta$ 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 $\theta$ 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 $\theta$.

# 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 $\theta$ if $f(\boldsymbol{x}|T(\boldsymbol{X}) = t)$ does not depend on $\theta$.

# 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 $\theta$. 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 $\theta$. Hence, by Definition 1.4, $T = X_1 + X_2$ is sufficient for $\theta$.

# 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 $\theta$ if $f(\boldsymbol{x}|\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t})$ does not depend on $\theta$.

## 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 $\theta$.

# 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 $\theta$.

# 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 $\theta$.

# 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 $\theta$.

# 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 $(\mu, \sigma)$.

# 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 $\mu$ and $\sigma$.

# 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 $\theta$.

# 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 $\theta$.