The statistics reference book for judges

Harry Gray

2020-10-07

Contents

Preface

This book by the Leverhulme Research Centre for Forensic Science provides a learning and reference resource for judges on the topic of evaluating forensic science evidence using probability and statistics. It aims to provide the stepping stones required to become comfortable with likelihood ratio-based evaluations of evidence, as endorsed by the European Network of Forensic Science Institutes, and the Forensic Science Regulator of the UK.

Why this book

High quality resources can already be found on this topic, which have been authored by the founders of systematic approaches to forensic evidence evaluation and which have been endorsed by national-level organisations. These include **?**, **?**, **?**, **?**, **?**, and **?**. This resource aims to complement and extend the existing texts in the following ways:

1. by providing more examples for each of the core concepts so that learners can see the theory in action
2. by providing worked exercises and solutions for learners to test and reinforce their understanding
3. by providing interactive material (in the online version of this resource) to enhance the learning experience
4. by being based online to allow feedback to be easily given and regularly actioned upon, and for content updates to be easily made as scientific thinking develops and current practices change - see the sections below on how to contribute.

This book follows the main content from the Royal Statistical Society's Fundamentals of Probability and Statistical Evidence in Criminal Proceedings book (**?**) and then portions of The Logic of Forensic Proof (**?**). We have also provided motivation for the use of probability and statistics by drawing from the

uncertainty communication literature, and in particular adapting the advocated model from **?** to forensic science. Put together, this resource provides an intermediate overview of why and how we use probability and statistics in evaluating forensic evidence, including examples and exercises to reinforce understanding.

This book can be used to supplement the online course content from [cite MOOC]. This way, learners can progress through more formalised training (with video, audio, and graded tests), and simply use this book as a resource to see further examples. The order and presentation of the material is different to what appears in the course, although most of the same ideas appear here.

The interactive examples in the online version of this book can be found as their own tool at [cite shiny app]. This is to provide a standalone resource for those who are not using the book e.g. a wider audience, and also so that those who prefer to use a printed version of this book can still have fast access to the examples they need.

The content presented here is not a comprehensive guide to the evaluation of all forensic evidence. For example in the UK (as of writing this in August 2020), the numerical evaluation of impression evidence (such as fingerprints) is not widely practised compared to say, how it is done in The Netherlands. In addition, each evidence type is currently at a different stage of scientific understanding. As a result, the exact statistical considerations for each evidence type differ, even if the general evaluation methodology is the same or similar. More information on the state of each type of evidence can be found in the judicial Primers series commissioned by the Royal Society and the Royal Society of Edinburgh (**???**).

Who this is for

This book is primarily aimed at judges, but in the UK that covers at least two distinct domains: criminal and civil. The skills required to understand the statistical evidence in both of these domains is sufficiently different to be given separate educational treatments. This book (as of the current version) is dominated by statistical content to help with forensic evidence in criminal cases. As time progresses and we receive feedback, and in particular if there is demand for it, we may add content to aid understanding of using the Ogden tables for calculation of damages due to injury, and also to help with understanding epidemiological evidence, e.g. for causation of illness.

There is a wider audience who might also benefit from this material. In a professional capacity, this includes legal counsel, students of law, and students of forensic science who are relatively new to statistics and its application in evaluating evidence. Our hope is that this can be an engaging resource to supplement professional development and formal study.

Outside of that, we hope that this resource might be an interesting introduction for members of the public who are curious about how forensic scientists

determine the value of scientific evidence and present that to the court.

We welcome feedback from anyone, which can be used to improve the content of this book and for the development of other resources too, see the notes on how to contribute and and how to contact us in the corresponding sections below.

How to contribute

One of the purposes of this book is to be reactive to feedback and be easy to contribute to. We strongly encourage feedback, be it negative or positive, so that we can change what is not working and keep what is working. There are two ways someone can contribute, detailed below.

Github

If you don't know what 'Github' is, then skip this section and go to the Email section instead.

All of the code to generate this book (and its content) can be found here [insert repo link] on Github.

Please use the Issue function to raise any problems or suggested content changes. If your change is minor and you can do it quickly yourself (e.g. a spelling correction, a minor rewording, etc.) then feel free to make a Pull request.

Any further correspondence can be addressed below.

Email

Feel free to email h.w.gray@dundee.ac.uk with your feedback on the book. When making a contribution, please be sure to reference specific sections of the book so that I can make a change. If you have disagreements with the content then feel free to raise these, but please be respectful and provide references so that I can best engage with your feedback. Remember to mention your name and affiliation so that I can add you to the list of contributors below.

Getting in contact

Please do get in contact with h.w.gray@dundee.ac.uk with any thoughts and suggestions about this book and future content. We particularly welcome thoughts of members of the UK judiciary as we work towards improving this resource and providing further resources.

Acknowledgements

Funders

The primary author of this book was funded by the David and Claudia Harding foundation on grant [enter grant number], and is part of a department which is funded by the Leverhulme Trust. Our thanks goes to these funding bodies, who have made this non-traditional academic output possible. It is our intention that this resource will contribute to raising statistical literacy within forensic science and will have done so by embracing emerging technology and positively disrupting the traditional way we engage others with statistical ideas.

Contributors

Roy Mudie at LRCFS has developed the R Shiny application which accompanies this book. This allows the html version of this book to contain interactive examples which we hope (and soon will test to see!) enhance the users' learning experience. The Shiny application can also be used as a standalone tool to engage with our statistical content. Roy has helped with many of the technical issues which have arisen behind the scenes and helped with how to visualise some of our content in a way that was intuitive. Thanks a lot for all of your work, Roy!

Software

This book was written using the **bookdown** package (**?**), which was built on top of R Markdown and **knitr** (**?**). This fantastic free and well-documented software has allowed this book to be developed hassle-free in a completely reproducible way. Thank you Yihui for your work to support open software and open science.

Updates and versions

About this version (1st draft)

The book is currently in first draft stage. This means that the content has finished its first review and has received minor edits from an internal review team.

Planned versions

Major versions in planning:

- Second draft, after more internal feedback
- Version 1, after feedback from internal students and some external practioners
- Version 2, after wider external feedback

Update history

04/2020 - Work on the book begins

06/2020 - Application for interactive examples created

08/2020 - First draft of the book

Future content

- Interactive exercises with feedback for each chapter
- Case Assessment and Interpretation protocol basics
- A Bayesian network primer

Chapter 1

Introduction

What is the value of scientific evidence after taking into account assertions made by the prosecution and defence? That is the question which underlies the material in this book.

The answer depends upon the type of scientific evidence in question,

Strides were made in this direction with the Case Assessment and Interpretation (CAI) protocol… [brief history of development of forensic science and evidence evaluation]

[make the point that as scientific technology and knowledge sharing both increase, we are able to quantify more and move away from subjective opinion]

[as technology continues to develop, so will quantitative analysis methods and this will make their appearance in the courtroom more commonplace, e.g. current scientific research is moving towards machine learning algorithms for pattern recognition, and whilst this might not have filtered through to standard forensic analysis pipelines yet, it is coming]

[there's a need to raise everyone's statistical literacy to keep the technology understandable and accountable to public demand; this is moreso for judges who gatekeep evidence and administer justice]

[this book aims to introduce key concepts in probability and statistics which relate to evidence evaluation, providing interactive examples and being responsive to feedback to meet the needs of its users]

[chain of investigation (some CAI); roles within the courtroom]

This book is organised as follows.

In Chapter **??** we set the basis for using probability and statistics for scientific evidence: uncertainty.

In Chapter **??** we show how the concept of probability can be used to quantify the uncertainty about events occurring.

In Chapter **??** we introduce statistics as a way of describing empirical data and using that to make inferences about uncertainty.

In Chapter **??** we describe formal statements about events, known as propositions. This allows us to begin to describe asserted events in criminal cases, which are essential when evaluating evidence.

Then, finally in Chapter **??** we tie together the ideas from the previous Chapters in order to arrive at quantifying the value of scientific evidence using the likelihood ratio.

Chapter 2

Uncertainty

In 2008, Kennedy Brewer was exonerated after 15 years in prison for the brutal murder of his then-girlfriend's 3 year old child. Expert testimony had claimed that marks found on the body of the child had been bite marks and had without a doubt been caused by Brewer. The expert's belief did not accurately reflect the uncertainty in the scientific understanding of bite mark evidence.

Becoming familiar with uncertainty is the basis for thinking about probability and statistics, and ultimately evaluating scientific evidence. In this chapter, we will define what is meant by uncertainty, how different types of uncertainty can be characterised, give an example of uncertainty in scientific evidence, and finally discuss the communication of uncertainty.

2.1 What we mean by *uncertainty*

In this book, we frame discussions about uncertainty around beliefs, since beliefs about facts are the focus of forensic science. Certainty describes a belief which can be guaranteed without any doubt. If there is any doubt in the belief then it cannot be called certain. We call those beliefs uncertain. Beliefs can have varying degrees of doubt and so uncertainty can be described on a spectrum with one end containing beliefs with high degrees of uncertainty and the other containing beliefs with low degrees of uncertainty. Many beliefs come with at least some degree of doubt, however small it may be, and so must be uncertain. This makes uncertainty unavoidable.

Communicating uncertainty is key to forming and calibrating realistic expectations. When uncertainty is communicated well and everyone has realistic expectations then it can be managed effectively in order to mitigate negative consequences. However, in practice there are understandable incentives against doing this. One example is that people in positions of authority do not want

to diminish the trust of those they have a responsibility towards by admitting that they do not know something, such as politicians. Another is when the uncertainty might cause a disproportionately negative reaction, e.g. the slight uncertainty that humanity will be erased in a nuclear war this year. This means that unfortunately, in many aspects of life, uncertainty is simply ignored or concealed.

One classical uncertainty has been thrust into the public domain due to the COVID-19 pandemic during 2020: medical diagnostic testing. Results from medical tests which aim to diagnose disease are, generally speaking, uncertain. Tests can return positive results for people who do not have the disease in question, known as false positive results, and tests can also return negative for those who do have the disease in question, known as false negative results. This means that for any group of people who are all tested, the number of positive tests may not accurately reflect the number of people who have the disease. Even highly reliable tests will result in many false results provided that enough people are tested. This can be a particular problem for diseases which are not well understood and this is emphasised when using tests which are new and whose reliability have not yet been fully validated. The resulting situation of a new disease with new tests is highly uncertain.

In society, there are many situations in which we can communicate and manage uncertainty better. This will come with building a better understanding of uncertainty and having honest and open discussions about it. We aim to contribute to that in forensic science, by framing scientific evidence evaluation around the fact that it is uncertain. In order to do this, it is useful to be a bit more specific about different types of uncertainty.

2.2 Types of uncertainty

Uncertainty in its broad definition from the previous section covers a spectrum of situations. There can be clear logical differences between uncertain beliefs based upon the type of uncertainty that is described. We focus on two types of uncertainty

1. **Aleatory uncertainty**: uncertainty due to variation.
2. **Epistemic uncertainty**: uncertainty due to a lack of knowledge.

Aleatory uncertainty describes situations with natural variability, sometimes referred to as **chance**. With this type of uncertainty, we cannot be certain about something because there is a component of randomness to it. For example, before a roulette wheel is spun, you are asked what number the ball will land on. Assuming the spin is not fixed to favour any particular number, then the result will be random in the sense that there is no strategy of guessing which will allow you to win more often than selecting a number at random. Due to

the randomness involved in the spin, your guess will be uncertain. This type of uncertainty cannot be reduced because it is inherent to the process that causes it.

Epistemic uncertainty concerns situations about which we have a lack of knowledge. The reason for this type of uncertainty is that we do not know all of the necessary information. Suppose I ask you to turn your back so that you do not see the result of the roulette spin. I spin the wheel and see the result, and then I ask you to guess the number. Your guess is now uncertain not because of randomness, but because you do not have the information that I have. There is no uncertainty for me but there is for you, and your uncertainty could be eliminated by learning what I know. Epistemic uncertainty can be reduced by obtaining more information.

The idea of uncertainty being personal is key to thinking about events which have happened in the past. Past events have either occurred or not, and so they are factually certain. However, an individual's beliefs about those past events can be epistemically uncertain because they have insufficient information. Individual's will base their beliefs on the information that they have available to them and, since this information and how it is valued could vary from person to person, it is possible to have varying degrees of epistemic uncertainty between individuals about the same factually certain event. Imagine that another player is introduced into the roulette example above. I ask you both to turn your back whilst I spin the wheel and I see that the result is a 10. I then tell you nothing about the result but I tell the new player that the result is an even number. You still have the same uncertainty as before but the new player has less uncertainty than you because they have more knowledge about the result; you both have different degrees of epistemic uncertainty about the same factual event.

Epistemic uncertainty is reflected in the courts through the phrase 'beyond reasonable doubt', the burden of proof for the prosecution in UK criminal trials. It is not 'without any doubt'; there is room for some uncertainty about the facts. Events have factually occurred but perhaps only the aggrieved and/or the defendant are certain about the truth. On the other hand, the fact finder begins a trial with a high degree of uncertainty about events with the presumption of innocence on behalf of the defendant. Evidence about the facts is presented to the fact finder in an attempt to reduce their epistemic uncertainty. Upon hearing all of the evidence, the fact finder makes a decision as to whether the prosecution sufficiently reduced their uncertainty to 'beyond reasonable doubt' in favour of guilt. When the fact finder is a jury, then each juror must have their personal epistemic uncertainty reduced to this level during the initial stage of deliberations.

Uncertainty is present in all parts of the legal system as it is in every other aspect of society. Now that we have discussed the types of uncertainty, we can be more specific about how it manifests in scientific evidence.

2.3 Uncertainty in scientific evidence

There are multiple sources of aleatory and epistemic uncertainty in forensic scientific evidence. As well as the unique uncertainties that arise from individual case scenarios, there are generic uncertainties which relate to the collection, analysis, and interpretation of different scientific evidence types. Some of these generic uncertainties are shared across evidence types, for example the transfer and persistence of materials on surfaces, but some are also specific to a particular evidence type, such as the challenges involved in the analysis of low template or mixture DNA evidence. In this section, we will consider evidence in general and the collective uncertainties that can be present. Then we work through a small example using a specific instance of fibre evidence.

We will consider the following questions when thinking about uncertainty in forensic scientific evidence:

1. What are we uncertain about?
2. What are the sources of uncertainty?
3. What is the level of our uncertainty?
4. What is the magnitude of uncertainty?
5. In what form is the uncertainty communicated?

These questions follow the model of communicating uncertainty presented in **?**. We will apply this model to forensic evidence and address each of these questions in the order in which they appear above.

What are we uncertain about? This is the **object** of uncertainty. This is most often the sequence of events which led to the evidence. Nested within this, there can also be uncertainties about the characteristics of the evidence, the analysis of the evidence, comparisons of the evidence to information within databases, expert opinion concerning the evidence, etc.

What are the sources of uncertainty? These are the **reasons** for uncertainty. Imprecise measurements, contaminated evidence, lack of scientific knowledge, database limitations, subjective interpretation, etc.

These first two questions make it clear exactly what we are uncertain about and why. In an expert witness report, they can appear as the rationale behind what the expert has been tasked to do and why they have done it.

What is the level of our uncertainty? This is whether the uncertainty relates **directly** to the object itself or **indirectly** through uncertainties about things which affect the object. For example, uncertainty about an item of evidence is direct but lack of confidence in the science underlying the evidence interpretation is indirect.

This particular question is important because these two levels of uncertainty can be evaluated differently. Direct uncertainty is often easier to quantify than

indirect uncertainty, which will be explained in more detail in the next chapter. In practice, this means that in expert witness reports it is easier to treat direct uncertainties with an evaluative likelihood ratio approach when compared to indirect uncertainties. Indirect uncertainties, such as a lack of scientific knowledge for a particular piece of evidence, may be listed in the expert's report as a verbal caveat for the expert's evaluation or may have been part of a decision not to evaluate the evidence using a likelihood ratio approach.

What is the magnitude of uncertainty? This is **how** uncertain we are. Does the measurement technology have a high or low level of precision? Is the database highly reliable or not? Is the expert's interpretation more or less reasonable compared to other experts in their field? How well does the evidence support the prosecution?

In what form is the uncertainty communicated? What tools are used to **communicate** the uncertainty to others. Is uncertainty mentioned in the expert report? Is it clear that uncertainties have been addressed or are they only acknowledged? Has a number been used to convey the uncertainty, or a verbal expression?

The final two questions above relate to evaluating and communicating uncertainty. In expert witness reports, this is where the expert presents their evaluative opinion having considered the evidence alongside the case circumstances and assertions made by legal counsel. If they have used a likelihood ratio approach, then this evaluation will be given on a numerical or verbal scale (in the UK). For certain types of evidence such as fingerprint identification, the expert may give their opinion as the inclusion or exclusion of an individual.

Forensic science is still a developing field. Expert interpretations which historically overlooked (and in some cases ignored) major elements of uncertainty in their discipline have been and continue to be updated by practices which are informed by scientific evidence. However, there is still a long way to go in some fields, e.g. understanding transfer and persistence of fibres. There is still a large amount of epistemic uncertainty. Notwithstanding this, there will always be a degree of aleatoric uncertainty due to the stochastic nature of real-world processes. This means that there might be reasonable disagreement between experts. These disagreements can still be settled in the mind of the fact finder in the same way as disagreements in other aspects of evidence: through cross-examination.

Uncertainty is (and will always be) an integral part of science, and by extension scientific evidence. This means that it is important for us as a community to work together to ensure a common understanding of uncertainty, both in terms of how we think about it and how we talk about it. That is why there is more of a treatment of it in this book compared to other similar texts. Most academic research thus far has focussed on how we should think about uncertainty, but we are yet to establish consensus for talking about uncertainty.

In the next section, we will apply these questions to characterise the uncertain-

ties in a practical example of forensic evidence analysis.

2.4 Example: fibre evidence

Fibre evidence is a term used for forensic evidence related to textile fibres which are shed from textile materials (such as an item of clothing) as a result of contact with a surface. A fibre that is recovered from a surface or different textile material, known as the questioned sample, can then be compared to fibres which were taken from the suspected textile material of origin, known as the reference sample, and analysed for similarity. If the questioned sample does originate from the textile material, then there are two possible methods of the fibres having been transferred there: direct and indirect contact. Direct contact means that the fibres were transferred from the textile material as a result of direct contact with the surface from which the questioned sample was recovered. Indirect contact means that the fibres were first transferred to some other surface/s via direct contact and then transferred from this surface to the surface from which the questioned sample was obtained via another direct contact (making it indirect contact with the original textile material). This is a useful evidence type because it can help to distinguish between types of physical contact which may be alleged between people's clothing or people's clothing and other surfaces, which gives the fact finder information about a possible activity having taken place perhaps in a specific location.

In this example, we consider some generic uncertainties associated with fibre evidence. After a suspect has been identified and taken to trial, the most important uncertainty for the court to consider is whether or not the suspect committed the crime in question. This is epistemic uncertainty: the suspect either did or did not commit the crime but the fact finder does not know which is true prior to seeing evidence (although innocence is presumed). Depending on the case circumstances, the fibre evidence alone will likely not be sufficient to reduce the fact finder's uncertainty about whether the suspect committed the crime or not. However, it may be helpful in reducing their uncertainty about other related assertions, for example, whether or not the suspect was present at the crime scene. This could be useful for the fact finder to combine with other evidence.

Suppose that the expert is tasked with comparing fibres recovered from the suspect's clothing to fibres recovered from the crime scene. What are some uncertainties that the expert might consider?

The expert addresses uncertainty about the evidence. The fibre evidence is the **object** of uncertainty in this example. The **reasons** for uncertainty about this evidence might relate to two credible alternative events put forth by the prosecution and defence, this is epistemic uncertainty. There could also be aleatory uncertainty because the same textile material will shed variable numbers of fibres even for the same type of contact under the same amount of pressure. These are **direct** uncertainties about the fibre evidence.

Claims about these direct uncertainties will often be supported by other available information. This can include population surveys about the occurrence of textiles in a particular location, or scientific studies examining the shedability of fibres after controlled levels of contact. Uncertainties about the reliability of the scientific methodology underlying this supporting information, or its applicability to a particular case circumstance are **indirect**, e.g. doubts as to how applicable a fibre survey from Beijing, China is to a case in Dundee, Scotland.

The expert will synthesise their uncertainty either quantitatively using a mathematical model or qualitatively using their expertise (for which there is more uncertainty about the appropriateness of the mathematical model or the calibration of the expert's opinion). The result might be a number, a range of possible numbers, or qualitative verbal statements, all of which describe the **magnitude** of a particular uncertainty. For example, the expert might determine that out of every 1000 similar case circumstances they would expect 200 of them to yield the questioned sample of fibres if the suspect had truly been at the crime scene. The expert might also determine that out of every 1000 similar case circumstances they would expect only 2 to yield the questioned sample of fibres if the suspect had truly not been at the crime scene.

It is recommended practice that these magnitudes of uncertainty (when quantified) are combined into a single quantity, known as the likelihood ratio. This quantity is then **communicated** in expert reports. In this example, the expert has determined that it is 100 times more likely that the fibre evidence would have been observed if the suspect had been at the crime scene rather than if they had not. The fact finder can then use this, as well as other evidence, to make a decision as to whether they believe the suspect was at the crime scene. When the magnitude of uncertainty is not determined quantitatively, then verbal statements of uncertainty are communicated instead.

In the example above, we only focussed on one possible object of uncertainty. In reality, there are more objects that the expert has to consider and combine when interpreting evidence. The reasons for each of these objects of uncertainty will be case-dependent and often complicated by the complex nature of real-life criminal activities. This makes assessing uncertainty challenging. The expert might reasonably ignore some uncertainties to make this task more feasible. Many possible uncertainties can be excluded by only considering the sequence of events put forth by the prosecution and defence.

2.5 Communicating uncertainty

Accurately communicating uncertainty in forensic science is crucial in order for the fact finder to fully assess the evidence. This depends upon the extent to which the scientist is able to assess the uncertainty so that they can then communicate it. As we stated in the previous sections, there is limited scientific knowledge about the properties of some evidence types, e.g. how traces of

materials are shed and transferred through contact, and so this can limit the scientist's ability to assess the uncertainty.

There is general consensus in the UK that a likelihood ratio, range of likelihood ratios, or an equivalent verbal statement should be used to communicate direct uncertainties about the evidence. Examples include statements such as

- single number: the recovered number of indistinguishable fibres provide 50 times more support for the assertion that the questioned activity took place in the living room compared the bedroom.
- range of numbers: the recovered number of indistinguishable fibres provide between 10 to 100 times more support for the assertion that the questioned activity took place in the living room compared the bedroom.
- verbal qualifier: the recovered number of indistinguishable fibres provide moderate support for the assertion that the questioned activity took place in the living room compared the bedroom.

The resulting statement is always relative, in that it gives the fact finder information about the scientific evidence in light of the prosecution's assertions versus those of the defence.

Indirect uncertainties in forensic evidence are usually given as a subjective verbal statement, or a factual statement describing the strengths and weaknesses of the quality of the evidence. No systematic approach is currently used for assessing and communicating indirect uncertainties in forensic science. In other fields, a number of systematic approaches have been developed to make this subjective judgement more consistent and transparent. One way to achieve this is by defining a checklist of characteristics which determine a category for the quality of evidence based upon how many items on the checklist the evidence meets. For example, the GRADE scale has s routinely used to assess the effects of medical interventions. It has categories which range from 'very low quality' to 'high quality' depending on characteristics of the evidence. This may be a blueprint for forensic science to follow to systematically communicate indirect uncertainties in forensic evidence.

Since assessing uncertainty plays an important role in evaluating forensic scientific evidence, it is important that there is common ground when it comes to communicating it. This can be achieved through using a standardised framework for assessing uncertainty that is logically coherent. One example of coherence in this context can be making sure that inferences made using uncertainties are consistent and obey transparent logical rules. This means that inferences made within the standardised framework are logical, can be accurately described to others, and are able to be assessed by others - all of which are important for communication. These desirable properties (as well as others) are contained within the mathematical language of probability, and that is why it is used

as a common means to handle uncertainty. The main utility of probability is not because it quantifies uncertainty, but because it is a complete and common framework for logical inference. Probability is the focus of the next chapter.

2.6 Try it yourself

2.7 End of chapter survey

2.8 More information

GRADE: https://www.gradeworkinggroup.org/

Chapter 3

Probability to describe uncertainty

We saw in the previous chapter that there is uncertainty in scientific evidence. By asking a few simple questions about the uncertainties in a particular context, we demonstrated a systematic approach to assessing the uncertainty which can easily be applied to forensic evidence. As part of this process, the expert assesses the magnitude of their personal uncertainty given the scientific expertise and experience that they have. An example we gave of this was when the uncertainty was quantified as follows: out of every 1000 similar case circumstances, the expert believes 200 would yield a fibre match if the suspect had truly been at the crime scene. This is useful because it has described exactly how uncertain the expert is about the evidence in what is likely to be the prosecution's version of events. In this example, the uncertainty has been converted into a probability, which has then been converted into a so-called natural frequency for the purpose of communication. In this chapter we discuss probability, describing what it is and using examples to demonstrate some of its useful properties as a framework for handling uncertainty.

3.1 Quantifying uncertainty

Quantifying uncertainty gives us a systematic way of assimilating and comparing uncertainties. This means that different personal uncertainties for the same object can be assessed consistently and also that personal uncertainties for different objects can be compared. This is because the framework of mathematics forces quantities to obey a coherent and consistent set of logical rules. The subset of mathematics which handles uncertainty is known as probability. The main benefit of using probability is the framework of logic that it enforces, rather than the quantification of uncertainty (although this is useful).

A probability is a number between 0 and 1 that describes the magnitude of uncertainty for the occurrence of an event. The probability must obey certain rules which we will show in subsequent examples in this chapter. A probability of 0 means that the event is impossible whilst a probability of 1 means that an event is certain. Uncertainty is described by probabilities which fall between 0 and 1. Probabilities of 0.5 describe an event whose occurrence is exactly as likely as its non-occurrence. Events whose occurrence is less likely than not should have a probability less than 0.5 on the scale, whilst events whose occurrence is more likely than not should have a probability greater than 0.5 on the scale. How close these probabilities are to 0 and 1 should reflect the magnitude of uncertainty in their occurrence or non-occurrence. Since we made the argument that uncertainty is personal because it depends upon an individual's beliefs, it follows that probabilities are personal too. In forensic science, personal probabilities are generally interpreted as an individual's degree of belief in the occurrence of an event. Not everyone who is familiar with probability theory agrees with this interpretation, but this historical debate is not important for the purposes of this book.

Quantifying direct versus indirect uncertainty is conceptually different. For example, if an expert is examining fibre evidence and is uncertain about the background prevalence of a particular colour of wool textiles in the local area, then they could consult a local population textile survey to help them determine this. The population survey could give information about the proportion of woollen garments worn in the local area, as well as the proportion of those which are the colour of interest. The expert can use these proportions to quantify their direct uncertainty about the background prevalence of that colour of wool fibre. However, if the expert is uncertain about the reliability of the methodology employed by the fibre survey, then quantifying this information is more challenging. That is why indirect uncertainties are often communicated using verbal expressions of evidence quality. We only consider quantifying direct uncertainties here, since indirect uncertainties about forensic evidence are usually verbally qualified in court.

Constructing probabilities to describe uncertainty is often done by assuming a probabilistic **model** for how the uncertainty is expected to behave in reality. Since the process that is being modelled is uncertain, the expectations might not be exactly what is observed in practice. This gives rise to the phrase "all models are wrong, but some are useful". The most useful models can accurately align an individual's magnitude of uncertainty to a quantitative probability, accepting that this can never be done perfectly.

3.2 Example: coin toss

The classic example for demonstrating probability is tossing a coin. The outcome of a coin toss is uncertain in most cases. We can consider some of the

questions from the previous section to describe this uncertainty.

What uncertainties are there? The uncertainty is whether the coin will land heads-up or tails-up as a result of the toss.

What are the sources of this uncertainty? There is a randomness to the flipping process. We also do not know if the coin is double-sided, e.g. has two heads instead of one heads and one tails. You may consider other sources here too, such as trust in the person flipping the coin to be acting fairly.

The potential double-sidedness of the coin represents epistemic uncertainty. We can eliminate this uncertainty by checking both sides of the coin before it is tossed.

The randomness of the flipping process represents a combination of aleatoric and epistemic uncertainty. The epistemic uncertainty comes from trust in the person flipping the coin and other factors which might unfairly influence the outcome of the coin flip. It is possible to mostly eliminate this uncertainty by learning about the person flipping it and overcoming them, e.g. by letting someone trustworthy flip the coin.

Assume that it has been checked that the coin has one heads and one tails, and that the person flipping the coin is acting fairly. There is now only aleatoric uncertainty about the outcome. This is an irreducible uncertainty of the coin flip; no further information can be learned which will make a guess of the outcome better than randomly guessing heads or tails.

What is the level of this uncertainty? This is direct uncertainty about the coin toss.

What is the magnitude of uncertainty? A probability of 1 means that an event is certain. Since the coin has one head and one tails, the coin toss must either result in heads or tails. This means that the event 'heads or tails' occurring must have a probability of 1. It also means that the event 'heads and tails' has probability 0; it is impossible to get both in a single toss. The logic of probability has dictated that any personal probabilities must agree with the above conditions in order to be coherent.

The event 'heads or tails', which has probability 1, is made up of two other events: the event 'heads' and the event 'tails'. In other words, the distinct outcomes 'heads' or 'tails' combine into the single outcome 'heads or tails'. These two distinct outcomes are known as **mutually exclusive**, since they cannot occur together. When outcomes are mutually exclusive, then their probabilities can be added together - this is one of the laws of probability. Without having yet quantified any uncertainty, the logic of probability states that the sum of the probabilities for heads and tails must equal 1.

Knowing that the probabilities of these two outcomes add up to 1 is helpful when assigning a personal probability to them individually. This is because we can consider their probabilities in relation to each other. Is getting a 'heads'

more or less likely than getting a 'tails'? We assumed that the person tossing the coin was doing so fairly, and so a rational belief with this information is that heads and tails are equally likely. If they are equally likely, then they must each have a probability equal to 0.5. Notice how this quantified uncertainty is the result of the logic of probability; the logic was the most important step because it constrained our beliefs about the coin to be coherent, then the quantification was secondary.

Now that we have beliefs for the outcome of the coin toss, we now have a probabilistic model. Checking the outcomes of this model can be a good way to confirm that it does align with our uncertainty. One way to do this, and also to communicate probabilities, is to frame them in terms of expected frequencies. This means assuming that some number of outcomes have been observed and to project what the model expects to occur for those outcomes. If we hold the belief that each probability is 0.5, then we expect that the outcomes should be evenly distributed between heads and tails.

3.2.1 Expected frequency tree

Below is an example of an expected frequency tree. The expected frequency tree displays the expected number of heads and tails corresponding to the probabilities of heads and tails for a fixed number of coin tosses.

A probability of heads of 0.5 means that we expect 5000 or every 10,000 tosses to result in heads, and the other 5000 to result in tails.

To view all of the interactive examples in this book, please visit Interactive Stats Book (http://127.0.0.1:5555).

The expected frequency tree in the example above shows what the probability model expects to occur for a specified probability of heads. It does not necessarily show what will be observed in practice because there is an unavoidable uncertainty about the outcomes (aleatory uncertainty). Tossing the coin, observing the outcomes, and then using that to refine the probability model is the topic of the next chapter.

As a final comment, notice how if we did not ignore some of the epistemic uncertainties about the coin toss (e.g. if we did not trust the coin flipper to be fair), then the probability model would need to be adjusted to align with this belief, e.g. by changing the probability of heads from 0.5.

3.3 Personal probabilities

We mentioned in the previous section that probability was personal because individuals have different knowledge and beliefs. In the previous example someone might believe that the person (or computer) tossing the coin was doing so

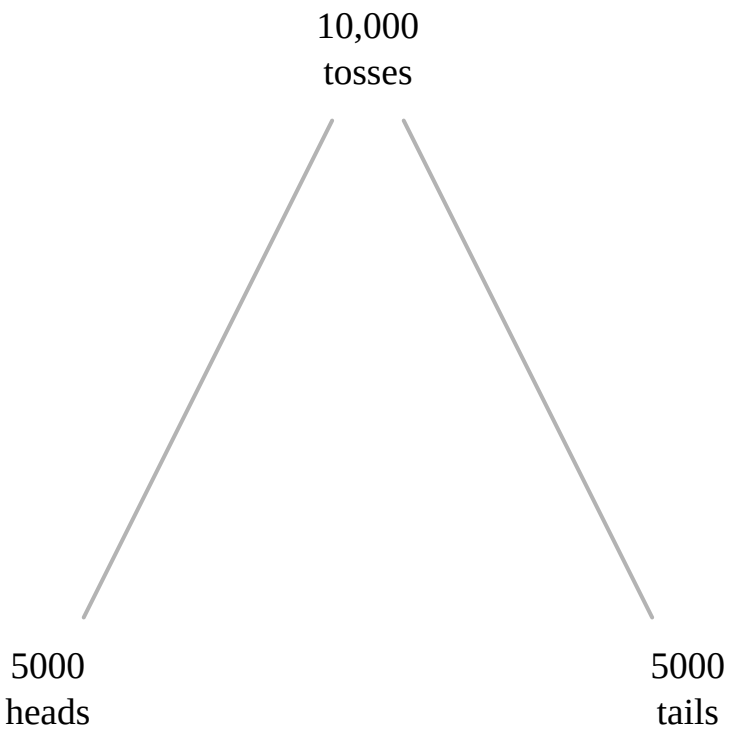


Figure 3.1: Out of 10,000 cosses, we expect 5000 to be heads and 5000 to be tails. The probability of heads is 0.5 and is equal to the probability of tails.

unfairly, e.g. that the flip was in favour of heads or tails. Their probability for heads should then be different to 0.5.

Some differences in beliefs might be more or less reasonable since they might be based upon more or less reliable information. Probability assignments based upon more reliable and generally agreeable information will be more **objective**. For example, the scientific evidence underlying the processing and analysis of full DNA profiles for single donors is well established and so experts are generally confident and in agreement about assigning match probabilities in this circumstance.

Probability assignments which are based upon less reliable or agreeable information will more be more **subjective**. For example, probabilities formed about the transfer and persistence of DNA on surfaces. Subjective probabilities are not necessarily bad because they can still be the best assessment of the available knowledge. This might happen for example when there is very limited published scientific literature about a particular scenario, but when the expert has case experience in that scenario. In other words, subjective probability assignments are not necessarily (and in criminal investigations should never be) arbitrary.

For repeatable events like the coin toss we can check and update our probabilities using empirical data, e.g. repeated flips of the coin. This can be seen as eliminating epistemic uncertainty about the biasedness of the flipper or gaining a better understanding of the aleatory uncertainty. With enough repetition, personal probabilities which were initially different between individuals can converge. Repeating the process in question makes personal probabilities more easily assessable, as beliefs which are coherent with or contradictory to available information become clearer.

In criminal cases, the opportunity to repeat events and gather scientific data from the case circumstances can limited (and repetition might not even be desirable for harmful processes). The events are uncertain and one-off, and so probabilities cannot usually be assessed by empirical repetition alone. They can still be informed by other empirical data however, e.g. be validated repetitions in similar (but not the same) circumstances.

3.4 Conditioning on information

Information which is used to construct a probability is called **conditioning** information. This is because we are constructing a probability which is conditional on that information, otherwise known as a **conditional probability**. The process of using information this way to construct such a conditional probability is itself called **conditioning**.

This means that probability is conditional on our current state of individual knowledge. However, to avoid being overly tedious we usually omit to say this each time we talk about a specific probability, e.g. we do not say the probability

of heads followed by listing all of an individual's assumptions and conditioning information each time it is mentioned. Instead, it is conventional to list initial assumptions when the probability is first mentioned. This listed information is known as **background information** since after it is listed it is assumed to be part of the background of any belief that is expressed. For example, for the coin toss we stated that the flipper was tossing the coin fairly. This assumption was a key part of the background information that logically led to a 0.5 probability of heads. This probability is not as rational for individuals who did not assume this same background information. Background information is not always made clear but it is always a component of probability assignments.

Most explicit applications of conditional probability are not focussed on background information, but on other extra information which is logically separate from and usually more important than the background information. The distinction between these two types of information is context-dependent as background information in some situations may be considered the important information in other situations. In any case, the important information is usually made clear when probabilities are expressed whilst the background information is usually contained within the assumptions underlying the probabilities. Any conditioning information is referred to as 'relevant' to a specific event if conditioning on it changes an individual's probability of that event occurring.

A very useful application of conditional probability is to condition on possible future information/events to see how this could affect probabilities of interest. For example, banks who lend money will need to consider the probability that loanees can be pay the money back (with interest). Possible future events which might affect that probability include an individual's employment situation. If the employment is potentially unstable and the individual's probability of repayment declines greatly if they lose their job, then the bank will be less likely to lend to them compared to if the individual's probability of repayment was not affected by their employment. Considering possible future events helps to make decisions in the present.

In the same way, one can condition on possible events in the past. This is particularly useful when there is epistemic uncertainty about what has happened in the past. With this type of uncertainty about the past, there will often be observed outcomes which could have been caused by multiple possible events. In such cases, one can consider the probability of the observed outcome conditioned in turn on each of the candidate causal events. These conditional probabilities can then be compared to determine which of the causal events most supports the observed outcome. This is how conditional probabilities are used for interpreting forensic evidence: we look at the events which the defence and prosecution assert and compare how likely the observed evidence would have been when it is conditioned on each in turn. We will revisit this idea in later chapters. For now, we consider an example of conditional probability by adding another coin toss.

3.5 Example: double coin toss

Suppose now that we toss 2 coins, labelled coin 1 and coin 2. Coin 1 is tossed first and the outcome is recorded. Then coin 2 is tossed and its outcome is recorded. The outcome of the toss of coin 1 can be conditioned on to see how it affects the probabilities of the outcomes of coin 2. This means that the uncertainty of interest is the toss of coin 2 conditioned on the outcome of the toss of coin 1.

Suppose that the background information is that both coins are tossed in such a way that guarantees a 0.5 probability of heads for each coin. This means that the uncertainties and probabilities for the toss of coin 1 are unchanged from the single toss example in section **??**single-coin-toss). The possible outcomes for coin 1 are 'heads' with probability 0.5 and 'tails' with probability 0.5. Assuming that coin 1 has been tossed and is heads, what is the probability that coin 2 will be heads? The background information states that this probability should be 0.5 because coin 2 (like coin 1) is tossed in such a way that guarantees this. This means that the probability of a tails for coin 2 conditioned on a heads for coin 1 is also 0.5 (since those two probabilities must sum to equal 1). The same reasoning applies for conditioning on coin 1 being tails too; there is an equal probability for coin 2 of 0.5 for heads and tails conditioned on coin 1 being tails.

Figure **??** shows this information in an expected frequency tree for the double coin toss.

The first branch of the outcomes is the same as in section **??**single-coin-toss). After these outcomes of coin 1, there is a probability of 0.5 for coin 2 being heads or tails. In expected frequency terms, for every 5000 tosses in which coin 1 is heads we expect 2500 heads and 2500 tails from coin 2. Similarly, for every 5000 tosses in which coin 2 is tails we expect 2500 heads and 2500 tails from coin 2. This also means that out of the total 10,000 double coin tosses, we expect there to be 2500 double heads. As a probability, this means that there is a $\frac{2500}{10000} = \frac{1}{4} = 0.25$ probability of obtaining a double heads.

This probability of 0.25 for the double heads can also be obtained using another rule of probability, known as the **multiplication rule**. This mathematical rule states that the joint probability of two events, say events A and B, can be obtained by multiplying the probability of event A by the probability of event B conditioned on event A. The order of events can also be switched so that one multiplies the probability of event B by the conditional probability of event A conditioned on event B. In the double coin toss example, that means the probability of a heads for both coin 1 and coin 2 could be obtained by multiplying the probability of a heads from coin 1 with the conditional probability of a heads from 2 conditioned on the heads from coin 1. These probabilities were 0.5 and 0.5 respectively, and so multiplying these together gave the joint probability as 0.25.

The probabilities of coin 2 in this example are unaffected by conditioning on the outcomes of coin 1. This is technically known as **independence** between the

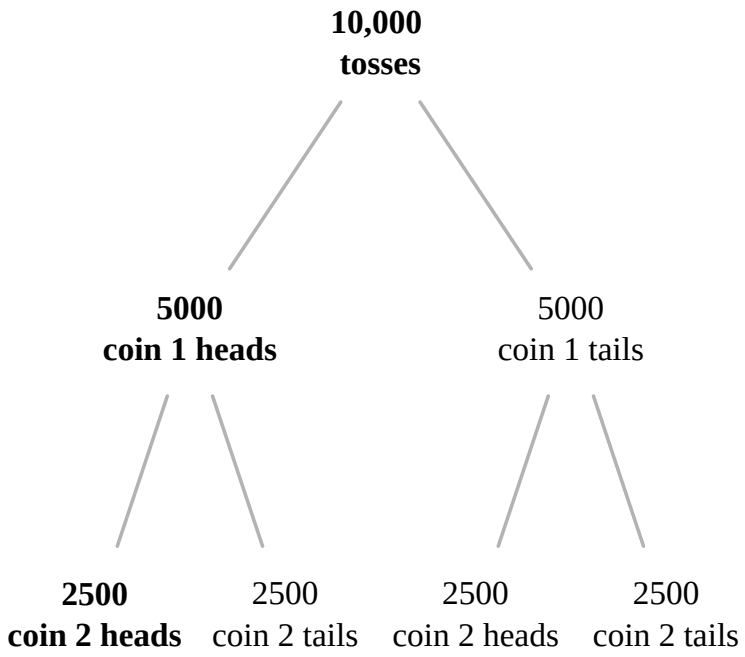


Figure 3.2: Out of every 10,000 double coin tosses, we expect 2500 to be double heads. The probability of getting two heads is 0.25.

coin tosses, and is an example of conditioning information that is not relevant. Independence between the coins was guaranteed by construction in this example because the background information made it so. In practical applications of probability, the independence between two processes should be made explicit and should be justified.

We can also consider two events for the same example which are not independent. Suppose there is a game that is decided by the result of the double coin toss. Player 1 wins if the coin tosses result in double heads and player 2 wins if the coin tosses result in double tails. If neither player wins then a draw is called and the coins are tossed again. Since double heads is just as likely as double tails, the probability is 0.25 for each, then both players are equally likely to win prior to the first coin being tossed. We can see this from the highlighted Figure **??** below.

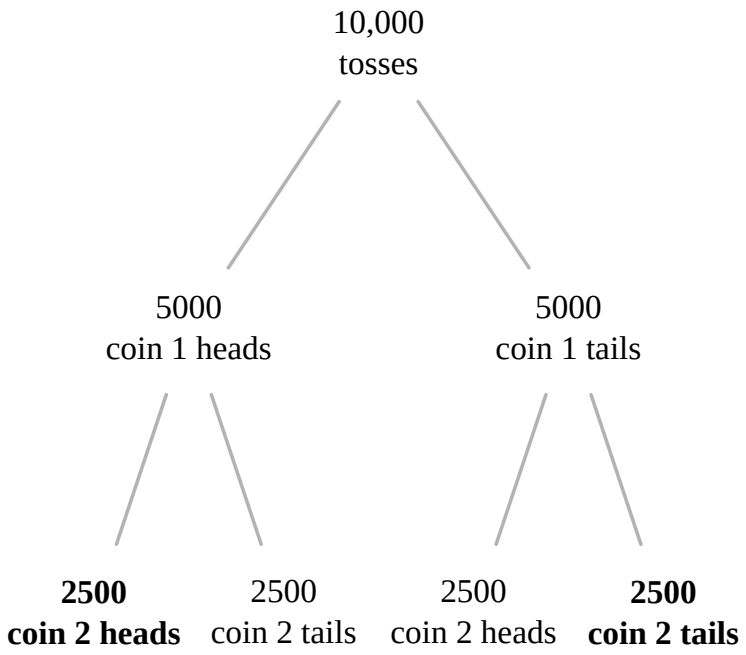


Figure 3.3: Out of every 10,000 tosses, 2500 are double heads and 2500 are double tails. This means that wthe players have equal probability of winning of 0.25 each. The remaining 5000 tosses result in a draw and so the probability of a draw is 0.5.

Out of 10,000 double tosses, player 1 is expected to win 2500, player 2 is expected to win 2500, and 5000 double tosses are expected to result in a draw. However, whilst the outcome of coin 1 does not affect the outcome of coin 2, it does affect the winning probabilities of this game. For instance after coin 1 has been tossed and shows a heads, there is now a 0.5 probability that player 1

wins, a 0.5 probability that there is a draw, and it is impossible that player 2 wins. Conditioning on coin 1 being heads has increased the probability of player 1 winning from 0.25 to 0.5, and it has decreased the probability of player 2 winning from 0.25 to 0. The opposite occurs when the outcome of coin 1 is tails. On the other hand, the probability of a draw remains unchanged at 0.5 before coin 1 is tossed, and 0.5 after it is tosses regardless of the result. In summary, the winning outcomes are not independent of the result of coin 1, but the draw is independent of it.

3.6 Odds

Odds are another way of expressing probabilities. Odds are given as a ratio of probabilities so that it is clear how much more or less likely one event is compared to another. This means that probability can be thought of as an absolute measure of uncertainty, whilst odds are a relative measure of uncertainty. Any two probabilities can be compared together to create odds. One pair which is commonly used is the probability of an event occurring versus the probability of it not occurring.

For example, the fair coin toss was just as likely to result in heads as it was to result in tails. This results in odds written as 1:1, which is spoken as '1-to-1' or more commonly 'evens'. Intuitively, odds of 1:1 means that we can separate the possible outcomes into a total of 1+1=2 parts which are equally probable. One part represents the probability of heads and the other part represents the probability of tails.

We can convert from odds to probability as follows. For the single coin toss we had odds of 1:1 for heads. The total probability, which must be 1, is made up of 2 equally sized parts. One of these parts represents heads, and so the probability of heads (and similarly for tails) is $\frac{1}{2} = 0.5$. Suppose we have odds of 1:9 for an event occurring versus not occurring. This means that the total probability, which must be 1, is made up of 10 equally sized parts. One of these ten parts represents the event occurring, and so its probability is $\frac{1}{10} = 0.1$. Odds which are used to represent probabilities before any conditioning occurs, are known as **prior odds**. The conversion from odds to probabilities is more challenging when the probabilities of both events do not sum to 1, but we do not consider that here.

Converting from probabilities to odds is much simpler as the important factor is simply how much bigger one probability is than the other. For example, the probability of player 1 winning the game in the double coin toss example by getting double heads was 0.25. That meant that the probability of player 1 not winning before any coins had been tossed was 0.75. Since the probability of player 1 not winning (0.75) was three times larger than the probability of player 1 winning (0.25), the prior odds of player 1 winning were 1:3.

Odds which are used to update prior odds by conditioning on more information are known as **posterior odds**. For example, as part of the double coin toss game we conditioned on the result of coin 1 being heads. Before the conditioning, the probability of player 1 winning was 0.25; the prior odds of player 1 winning were 1:3. Conditioning on coin 1 being heads resulted in the probability of player 1 winning rising to 0.5. This means that the posterior odds (i.e. **after** observing coin 1 being heads) of player 1 winning were evens (1:1).

Another reason that odds are useful is because of the simplicity that they give to a very important mathematical result, which we discuss in the next section.

3.7 Bayes' rule

Bayes' rule is a mathematical rule which links the **prior odds** to the **posterior odds**. Suppose we are considering odds of two events A and B, and we observe some new information from an event E. It is a natural idea to update the prior odds for A and B with the new information gained from E in order to obtain posterior odds - the odds obtained by conditioning on E. Bayes' rule states that this is done as follows:

$$\text{posterior odds} = \frac{\text{probability of E conditioned on A}}{\text{probability of E conditioned on B}} \times \text{prior odds}.$$

This rule states that the posterior odds of A and B conditioned on E are a product of the prior odds of A and B multiplied by a ratio of probabilities for E conditioned on A and B, respectively. This ratio is known as the **Bayes factor**, or **likelihood ratio** in this instance. From the formula, the likelihood ratio behaves as the updating factor for the prior odds due to the event E. In this sense it describes the relative support of events A and B to the event E: how much more (or less) probable was event E when conditioned on A compared to when it is conditioned on B? Understanding the likelihood ratio is the main goal of this book and it is revisited in Chapter **??**. Bayes' rule describes how prior odds must be updated in light of new conditioning information.

Bayes' rule gives us another mathematical expression that probability assignments must obey. This means that posterior odds or probabilities must be assigned coherently in that they must equal the prior odds multiplied by the likelihood ratio. This is useful for example when odds are based on a significant amount of epistemic uncertainty, since they can be hard to accurately quantify. Bayes' rule ensures that any quantified belief is still coherent. It does not remove subjectivity from the probability or odds assignment, but it does remove subjectivity from how that probability or odds assignment should be updated in light of new information.

Bayes' rule allows us to switch the conditioning information as we move from the likelihood ratio to the posterior odds. Probabilities for E which are conditioned

on A and B, the likelihood ratio, are updated to probabilities for A and B conditioned on E, the posterior odds. This switching of conditioning information is known as **transposing the conditional**. Bayes' rule gives us the correct way to transpose the conditional, i.e. by using the prior odds. Incorrectly transposing the conditional is a tempting mistake which is easy to make in practice. The most famous example of this in the legal domain is known as the **prosecutor's fallacy**, which we will define in a later chapter.

Despite seeming to be logical, Bayes' rule can lead to highly counter-intuitive results.

3.8 Example: guessing coin 1

The double coin toss example above was extended to a game in which player 1 won if two heads were tossed and player 2 won if two tails were tossed. They drew if there was any combination of heads and tails. When thinking about the probabilities of winning, we conditioned on coin 1 showing heads and also on coin 1 showing tails. In each conditioning, one of the players' probabilities reduced to 0 and the other's increased to 0.5.

Suppose the game is altered slightly so that player 1 now tosses the coins and doesn't show player 2 which sides were facing up. As part of the background information in this example, assume that player 1 always tells the truth and is tossing the coins in such a way to guarantee 0.5 probability of heads in either coin. Player 1 only tells player 2 whether they have won or not, i.e. whether double tails were tosses or not. If player 2 does not win, then they are given another chance to win by guessing the result of coin 1's toss. If they guess correctly, they win this new game. What should they guess?

Firstly, what are the uncertainties for player 2? This is a classic situation of epistemic uncertainty. Before the coin tosses, there is aleatory uncertainty - there is an unavoidable randomness to the future outcomes of these coin tosses. Focussing only on the uncertainty of coin 1, both heads and tails have equal probability of 0.5. These are prior odds of evens, or 1:1.

After the coin tosses, there are two possible situations. Either player 2 is told that they have won, in which case there is no uncertainty for them any longer because they know that coin 1 must have resulted in tails. Or, player 2 is told that they did not win, and now they have epistemic uncertainty about the results of both coin tosses. Player 2 is now offered the opportunity to guess coin 1 to win. This is direct uncertainty about the toss of coin 1 and so can be quantified. The outcomes of the double coin toss are presented as an expected frequency tree in Figure **??** below.

Player 2 knows that one of these three outcomes must have occurred: either coin 1 was heads and then coin 2 was either heads or tails, or coin 1 was tails and coin 2 was heads. For every 10,000 double tosses, these outcomes are expected

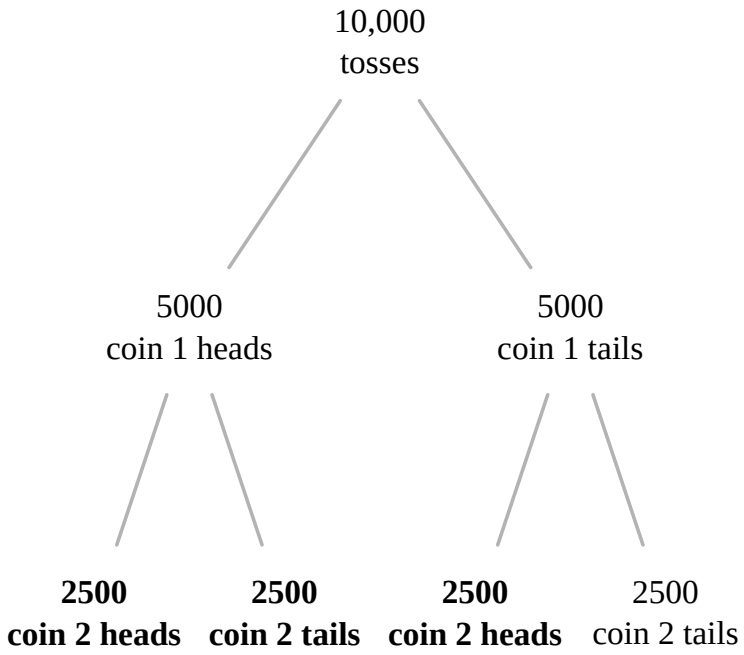


Figure 3.4: Player 2 expects not to win in 7500 out of the original 10,000 tosses. Out of those 7500 non-winning double tosses, 2500 came from coin 1 being tails, and 5000 came from coin 1 being heads. The posterior odds of coin 1 being heads are 2:1.

in 7500 cases. Out of these 7500 non-winning double tosses, coin 1 is expected to be heads in 5000 of them, and coin 1 is expected to be tails in 2500 of them. This represents posterior odds, having conditioned on the information that the outcome was not double tails, that coin 1 is heads of 2:1, since 5000 is twice as large as 2500. Player 2 should therefore always guess heads when given the choice in this game, since the odds will be in their favour. This result might seem counter-intuitive at first glance. The key thing to understand is that the information that player 2 has not won through getting double tails should update their belief about the possible outcomes of the tosses. We will now apply Bayes' rule to this problem to see this more clearly.

We can verify that posterior odds of 2:1 in favour of coin being heads are correct by using Bayes' rule. Event A from Bayes' rule in this example is coin 1 being heads and event B is coin 1 being tails. Event E is the knowledge that the result was not a double tails. To apply Bayes' rule, we need to calculate the prior odds of coin 1 being heads compared to coin 1 being tails and also the likelihood ratio for this situation.

The likelihood ratio in this example considers the following question: how much more (or less) probable is it to **not** toss double tails if coin 1 is heads compared to if coin 1 is tails? To calculate this, we need to first calculate two probabilities. The first is the probability of not tossing double tails conditional on coin 1 being heads. This probability is 1. If coin 1 shows up heads, then it is certain that a double tails will not be tossed.

The second probability we need is the probability of not tossing a double tails conditional on coin 1 being tails. Out of every 5000 tosses in which coin 1 is tails, we expect 2500 to lead to coin 2 being heads and 2500 to lead to coin 2 being tails. The probability is 0.5. Using these two probabilities results in a likelihood ratio of $\frac{1}{0.5} = 2$; an outcome other than double tails is twice as likely when coin 1 is heads compared to when coin 1 is tails.

The second component of Bayes' rule that we needed was the prior odds of coin 1 being heads. We reasoned that these odds were evens because before we toss coin 1 there is an equal probability of heads and tails. The prior odds of coin 1 showing heads are 1:1.

Bayes' rule dictates that the posterior odds must be equal to the prior odds multiplied by the likelihood ratio. With a likelihood ratio of 2 and prior odds of 1:1, we obtain posterior odds of 2:1 in favour of heads; coin 1 is twice as likely to be heads than tails when we know the outcome was not double tails. This result is the same as when we reasoned without using Bayes' rule, and so we have verified that it satisfies Bayes' rule.

In this example we could calculate the posterior odds easily without using Bayes' rule, this meant that we could verify the probability assignment. In many real situations, the posterior odds are hard to quantify in a way that guarantees that they satisfy Bayes' rule without using Bayes' rule itself to derive them. This is where Bayes' rule is most powerful, because all that is required is to quantify

prior odds and the likelihood ratio in order to obtain logically coherent posterior odds.

3.9 Reliable probabilities

Using our notion of probability, anyone is capable of assigning a probability to their personal uncertainties. But even if everyone does this to the best of their ability, some people simply have more information about some events than others. This is clear from the idea of experts versus non-experts: the expert's probability assignment will be more **reliable** than that of the non-expert. Reliability here refers to the fact that the expert's knowledge is closer to all the available knowledge about a given topic, and so their best assessment is better than an uninformed non-expert's. This is the idea of the expert witness. Reliability of probability assignments is a key part to interpreting and using expert evidence.

We mentioned in a previous section that background information through assumptions was important since it acts as conditioning information that informs the probability. In practice what happens is that assumptions are made and stated upfront, so that they are inherently conditioned on when any probability is stated. We did this in the coin toss example when we assumed that the coin tosses were done in such a way so as to favour neither heads nor tails. In practice this might not be able to be guaranteed, and so instead the assumption might be justified by a reasonable 'default' state of knowledge, i.e. by assuming that the coin tosses are fair unless there is reason to believe otherwise. This is often done in practice when assigning probabilities to real events, e.g. in the expert report or its supporting information. Assumptions like this in practice might be considered widely reasonable in the scientific community and therefore accepted by most experts.

In all cases, it is important that assumptions are made transparent so that when they are unreasonable they can be rightly challenged. This makes probability assignments **assessable**. They can then in principle be assessed for their reliability. We saw that independence was an important assumption that is often made. It is done so because it makes complex probability calculations easier. However, it can result in seriously inaccurate probabilities when it is incorrectly assumed.

Another thing that can be checked is the general **calibration** of personal probability assignments. This is a measure of how accurately an individual's quantified uncertainty reflects the true uncertainty. This can be done by asking experts to assign probabilities to events whose probability is known (but unknown to the expert). For example, an expert who assigns a probability of 0.4 to an event whose actual probability is 0.45 is calibrated better than an expert who assigns a probability of 0.8 to the same event. Greater calibration leads to greater reliability. Experts in forensic science demonstrate this in some

Table 3.1: Labelling statistics based on the assigned label and the underlying truth.

Truth	Labelled positive	Labelled negative
Positive	**True positive**	**False negative**
Negative	**False positive**	**True negative**

capacity by performing competency tests. This involves conducting simulated examinations of evidence when the underlying truth is known to the examiners, but unknown to the experts.

Some probabilities can be empirically validated using repetitions of the same event. When this cannot be done, e.g. for one-off events, then information from similar events can be used to inform that probability assignment. This is part of a wider idea of using empirical **data** to construct more reliable probabilities. If there are known data relating to an event of interest, then an individual's best assessment of the probability of that event must incorporate that data. This idea moves us from theoretical uncertainty to observed uncertainty and from probability to statistics. Statistics is the focus of the next chapter. In the remaining sections, we apply the ideas introduced in this chapter to more realistic practical scenarios.

3.10 False positives

False positives and false negatives are terms to describe mistakes in uncertain categorical assignments. Typically these are binary assignments where something or someone is either labelled as a **positive** case or a **negative** case, and the truth about them actually being a positive or negative case is unknown.

If the truth is that they are a negative case, but they were mistakenly labelled as a positive case, then the assignment is a **false positive**. If the truth is that they are a positive case, but they were mistakenly labelled as a negative case, then the assignment is a **false negative**. If the labels were correct, then the assignment was a **true positive** or **true negative**, respectively. This information is presented in Table **??**.

For example, when testing someone for a specific disease, we are uncertain about whether or not they have the disease before applying the test. The test results categorise them as either positive or negative for the disease, but it is never absolutely guaranteed to be correct. Even the most reliable tests make mistakes sometimes, even if that's only very rarely. The test result should decrease our uncertainty about whether the tested person has the disease or not, but it can't totally eliminate it. The best tests will greatly decrease our uncertainty, and the not-so-good ones won't change it much.

Table 3.2: The number of people who are affected by the disease and their diagnostic test results.

Disease	Test positive	Test negative	Total
Present	99	1	100
Absent	495	9405	9900
Total	594	9406	10000

If many assignments of positive/negative have been made under controlled conditions, e.g. when the underlying truth of positive or negative is known, then one can determine the **rate** of true/false positives/negatives. This rate corresponds to the probability of each entry in Table **??** occurring.

The probability of a false positive occurring is called the **false positive rate** and the probability of a false negative occurring is called the **false negative rate**.

The probabilities of true assignments have different names. The probability of a true positive is called the **sensitivity** and the probability of a true negative is called the **specificity**.

The **base rate** of a characteristic is the probability that when we randomly select an object from the population of interest, then that selected object has the specified characteristic. This is commonly called the **prevalence** when the characteristic that we are interested in is a disease.

3.11 Example: diagnostic tests

The following example is adapted from **?**.

The risk of a disease is 1% in a relevant population of 10,000 people. This means that the disease affects 100 people out of the total 10,000, and it does not affect the other 9,900.

A diagnostic test has been created for this disease. The test has a sensitivity of 99%; out of the 100 people who have the disease, 99 of them have a positive test result. The final 1 person tests negative despite having the disease. This person receives a false negative result.

The test has a specificity of 95%; out of the 9,900 people who do not have the disease, 9,405 have a negative test. The other 495 people test positive despite not having the disease test. These people receive false positive results. This information is displayed more clearly in Table **??**.

This test has high sensitivity (99%) and high specificity (95%), which makes it sound reliable. However, remember what these terms mean: the probability

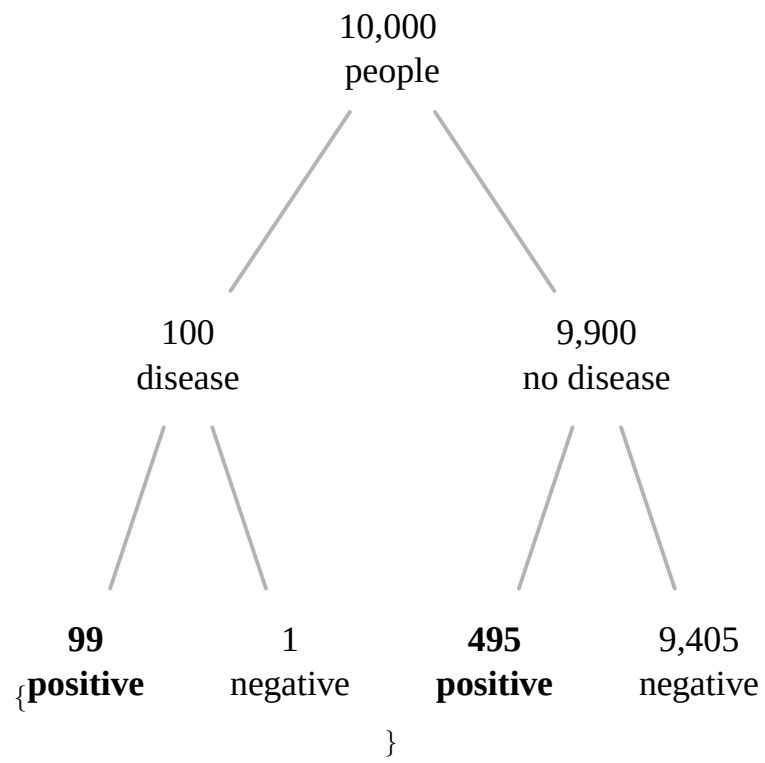
of testing positive given that you do have the disease (sensitivity), and the probability of testing negative given that you don't have the disease (specificity). This probability is conditioned upon knowing whether the person has the disease or not.

In practice, people do not know whether they have the disease or not, and that is why they get tested. This means that this is not useful conditioning information in practice. The information that people do have is whether their specific test result was positive of negative, and so this is the information that the probability should be condition on. What's the probability of actually having the disease given the result of the test?

Look back to the columns of Table **??**. Consider the negative results first. A total of 9,406 people from our population of 10,000 tested negative. Out of these 9,406 who tested negative, 9,405 did not have the disease. There was only a single individual who tested negative despite having the disease. This means that a negative test result is a great (but not perfect) indicator for not having the disease.

Now consider the positive results. A total of 594 people from our population of 10,000 tested positive. Out of these 594 who tested positive, only 99 (~17%) actually have the disease. The large majority of people who tested positive, 495 (~83%) of the 594, do not really have the disease. This can be seen clearer in Figure **??**.

\begin{figure}



{

}

\caption{Out of the 594 people who test positive (shown in bold font), 99 (~17%) have the disease. } \end{figure}

If a randomly selected individual from this population tests positive, then it is highly likely that they do not have the disease. A positive result for this test is a terrible indicator of whether someone has the disease. This is a counter-intuitive result that can be explained using the logic of Bayes' rule.

This phenomenon is caused by the very low **base rate** (prior probability) of the disease. This is the same as the risk of having the disease for people within the population, which was 1%. A randomly selected individual has a very low probability of having the disease prior to being tested. The test has very high sensitivity and so it is able to detect almost all of the true positives. The issue was that it tested many people who didn't have the disease, and so even a low error rate led to many false positives. Due to the **base rate** being so low, the number of true positives (99) was much smaller than the number of false positives (495). This meant that the positive results largely consisted of false positives.

3.12 Example: doping

The following example has been adapted from **?**.

Table ?? and Figure ?? present the technical diagnostic testing information in a format which is easier to understand and base decisions on. However, it is often not presented this way in practice and so must first be 'translated' from the raw numerical information.

> A test designed to detect athletes who are doping is claimed to be '95% accurate'. If an athlete is doping then the test returns positive 95% of the time, and if the athlete is not doping then the test returns negative 95% of the time. It is suspected that around 1 in every 50 athletes dope. An athlete tests positive for doping using this test during a random drugs screening. How likely is it that they are really doping?

The answer is around 28%, pause for a moment and see whether that answer comes to you from reading the above text. After reflecting on the text, continue through the example below, where we present this same information in a more familiar format.

We can convert some of the written information into our technical definitions. The second sentence states that the sensitivity and specificity are both 95%, although those words are not explicitly used. The base rate for doping is given as approximately 2%.

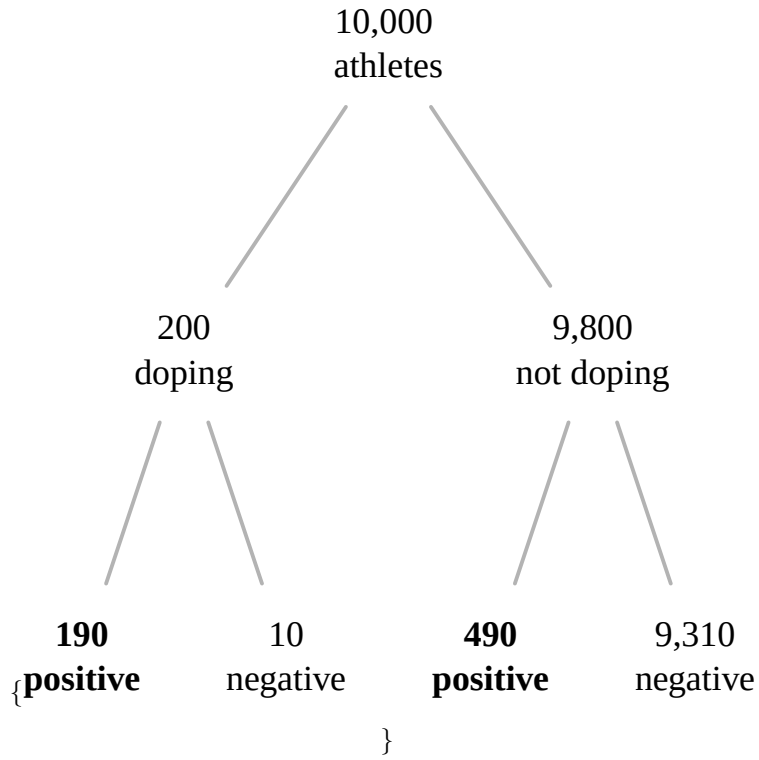
We haven't been given a relevant population size to use natural frequencies to describe these rates, but we can imagine one in order to aid our understanding. Since we are going to use a hypothetical population of athletes, we will have to talk in terms of what we would expect from such a population and so can use expected frequencies.

Assume, for clarity, that we have a relevant population of 10,000 athletes. Using the base rate, we expect 200 (2%) of these to be doping and 9,800 (98%) not to be doping. The sensitivity tells us that out of the expected 200 athletes who are doping, the test is expected to return positive for 190 (95%) of them and negative for 10 (5%) of them. We expect 10 false negatives.

Out of the expected 9,800 athletes who are not doping, the specificity tells us to expect 9,310 (95%) to test negative. We expect 490 (5%) of these non-doping athletes to test positive; we expect 490 false positives.

We expect a total of 680 positive tests and 190 (~28%) of those positive tests to be from an athlete who is doping. The answer to our original question is that given a positive test result, we expect the athlete to be doping roughly 28% of the time. This information is presented in the expected frequency tree in Figure ??.

\begin{figure}



{

}

\caption{Out of the 680 athletes who test positive (shown in bold font), 190 (~28%) are doing. } \end{figure}

3.13 Try it yourself

3.14 End of chapter survey

3.15 More information

Chapter 4

Statistics to infer uncertainty

In the last Chapter we spoke about probability as a means of quantifying our uncertainty. Probability was introduced as personal in the sense that it depends upon the information and beliefs of an individual. However, these beliefs are not completely arbitrary as they should always be a best assessment of the available information. One way to improve this assessment and increase and share available information is to gather data from empirical observations. These data can then be used to inform our beliefs and create more reliable probabilities.

The branch of mathematics that is concerned with describing and learning from empirical observations is known as statistics. In this Chapter we will look at statistics, how it relates to probability, what are some problems that statistics can be useful for, and how this is applied in forensic scientific evidence.

4.1 Learning from data

When we were thinking about probability, we were quantifying our theoretical uncertainties. This involved us thinking about our uncertainties to map out what we believe could occur in theory, sometimes creating a probabilistic model for this, and then using that theory to form expectations about would occur in practice. In the coin toss example, my probability of heads being assigned as 0.5 meant that I expect 5000 heads out of 10,000 tosses of that coin. But, what I expect to occur is rarely what we observe in reality.

Statistics is all about doing this process in reverse. We use observed events in the real world in order to try to learn about what kind of probabilistic models

might have lead to them. For example, suppose we tossed a new coin 10 times and observed 7 heads. What should we infer from these tosses about the probability of heads for this new coin? One reasonable answer would be a probability of heads of 0.7 because that is the proportion of heads that we actually observed. Another reasonable answer is still 0.5 because, although this means that we would expect 5 heads from 10 tosses, we are not guaranteed to always observe 5 heads for every 10 tosses. We might ask how rare it would be to observe a sequence of 7 out of 10 heads when the probability of heads is 0.5 compared to when the probability of heads is 0.7. How we might then choose between these inferences about the probability of heads is a question that statistical theory can help with.

Probability and statistics are complementary. Probability is useful because it can quantitively describe our theoretical uncertainty. Statistics is useful because it can quantitively describe the uncertainty that is observed in practice. This empirical uncertainty can then be used to compare to and refine our theoretical uncertainty. Where possible and practical, this process can then be repeated many times to reduce epistemic uncertainty and characterise aleatory uncertainty.

This results in probabilistic models which are continually improved by gathering **data** about them from empirical observations. In the previous short example, when we tossed the new coin 10 times we generated data about the coin toss. The data was the ten coin tosses and their results. We can summarise data or perform mathematical operations on it to create quantities known as **statistics**. For example, by summarising that 7 out of the 10 tosses (70%) resulted in heads, we have a statistic which describes the proportion of heads in that coin toss data.

We then used this data, and the statistic computed from it, to think about the probability of heads occurring from the toss of the new coin. We could toss the coin many more times to gather more data, compute more statistics, and reassess our probabilities. This is called learning from data. Using statistical theory makes the magnitude of learning match the amount of information in the data (subject to our beliefs); without using statistical theory this would not be guaranteed.

Using probability and statistics together in this way has many similarities with the scientific method; generating hypotheses and then performing experiments to gather evidence to test those hypotheses against. This methodology is why it is sometimes known as **statistical science**. In most of the processes in the real world we do not know the underlying models which cause them and so we use statistical science to learn about them. This can be done to achieve many goals, and we restrict ourselves to the following:

- describing empirical observations,
- inferring general conclusions from empirical observations,
- evaluating empirical observations.

We get to these topics in the next few sections but first we need to introduce a couple of important concepts.

4.2 Populations and samples

Two key concepts in statistics are populations and samples. A **population** is every possible event/characteristic/individual in a certain group of interest. Some examples of populations are

Population 1: all of the people living the UK,

Population 2: all burglaries in the Greater London area of England,

Population 3: all the footwear of people living in a small local region.

Having information about these populations can have important practical impact. For example, knowing the locations of all burglaries in Greater London might reveal patterns such as hotspots. The Metropolitan Police could use this information to put preventative measures in place.

Collecting data about the entire population is known as a **census**. The UK census is one such example, gathering data on people and households within the UK, occurring once every 10 years. Censuses are often very difficult or impossible to fully complete. This can be for logistical reasons such as it being too expensive and resource intensive to conduct. It can also be because the population is unable to be fully surveyed. Population 3 above would be highly relevant for criminal investigations within the specific local region, e.g. for comparing footwear marks found at local crime scenes. However, it is practically impossible to gather and record all of this information. One reason for this would be legal restrictions preventing the systematic collection of this data.

For these reasons, it is common to take a **sample** from the population instead. A sample is a subset of members out of a population. Since samples don't aim to gather information from the entire population, they are much simpler and cheaper to conduct. As well as saving resources, the aim of taking a sample is to be sufficiently representative of the population. Below are some examples of samples:

Sample 1: a survey of people within the UK,

Sample 2: reported burglaries in a Met. Police database,

Sample 3: voluntarily submitted footwear marks from a local area.

Sampling has many different types as it is such an important aspect of statistics. In this text, we will only mention random samples and convenience samples. A **random** sample is a sample drawn from a population in such a way that every member in the population has an equal probability of being

sampled. In sample 1 above, the survey can be conducted by randomly contacting potential respondents from a representative database of people.

A **convenience** sample is a sample drawn from a population in which only the most available members of the population are sampled. This means that convenience samples might not be representative of the population as a whole. For sample 2 above, it might be reasonable to say that these reported burglaries are fairly representative of all burglaries in Greater London i.e. population 2. This assumes that almost all burglaries are reported and logged and that there are no systematic differences in the burglaries which are not being reported (e.g. in their location).

Sample 3 is also a convenience sample, this time of submitted footwear. The degree to which the submitted footwear marks are representative of all footwear in a local area (e.g. population 3) depends upon the representativeness of the volunteers who submit them. For example, if a certain age group of people, say teenagers, is over-represented in the volunteers then the sample will be **biased** in favour of them. This means that when trying to count a simple measure such as the most common shoe size in the local area, the data will be biased in favour of shoe sizes which are probably smaller than the true most common size. This is because the teenagers will still be growing. If desired, statistical methods can be applied to take this bias into account, e.g. by weighing the teenager group's submissions proportionally less compared to other age groups. This is one way polling companies attempt to make a potentially biased sample more representative when trying to determine election outcomes.

Sampling adds another layer of uncertainty to consider, which lends itself to working with probability. In any sampling method, there is uncertainty about how representative the sample is of the population from which it is drawn, and also a natural randomness as to which members are sampled. This represents both epistemic and aleatory uncertainty. The epistemic uncertainty arises because we do not know some characteristics of the population (which is why we are sampling from it). Learning more about the population reduces this uncertainty, and so the sampling procedure itself can reduce this.

The aleatory uncertainty is because there will always be randomness when sampling. Randomness is clearly present in a random sampling procedure, but it is also present within convenience samples. This is because there is a randomness as to which convenient members are sampled. **Sample size** is an important factor for characterising the aleatory uncertainty associated with sampling. The more we perform a sampling procedure in a population, the more we learn about its inherent variability. But sample size isn't everything: a small but carefully constructed random sample can be more representative of the population than a large but careless convenience sample.

Now that we have seen what populations and samples are, we can look at the first use of statistics from our list: describing observations.

4.3 Describing observations

Using statistics to describe observations is probably their most familiar application. This is because they concisely and precisely report and summarise data, and this plays a central role in our ever growing data-centric society. These type of statistics are referred to as **descriptive statistics**. Some common examples are the average height of males or females in the UK, median household income, most popular baby's name for a girl etc. Descriptive statistics give us information about a particular set of data. Two common categories these fall into are **central tendency** and **dispersion**.

Measures of central tendancy aim to tell us a number around which the data tend to cluster. There are three common measures used for this: the **mean**, **median**, and the **mode**. The **mean** describes the average value of the data. The **median** describes the value in the middle of the data after we put all of the numbers in order. The **mode** describes the most common value on the data.

Dispersion aims to tell us how closely the data cluster around the measures of central tendency. The most common of these known as the **standard deviation**. The **standard deviation** tells us on average how close the data are to the mean. Larger values indicate that the data is not well represented by the mean since they are spread widely around it. Smaller values indicate that the mean is a good representation because the data are tighly packed around it.

When naming specific descriptive statistics, it is good practice to mention whether that statistic is describing a sample of data or a population. This makes it clearer and more interpretable to the consumer of the statistic because they can think about it in context, e.g. "could the statistic be biased by convenience sampling?" When a statistic describes a population, it is referred to using words such as *population*, *true*, *parameter*, e.g. the population mean. When a statistic describes a sample it is usually referred to as a **sample statistic** or using the word *empirical*, e.g. the sample mean. When this distinction is not made, there is the risk of confusing a sample for a population, which ignores uncertainty and leaves the potential to make an unwarrented and inaccurate generalisation. If in doubt, assume everything is a sample. Even for census data this can be a safe assumption since some forms are bound to not be completed or go missing after completion (even if this is a really small number).

The following example looks at descriptive statistics in action.

4.4 Example: GSR publications

In 2020, a systematic analysis of all academic publications related to gunshot residue (GSR) particles was published by **?**. Among other things, the study

Table 4.1: Academic publications related to gunshot residue analysis from the years 2010-2018.

Year	Publications
2010	20
2011	27
2012	45
2013	38
2014	37
2015	31
2016	41
2017	47
2018	45

looked at trends over time in the number of publications which are related to GSR. A small subset of the data, containing information about the number of publications per year between 2010-2018, can be found below in Table **??**.

Using this data, we can calculate some descriptive statistics. Let's assume this is a sample of the total population of GSR-related publications during the same timeframe. This is a safe assumption because some publications (albeit very few) may have been missed during the literature search.

To calculate the **sample mean**, we calculate the total number of publications and divide this by the number years for which we counted publications. Using the GSR publication data, this gives:

$$\text{sample mean} = \frac{20 + 27 + 45 + 38 + 37 + 31 + 41 + 47 + 45}{9} = \frac{331}{9} \approx 36.8,$$

which means that the sample mean number of GSR-related publications per year during the period 2010-2018 is roughly 36.8. In other words, if we wanted to report a single number for the rate at which GSR-related publications were published each year between 2010-2018 in this sample, then that rate would be 36.8. You can confirm this by multiplying 36.8 by the number of years (9) and compare that to the total number of publications.

Notice how 36.8 isn't itself a valid number of publications for any given year - the average doesn't exist as its own data point, it is simply a statistic which describes the data points. This highlights that there isn't such a thing as an "average year" for instance, which is sometimes how the average is reported.

How much spread around this sample mean is there in this data? We can answer that by calculating the **sample standard deviation**. The idea is to see a single number descriptive statistic for how different the number of publications are each year compared to the sample mean. The closer the

number is to 0, the tighter the data are grouped around the sample mean. The further away from 0, the more spread there is.

To show some intuition for the sample standard deviation and how it is computed, we can plot the number of publications each year and highlight the difference between these and the sample mean.

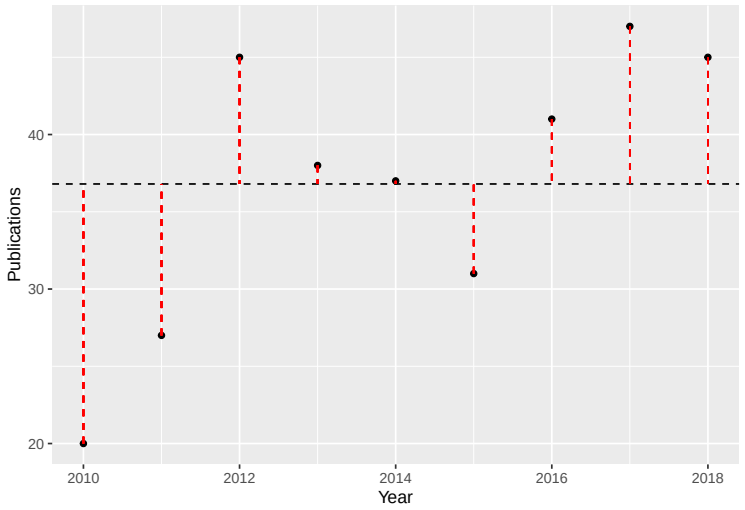


Figure 4.1: The number of GSR-related publications between 2010-2018. The black dashed line represents the mean number of publications per year in this sample. The red dashed lines show the difference between the sample mean and the observed number of publications.

The standard deviation reduces the differences in Figure **??** into a single number summary. The first couple of steps involve calculating the differences, squaring them (this mean multiplying them by themselves), and then adding all of these squared differences together. The squaring operation means that each individual difference becomes a positive number and so we end up adding together a collection of positive numbers. If any of the numbers were negative, then adding them together with positive numbers would decrease the total sum of differences. Squaring the differences also means that bigger differences become much bigger. This means that big differences contribute much more to the total sum of differences. We can see this by calculating the differences and their squares in Table **??** below.

If you look at the year 2013 as an example, you can see that its difference from the sample mean is only 1.2. Multiplying this by itself gives $1.2 \times 1.2 = 1.44$, and so the squared difference is only a small increase from the original difference. Compare that to the year 2010. For that year, there is a difference of -16.8 compared to the sample mean, but when this is squared we get $(-16.8) \times (-16.8) = 282.24$. The years 2013 and 2010 are not too far away in

Table 4.2: Number of GSR-related publications per year. The differences represent the number of publications that year minus the sample mean number of publications 36.8. These differences are then squared in the squaredDifferences column.

Year	Publications	sampleMean	Differences	squaredDifferences
2010	20	36.8	-16.8	282.24
2011	27	36.8	-9.8	96.04
2012	45	36.8	8.2	67.24
2013	38	36.8	1.2	1.44
2014	37	36.8	0.2	0.04
2015	31	36.8	-5.8	33.64
2016	41	36.8	4.2	17.64
2017	47	36.8	10.2	104.04
2018	45	36.8	8.2	67.24

terms of their difference to the mean (1.2 compared to -16.8), but they are very far away on the squared difference scale (1.44 compared to 282.24). Notice how squaring decreases differences which are less than 1 (this happens in 2014).

Now that we have the squared differences we add them all up to get a total of 669.56. The next step is to divide this total by a number so that it approximately represents the squared differences per year. This is where the calculation of the standard deviation can be different for a population versus a sample.

For a population, we would divide the sum of squared differences by the number of years which contributed to the sum, 9 (just as is done for both the population and sample mean). This is the more intuitive quantity because it exactly represents the squared differences from the sample mean per year.

However, for a sample, there are two ways we can do it. We can either divide the sum of squared differences by the same number as for the population standard deviation, 9, or we can divide by this number minus 1, which is 8. The former is known as the biased sample variance, and the latter as the unbiased sample variance. We focus on the unbiased sample variance in this book.

The reason that the unbiased sample variance divides by this less-intuitive number involves more complex statistical theory which is not covered here. One intuition behind it is that the number we divide by is like a currency. Each data point gives an extra unit of currency to spend. We pay a price of this currency in the computation of a new statistic that uses the original data as well other sample statistics based upon that data. The price paid is equal to the number of extra statistics we use in the computation of the new one.

To compute the mean (population or sample), we added the data for the number of publications per year together as the first step. Since this involved no statistics, we paid no price and so could divide by the number of data points. To compute the population standard deviation, we would add the squared differences from the population mean together. Since this involved the population mean, and not a sample statistic, no price is paid and so we could just divide by the number of data points. For the sample standard deviation however, the squared differences are computed based upon the sample mean. Since this is a sample statistic, we pay a price equal to one. We therefore divide the sum of squared differences by the number of data points minus 1.

Back to the GSR dataset, when we divide the sum of squared differences by $9 - 1 = 8$ we get 83.695. The squared differences and the result after dividing their sum by 8 is shown in Figure **??** below.

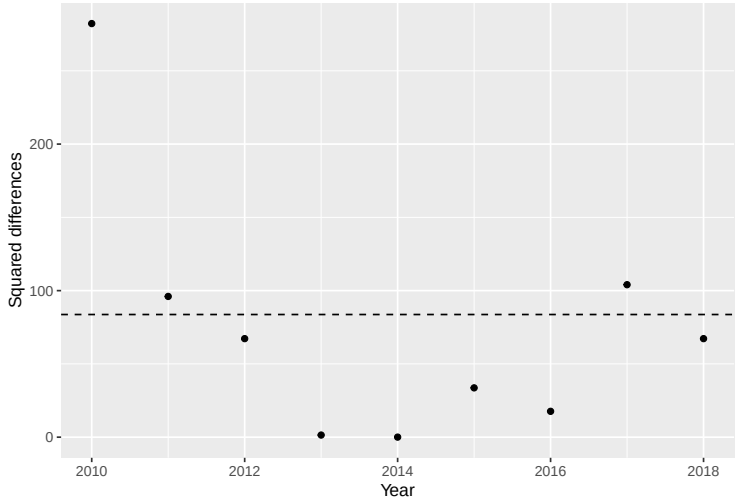


Figure 4.2: The squared differences between the actual number of publications each year and the sample mean. The dashed line shows the sum of the squared differences 669.56 divided by 8.

At this point, the number 83.695 represents the typical squared difference between the data points and the sample mean. It is more interpretable to convert that from the squared scale of differences back to the original scale. This is the final step of computing the unbiased sample standard deviation. The mathematical operation which does this conversion, the 'opposite' of squaring, is called the square root. The square root takes a number and finds which other number is squared to make it. The square root of our quantity of interest is $\sqrt{83.695} \approx 9.15$. This is the unbiased sample standard deviation.

To recap, the steps we took to compute the unbiased sample standard deviation were:

Table 4.3: Academic publications related to gunshot residue analysis from the years 2010-2018 in ascending order.

Year	Publications
2010	20
2011	27
2015	31
2014	37
2013	38
2016	41
2012	45
2018	45
2017	47

1. compute the sample mean,
2. compute the differences in the number of publications each year compared to the sample mean,
3. square the differences,
4. add them together,
5. divide by the number of years which contributed to the sum minus 1,
6. compute the square root.

A sample standard deviation of 9.15 means that there is a typical spread of 9.15 around the sample mean of 36.8 in this dataset.

The **sample median** is best seen by rearranging Table **??** so that the number of publications are in order from lowest to highest, rather than in chronological order by year. This is shown in Table **??**.

Since we have 9 values, the middle value is the one which has 4 values lower than it and 4 values higher than it. This is seen by inspecting Table **??**. The median number of GSR-related publications in this sample during the period 2010-2018 is 38. This is quite similar to the sample mean of 36.8, and so these two measures of central tendency are in general agreement with each other.

The median isn't always this easy to calculate. When there is an even number data points then we have to decide upon a consistent method to select the median value. For example, if we excluded 2010 from our sample of GSR-related publications, then we have 8 years of publication data. There are then a range of values between 38 and 41 which leave 4 observed values above and below them. However, the technicalities of these situations are beyond the scope of this content.

The final measure of central tendency to calculate is the **sample mode**. This is the most common number of GSR-related publications that appears in our

sample. We can see from Table **??** that one value appears twice, and all others appear once. The sample mode is therefore the value which appears twice, which is 45.

All of the above statistics should be subject to a fair amount of criticism for this dataset. The least reliable here is the mode. With just two occurences, including another single year with 20 publications into the dataset would change it from 45 down to jointly 45 and 20. All of these descriptive statistics miss the main message of this dataset, which can be seen from Figure **??**: there appears to be an upwards trend in the number of GSR-related publications during this time period. There are descriptive statistics for this too, but the main message is that descriptive statistics alone, whilst very useful, might not tell the whole story.

Depending on the question of interest, it could be meaningful to compare descriptive statistics from this sample to those of another sample of publications, e.g. the same GSR publications from the 9-year period 2001-2009. Such comparisons often go beyond just describing the data, and begin to start to infer conclusions from them. This brings us on to the other type of statistics: inferential statistics.

4.5 Inferring from observations

Inferring from observations means drawing more general conclusions about populations based upon information gathered from samples. This is known as **statistical inference**, and it is done using **inferential statistics**. Inferential statistics go beyond describing information in a sample, and that is what makes them different from descriptive statistics.

A popular application of statistical inference is **estimation**. Estimation aims to learn about some property of a population of interest. A common type of estimation is **parameter estimation**, in which certain parameters of the population are estimated using probability models and statistics from samples. We saw an example of this at the beginning of this chapter. The example stated that a coin was flipped 10 times and 7 of those resulted in heads. The sample proportion of heads, a descriptive statistic, is 0.7. If we use this sample statistic to make an inference about the population (e.g. the population proportion of heads, or the probability of heads is 0.7) then we say it is an **estimate** and in this use it is an inferential statistic. This estimate is a single number summary, and we call such estimates **point estimates**. In the example above, we used the sample proportion of heads as a point estimate for the population proportion of heads.

As well as estimation using point estimates, we can perform estimation using **interval estimates**. Rather than using a single number summary, interval estimates give a range of values as their estimate. For example, given that the

sample proportion of heads from 10 coins flips was 0.7, we might take an educated guess that the population proportion of heads is between 0.4 and 1. We could make this range more narrow, say 0.6 to 0.8, but as we make the interval more narrow, we also increase our uncertainty. Logically, we have to be less certain that the population proportion of heads lies between 0.6 and 0.8, compared with 0.4 and 1, since the former interval is fully contained within the latter interval. A common interval estimate is a confidence interval, where the size of the interval depends upon a level of confidence we wish to have (amongst other factors).

Estimates are carefully selected so that they have desirable statistical properties and are supported by statistical theory as being "optimal" in some sense. For example, the sample proportion of heads can be thought of as the point estimate which maximises the probability of observing 7 out of 10 coins as heads (with some reasonable assumptions). In addition, a correctly constructed 95% confidence interval has the property that if enough samples are gathered and intervals calculated for each sample, then 95 out of every 100 of them will contain the population parameter. Individuals may disagree about what are "reasonable" assumptions or which criterion should be maximised and so there can be reasonable discrepencies between estimates — there may be no "correct" answer. Transparency about these matters is key to intepreting and evaluating inferences.

From this basis of estimation we can branch out into other forms of inference such as statistical hypothesis testing, making predictions about the future, and, most relevant to forensic science, evaluating scientific evidence. What is important before making any statistical inference is the data upon which the inference will be based. The quality of the inference can only be as good as the quality of the data. Assessing the quality of data is the topic of the next section.

4.6 Data sources

A data source is something which contains data. This broad definition covers anything from extensive national criminal databases, which contain information about convicted criminals, to handwritten personal notes about the outcome of coin tosses. Examples between these extremes include academic journal articles detailing experiments related to forensic problems and their results, and proprietary in-house databases with images of footwear marks from crime scenes managed by commerical forensic service providers. Clearly there is a range in quality for these sources of data. A range in quality for the data means that the resulting statistical inferences will necessarily vary in quality. That is why it is important to be able to recognise the characteristics which contribute to a high quality data source when presented with any statistical inference.

Statistics can be computed on any source of data. The source of data need not be personally collected. As we saw in the GSR example, the source of data can be second-hand from someone else's data collection. This grants statistics a wide range of applications, and means that so long as the data is accessible, then it can be analysed. When data is collected and analysed by the same individual or research team, then the experimental protocal and quality of the data is known by them. However, when data is analysed second-hand, then the experimental design is unknown and so is the quality of data. This makes transparancy and trust key. Transparancy can be achieved by detailing the conditions of the experiment or sampling methodology which are meaningful to the quality of the data and resulting data analysis.

quality: there a few main components to think about when it comes to assessing the quality of a data source.

1. variability. variability refers to how much the data vary. this can be because the target population varies or because the data capture protocol influenced variability. Variability can be better understood through repeated sampling.
2. size. size refers to the size of the data source.
3. validity. validity refers to the integrity of the data, its collection strategy, its measurement strategy, any verification, replication/reproduction, and accreditation.
4. representativeness. this relates to how representative the collected data is of its target population or overall purpose.

all of these components are related to each other. for example, a data source which measures something with very low variability will require a smaller number fo measurements to capture this variability, and therefore a smaller overall size.

In addition a big data source which has no validity is of little use, as is a data source which does not sufficiently represent its target population.

As well as assessing the quality of a data source, another factor to consider is its relevance to the circumstances in which it is being used. This will vary on a case-by-case basis, depending on the question that is being addressed. After the question has been fixed, then the **relevant population** can be identified and hopefully there are data sources which have sampled from it, and if not then an experiment can be designed to sample from it.

For example, if we wanted to investigate the claim that life expectancy in the UK is increasing, then the relevant population is the ages of the people of the UK measured over a period of time. A relevant data source would be one which contains this information. Official national statistics based on UK

census data would be a relevant and high quality data source. Anecdote from an individual in the UK about many generations of their personal family history is also relevant, but lower in quality than official statistics due to its smaller and biased sample.

Data sources can be highly relevant but low quality, and have low relevance but high quality. Both of these properties need to be assessed in order to evaluate the overall value of a data source to a question at hand.

In order to fully assess the quality and relevance of a data source, it needs to be accessible to those assessing it. Maximum accessibility is achieved when data are published as open access, which means that anyone can theoretically gain access. This isn't widely practiced in reality. What often occurs is that access is given to a restricted group of individuals. These individuals might be accredited with a governing body to assure their quality and integrity, or members of a private group (e.g. forensic service providers) who own the data. This involves a level of trust in the accreditation process achieving its aims.

Due to the varied nature of crimes, sometimes the only available data source in a forensic context might be the experience of an examiner. This presents challenges to accessibility, assessibility and quality, but it is the best information that might be available for a particular case. In a criminal case, comments will be made about the expert's conclusions in the their expert witness report. These can then be explored by both parties in pre-trial meetings. Any important unresolved matters can be left for verbal evidence and cross-examination.

4.7 Example: Ogden Tables

Civil cases involving personal injury and fatal accidents will likely have to calculate the financial damages experienced on behalf of the claimant. This means calculating future costs and losses, which would not have been incurred had it not been for the injury in question. Such losses include lost income and pension benefits when employment has been affected, cost of equipment now needed to manage the effects of the injury, etc. Factors which need to be considered in order to value the losses include age-to-retirement of the claimant, life expectancy of the claimant, time for which specialist equipment might be required etc. These calculations are complicated.

Actuarial tables containing values needed to calculate damages were permitted for use in The Civil Evidence Act 1995. The tables are informally known as the Ogden Tables, after the chair of the Working Party who introduced them. For civil disputes of personal injury, the Ogden Tables are the primary data source for calculating damages. The values in them have been calculated by UK government actuarial scientists using current mortality rates and are presented for a range of factors, such as claimant age. This means that the

complicated parts of the damage calculations have already been done by specialists and so in many common claims, only a simple calculation by reading values from the tables is required.

The Ogden tables are an example of a high quality data source. The values contained within its tables are based upon official statistics, which have been verified and agreed by accredited government statisticians. This makes the data underlying the tables accessible and highly validated. The calculations which lead to the values in the table are not openly available but have been performed and agreed by expert committees, this lends them validity too. The data source is representative since the target population are members of the UK and that is whose official data has been used in the calculations for the values in the table.

Official statistics will often be aggregates of many data sources. This means that lots of data has been used even though single number summaries (e.g. medians) are all that are presented. This gives some guarantee that the data source is of an appropriate size. Not much information is presented in the tables about variability, as the values are single number summaries. A level of trust in the process of calculating these values is needed for this. This trust is easier to achieve with this data source due to the recognised and accredited nature of the organisations involved in creating the tables.

The relevance of the Ogden tables depends upon the case at hand. The tables have been constructed to cover a range of common scenarios, e.g. claimant age and some common pension ages. When the claimant in question is covered by these common factors then the tables are a highly relevant (and endorsed) source of data. When the claimant in question is not sufficiently covered by these factors, then the tables cannot be used and are thus not a relevant source of data. This could occur because the claimant holds a more complex pension plan. In this situation the guidance in the is to contact another data source, a Fellow of the Institute and Faculty of Actuaries.

4.8 Expert witness evidence

The statistical ideas presented in this chapter will be present in expert witnesses' reports, but they might not be as explicit as presented here. Below are some of the ways that they might appear in expert witness evidence in criminal cases.

Descriptive statistics can be used to describe evidence recovered as part of the case. This could include a numerical summary of trace evidence, such as the physical characteristics of recovered glass fragments or clothing fibres. The purpose of this could be to generally inform the court, or to provide a comparison benchmark for other findings such as information used from other

data sources. If comparison is the purpose, then these descriptive statistics could be formally compared using inferential statistics.

Inferential statistics can be used when the expert is giving their evaluation of the evidence. This is known as their evaluative opinion. Evaluative opinions take the form of a likelihood ratio (introduced in Section **??**) in some situations, but might also be a qualitative statement. When the opinion, quantitative or not, is based upon probability assignments, then inferential statistics will have been part of that process. This is because some form of observed uncertainty (either recorded samples or recalled by the expert's experience) will have been generalised to describe the epistemic uncertainty. For example, the opinion could use observed frequencies of certain portions of DNA to determine how likely a DNA profiling match would be for a specific recovered DNA profile and a randomly selected individual from a certain population demographic. The opinion might also be that the expert excludes an individual as being the source of a latent fingerprint based on a qualitative assessment of the characteristics of the print and their expert experience. The inference being made in this case is that the individual does not belong to the population of interest.

It can be easy to confuse descriptive and inferential statistics when they are presented as written findings, as this explicit terminology is not used in expert reports. Inferences which are not clear can be clarified during examination of the evidence. For example, consider an expert who has tested a recovered firearm's potential to accidentally misfire due to sudden force triggering a faulty firing mechanism. To test the suspect's statement, the expert attempted to smash the gun against a car window with considerable force in 15 replicated experiments. The faulty firing mechanism triggered the gun to accidentally misfire on the 15th experiment. When the expert reports this it is important to distinguish whether they are only attempting to describe the results of the experiment, or if they are making an implicit inference that the gun has a misfire rate (probability of misfiring) of 1 in 15 when making forceful contact with a surface. The latter is a statistical inference and it should be questioned whether the inference is sufficiently substantiated by the experiment that the expert performed.

Data sources will frequently be used in expert reports. Common ones include recovered evidence, in-house databases containing information which is relevant to forensic investigations, journal articles which can report case information or findings from controlled experiments, and details of the expert's recalled experience. These data sources can be explored by following the criteria from Section **??** relating to quality and relevance.

The previous two chapters have given a conceptual overview of the mathematical theory required to begin considering case situations. These situations require more than just numerical theory as they also involve investigative information too. In the analysis of forensic scientific evidence, this means merging probability and statistics with contextual case information,

including statements from both the prosecution and defence. The prosecution and defence's proposed versions of events are required to fully assess evidence that has been recovered. These statements, which are the disputed aspects of the case, are formalised into what are known as **propositions**. Propositions are used as anchors to focus the interpretation of the evidence, and they are the topic of the next chapter.

4.9 More information

4.10 Exercises

Chapter 5

Propositions to frame uncertainty

This Chapter introduces propositions

5.1 Facts-in-issue

To motivate the use of propositions, we begin with a discussion about facts-in-issue. Facts-in-issue are the relevant facts of a specific case which are contested by each side of the legal dispute. In criminal trials, this typically manifests as the defence contesting facts which the prosecution have asserted.

There is epistemic uncertainty about the facts-in-issue and so relevant evidence is presented to the court. The fact finder then considers this evidence to determine the truth of the facts-in-issue, and ultimately the outcome of the case.

It is clear from this that the language and presentation of facts-in-issue must be precise and accurate. The presentation should also be done in a way that allows the fact finder to clearly discriminate between what either side of counsel is asserting. To achieve this, forensic scientists formalise the facts-in-issue into what are known as propositions.

5.2 Propositions

A proposition is an assertion about a factual state of nature. The 'factual' part of that is key since it means that the assertion is capable of being either true or false. Thus, propositions are capable of either being true or false.

Although not always correct, we will speak of propositions as not only **capable of being** either true or false, but **as** being either true or false.

The truth of propositions does not depend upon people's belief about them (unless the proposition is specifically about their belief). Consider the following:

1. The Earth orbits the Sun.
2. The Earth is flat.

Proposition 1 is true and proposition 2 is false, and we have substantial scientific evidence to prove this. At a certain point of time in history, however, common belief about those propositions was false: people thought that the Sun orbited the Earth, and that the Earth was flat (and some still do…). It was true that the Earth orbits the Sun both before and after humans had enough evidence to believe it was true.

Propositions can still evoke epistemic uncertainty; they are true or false but we might not know which. In this case we can probabilistically reason about them. Consider another example:

3. Poker Player B is holding an ace in this round.

In this hypothetical poker game, it's self-evident that the proposition is true if Player B is holding an ace and it is false if Player B is not holding an ace. If you are Player B then you will know the truth of this and the proposition is trivial. However, if you are not Player B then you won't know the truth of this; you have epistemic uncertainty.

To play your hand well, you will have to come to a reasonable belief about the probability of this proposition (and others) and make decisions about checking, raising, or folding etc., whilst accounting for this uncertainty. Note that in this example you might **never** know the truth of this proposition. Player B might never turn over their cards nor reveal the answer. The proposition was still either true or false regardless of this.

The truth of some propositions may change as situations change. Consider the following:

4. Boris Johnson's Conservative party govern the UK.

As of me writing this, the proposition is true. The UK still exists, the Conservative party exist under that same name and are led by Boris Johnson, and his party governs the UK after a landslide general election win. This proposition won't be true forever, and may not even be true by the time you're reading this. It can help to be more specific when formulating propositions to overcome this dependence on time or other factors. Changing the proposition instead to:

5. On 20th May 2020, Boris Johnson's Conservative party govern the UK

gives a more specific proposition that is now unambigious about the temporal factor which may affect it. However, the proposition is now less general since it only refers to one specific day. Whether this lack of generality is worth the extra specificity or not depends on the question at hand. When the question is provided by investigators and given to forensic scientists, then the scientists have to balance these factors when formulating appropriate propositions.

5.3 Competing propositions

When formulating propositions from facts-in-issue, the different sides of the adversarial system make different assertions about the circumstances which lead to the same factual observation.

For example, as part of a case there might be a DNA sample at the crime scene which was found to match the DNA profile taken from the suspect after their arrest. The prosecution might assert that the defendant was at the location in which the DNA sample was recovered. The defence, on the other hand, might accept that this is the defendant's DNA but instead assert that it was indirectly transferred there, e.g. from contact with another individual who was provably at the crime scene. Both of these assertions would reasonably lead to the defendant's DNA matching the crime scene sample.

Translating these assertions into propositions gives so-called **competing propositions**. Competing propositions typically come in pairs, one put forward by each side of legal counsel. They are often denoted as H_p and H_d to represent 'prosecution/defence hypothesis', and we'll use this from now on.

Case scenario: Police were called to the house of a known local drug-dealer after reports by neighbours of a suspected gunshot heard from within. They arrived at the scene to find a middle-aged male, the known drug-dealer, shot dead in a bedroom. A window at the back of the house on the ground floor was found to be smashed, with the glass having fallen towards the inside of the house. They found open death-threats on the deceased's public social media account from two males. The males are known to police, having previously been involved in drug-related crimes too. Both are arrested at their known addresses, at which time Suspect 1 confesses to having broken in to the house but not having fired the weapon. He refuses to name any accomplices. Suspect 2 refuses to comment, despite glass-covered clothing being found within a plastic bag in a bin at his residence. After further testing, this clothing was also found to be covered in gunshot-residue particles. No weapon is recovered.

Here are three facts-in-issue of the (probably) many for this case at this preliminary stage.

Fact-in-issue 1: whether or not Suspect 2 killed the victim with the intent to kill or cause grievous bodily harm.

- H_p^1: Suspect 2 committed murder,
- H_d^1: Suspect 2 did not commit murder,

Fact-in-issue 2: whether or not Suspect 2 fired the gun in question

- H_p^2: Suspect 2 fired the gun,
- H_d^2: Suspect 2 did not fire the gun,

Fact-in-issue 3: whether or not the glass fragments found on Suspect 2's clothing originated from the smashed window

- H_p^3: The glass fragments originate from the smashed window,
- H_d^3: The glass fragments originate from a source other than the smashed window.

We have extracted the information from the case scenario and formulated potential facts-in-issue. These facts-in-issue isolate individual issues that appear in the case scenario. They also make a clear statement about what is the potential issue which surrounds each fact, e.g. was it or was it not the suspect?

It is clear that for each fact-in-issue, we have created a potential pair of competing propositions. Each proposition makes a direct assertion about the associated fact-in-issue, considers each side of the issue (prosecution or defence) separately, and is either true or false. Using this process reduces complex case scenarios into individual unambiguous points of focus, around which the assessment of evidence can be anchored.

There are some other key properties of these propositions which are worth exploring. Let's revisit the following competing pair of propositions:

- H_p^2: Suspect 2 fired the gun,
- H_d^2: Suspect 2 did not fire the gun.

One property of these propositions is that when one is true, the other is necessarily false (and vice versa). They cannot both be true. This is known as **mutual exclusivity**. Mutual exclusivity occurs when two propositions make assertions which have no logical overlap. Suspect 2 firing the gun and not firing the gun cannot both be true, they are completely separable assertions.

Another property of these propositions is that either one or the other must be true – there is no scenario in which both are false. This is known as

exhaustivity. Exhaustivity means that the propositions cover the entire set of all possibilities for the event in question. We know that a gun was fired, so either Suspect 2 fired the gun or they did not. The proposition H_p^2 covers one specific event and H_d^2 covers everything else.

When competing propositions are exhaustive and mutually exclusive then they cover all possible events, meaning that one of them must be true, but they cannot both be true. This means that one of the propositions must be true and the other must be false. This 'either one or the other' property makes such competing propositions easier to evaluate, since knowing the truth of one determines the truth of the other.

These propositions were guaranteed to be mutually exclusive and exhaustive because of the way we constructed them. First we specified H_p. Then, we **negated** H_p to get H_d in each pair of propositions. A simpler way of phrasing this is to say we formed 'H_p' and 'not H_p'. Take the glass propositions as an example:

- H_p^3: The glass fragments originate from the smashed window,
- H_d^3: The glass fragments originate from a source other than the smashed window.

If H_p^3 is not true, then the glass fragments must originate from a source other than the smashed window, and so we get H_d^3. This makes the propositions necessarily mutually exclusive, 'H_p^3' and 'not H_p^3' cannot both be true because they are logical negations. It also means that they are exhaustive, since H_p^3 or its negation ('anything else') must be true.

The logic of negating propositions is in line with the burden of proof in UK criminal cases. It is for the prosecution to prove beyond reasonable doubt that their assertions are true. The defence, in proving reasonable doubt about the prosecution assertions need only show that the opposite of the prosecution story has reasonable credibility. Negating prosecution propositions is one logical way of doing this.

There are plenty of situations in which simply negating the prosecution propositions is inappropriate however, e.g. when the defendant has a genuine alternative narrative for events. In this situation, the defence narrative might be used for H_d, e.g.:

- $H_d^{3^*}$: the glass fragments originate from a glass barrier at a shooting range that Suspect 2 has recently been performing construction work on.

This proposition is still competing with H_p^3, and H_p^3 and $H_d^{3^*}$ are still mutually exclusive. However, they are no longer exhaustive. This is because all possible alternatives to H_p^3 are no longer considered by $H_d^{3^*}$. In other words, it is now

possible for both H_p^3 and $H_d^{3^*}$ to be false. This could happen if the suspect lies or is mistaken about their alternative version of events.

These properties of propositions are important when we begin to consider probabilities associated with them in Chapter [...]. But first we will look at different categories of proposition.

5.4 Hierarchy of propositions

It's helpful to further categorise propositions based upon the type of assertion that they make. This categorisation makes the proposition's claims explicit and clear, which helps us understand how the proposition can help to answer a question.

Propositions are mostly grouped into four categories (or levels): offence, activity, source, and sub-source. This is known as the **hierarchy of propositions**, since each category is generally more or less important than the others when considering the case as a whole. Examples of competing propositions for each of these levels are shown in Table **??** below.

Offence level propositions directly refer to a criminal offence. From Table **??** we have:

- H_p^1: The defendant committed murder,
- H_d^1: The defendant did not commit murder.

We saw these propositions in the case scenario of the previous Section. These competing propositions refer to the **ultimate issue** in the case. Others, such as H_p^2 and H_d^2 might not be the ultimate issue of the case, but will still determine the fate of the defendant. For this reason, they are the most important level of proposition.

Offence level propositions are considered by the fact finder in a case and rarely the forensic scientist. This is because propositions about criminal offences fall outside the domain of an scientific expert witness, who gives evidence only on scientific findings within a case. There will be other non-scientific evidence in the case that will need to be considered too. The scientist can indirectly assist in finding the truth of offence level propositions through assisting the court on other related propositions which do fall within their expert domain.

Activity level propositions relate to activies and do not assert criminal liability. This leaves them one step below the offence-level propositions in terms of importance to the court; the activity may have been criminal but it also may not have been. In Table **??** we have:

- H_p^4: The defendant kicked the claimant,

Table 5.1: Propositions levels and some examples of competing propositions.

Level	Examples
Offence	H_p^1: The defendant committed murder
	H_d^1: The defendant did not commit murder
	H_p^2: The defendant assaulted the claimant
	H_d^2: The defendant did not assault the claimant
Activity	H_p^3: The defendant fired the gun
	H_d^3: The defendant did not fire the gun
	H_p^4: The defendant kicked the claimant
	H_d^4: The defendant was not present when the claimant was kicked
Source	H_p^5: The glass from the defendant's clothing originated from the smashed window
	H_d^5: The glass from the defendant's clothing originated from some other source
	H_p^6: The blood on the defendant's shoes came from the claimant
	H_d^6: The blood on the defendant's shoes came from someone else
Sub-source	H_p^7: NA
	H_d^7: NA
	H_p^8: The DNA on the claimant's clothing came from the defendant
	H_d^8: The DNA on the claimant's clothing came from someone else

- H_d^4: The defendant was not present when the claimant was kicked.

If H_p^4 is true, then this still may have been done within the law. However, if H_p^4 is proven true to the fact finder, then it is a large step closer to answering the ultimate issue – all that remains is to show that the activity amounted to assault, thus addressing H_p^2. This means that activity level propositions still have a high degree of importance.

Forensic scientists do consider activity level propositions where scientific findings can assist the court on them. In fact, this is the advised level of proposition where it is possible. In order to consider activity level propositions, it is often necessary for the scientist to take in to account contextual information from case circumstances as well as the available scientific findings. If the necessary contextual information is unavailable, then the scientist may find themselves unable to assist the court regarding alleged activities.
Consider the propositions:

- H_p^3: The defendant fired the gun,
- H_d^3: The defendant did not fire the gun.

Finding gunshot residue particles on the defendant's clothing might not be enough information for the scientist to assist the court on these propositions. They would have to consider other explanations for the defendant having gunshot residue on their clothes first. If the defendant was in the military, part of a shooting club, or had any other reasonable explanation for having this residue on their clothing then the scientist would need to know. They might then require further information about the last time the defendant was in contact with firearms. The amount of particles found on the clothing, and the location in which they were found (e.g. cuffs), might then be considered in light of the defendant's response. Results from the scientific literature about the background abundance and transfer and persistence of gunshot residue particles might then need to be used in addition to the scientific findings from the case.

[add a picture to the above pls, it is sooo dry]

Source level propositions relate to the origin of traces and impressions found, e.g. at a crime scene. The source level makes attributions such as blood to an individual, fibres to a garment of clothing, footwear marks to a piece of footwear, etc. These propositions are relatively simple compared to activity level since they do not comment on questions such as "how?" or "why?" the traces and impressions may have come to be where they were recovered from. The answer to these questions are often the most consequential because they are logically closer to the ultimate issue. Since the activity level addresses these questions and the source level does not, the source level is placed below the activity level in the hierarchy.

Consider the competing propositions from Table **??**:

- H_p^5: The glass from the defendant's clothing originated from the smashed window,
- H_d^5: The glass from the defendant's clothing originated from some other source.

These propositions have been derived from evidence that has been found. This is a common situation for source level propositions: trace evidence is found and we need to know its origin. It is also very common for forensic scientists to consider source level propositions. This is because the science is well-equipped to shed light on them. For H_p^5 above, the scientist can compare analytic measurements from the recovered glass fragments to other fragments known to come from the smashed window. For H_d^5, the scientist will need to compare analytic measurements of the recovered fragments to fragments from other known sources.

Although the source level propositions are below the activity level in the hierarchy, this is not to say that they are unimportant. In fact, in individual cases, they may be the most important as they might effectively address the most important activity level propositions. For example, if the fact-finder decides that H_p^5 is true based on the expert's evidence, then they may not require further evidence to think that the following activity level proposition is true:

- H_p^{5+}: the defendant was present when the glass was smashed.

In other words, the truth of the source level proposition could have a logical domino effect that leads to the fact-finder being sure of the truth of the most important propositions.

The **sub-source level** is a particular version of the source level. It makes source attributions of more general traces, such as DNA to an individual. The difference between source and sub-source is that the source level states the specific cellular material (e.g. blood) that DNA would be recovered from, whereas sub-source simply states that DNA was found. This is the result of highly sensitive modern measurement technologies. Taking swabs of items of interest at a crime scene can detect very low levels of DNA even though there was no obvious biological material that was swabbed.

This distinction can be key in case circumstances because it leaves open questions of how the DNA came to be there. The sub-source level propositions in Table **??** were:

- H_p^8: The DNA on the claimant's clothing came from the defendant,
- H_d^8: The DNA on the claimant's clothing came from someone else.

DNA found on clothing may be explained by direct contact and indirect transfer, which might be reasonable in any given case. This means that more

obstacles are encountered than source level propositions when trying to move to the activity level. If blood has been found on an item then there are far fewer mechanisms of it coming to be there compared to finding DNA alone.

The hierarchy of propositions and the degree to which the scientist can aid the court in addressing them using science alone is represented in the Figure [...].

It's worth mentioning again that these are conceptual frameworks with which to consider propositions. Individual case circumstances will determine how important each level of proposition is and the degree to which the scientist can consider them using science alone.

What role does evidence play here? We have seen that source and sub-source level propositions are formed from evidence and that there is a logical relationship between all levels of proposition. In the next section, we will add evidence into the mix.

5.5 Example: proposition levels

Practice identifying the correct proposition level with the examples below.

5.6 Probative value of evidence

The **probative value** of evidence is the degree to which a piece of evidence can discriminate between competing propositions. If a piece of evidence cannot discriminate between propositions then it is not probative for those propositions. But this does not mean it is not probative of any propositions. Consider the following propositions from Table **??**:

- H_p^5: The glass from the defendant's clothing originated from the smashed window,
- H_d^5: The glass from the defendant's clothing originated from some other source,
- H_p^6: The blood on the defendant's shoes came from the claimant,
- H_d^6: The blood on the defendant's shoes came from someone else,

and the following three pieces of evidence:

- E_1: glass analysis reveals a "match" between the fragments found on the defendant's clothing and the smashed window,
- E_2: DNA analysis of the blood on the defendant's shoes does not match the claimant,

- E_3: gun shot residue particles were found on the defendant's clothing.

Clearly evidence E_1 is highly probative of H_p^5/H_d^5. Intuitively, we expect a match if H_p^5 is true. We might also come across a match if H_d^5 is true. This could occur when the same type of glass has been used for both the smashed window and the other source, but we might reasonably expect this to be less likely than if the glass fragments really did come from the smashed window. One factor which will affect our expectation is how common that type of glass is in other uses, e.g. are all household windows made from that type?

Evidence E_2 is not just probative of H_p^6/H_d^6, but it is almost completely determinative of it. We absolutely do not expect a "no match" if the claimant truly is the source of the blood; H_p^6 is incompatible with the evidence. The only situation in which this is not true is if there has been a mistake with handling and analysing the sample. With sufficient evidence gathering and analysing protocols in place, it is assumed that this is not the case unless there is reasonable evidence to suggest otherwise. On the other hand, we absolutely expect "no match" if the claimant is not the source, i.e. if H_d^6 is true.

Evidence E_3 is not probative of either pair of competing propositions with the lack of contextual information that we have given here. However, it's not hard to imagine that it could be highly probative of the propositions about firing a gun, e.g. H_p^3/H_d^3, if case circumstances reflected this.

As soon as forensic evidence has been recovered then it is trivially certain. The uncertainty that does arise comes from assessing the circumstances which may have led to the evidence being left where it was found, or the origin of the trace/impression which was recovered. This means considering the evidence in light of the competing propositions. We did this qualitatively in the discussion above with E_1, E_2, and E_3 and the propositions H_p^5/H_d^5 and H_p^6/H_d^6. Since the truth of the propositions is not known by the those investigating, there is uncertainty. And since there is uncertainty, it can be described quantitatively using the language of probability.

Repeating the points above, we can say that the probability of E_1, E_2, and E_3 is 1; since we observed it we know it to be certain. But what we really need is to condition the observed evidence E_1, E_2, and E_3 on the propositions that counsel have put forward H_p^5/H_d^5 and H_p^6/H_d^6. This will help the fact finder determine the truth of the propositions. The probabilities for the evidence change when we condition on propositions as they reflect our uncertainty about observing the evidence under these circumstances. Let's revisit our previous qualitative analysis.

What's the probability of E_1 assuming that H_p^5 is true? It is 1, we would consider it certain to obtain a match if the recovered fragments do come from the smashed window. What's the probability of E_1 assuming that H_d^5 is true? This is uncertain. We said that this would depend on factors such as how common the glass type of the smashed window was in other sources of glass

that the suspect might reasonably have come into contact with. The ability of the evidence E_1 to discriminate between H_p^5/H_d^5 depends on the probability of E_1 assuming that H_d^5 is true. If this probability is low then it can discriminate well between the propositions because E_1 would be far more likely assuming H_p^5 were true rather than if H_d^5 were true. If the probability is high (near to 1), then it cannot discriminate well between them because the probabilities are of similar magnitude and so the evidence is similarly likely assuming each proposition to be true. In other words, E_1 could have high or low (or anywhere inbetween) probative value.

What's the probability of E_2 assuming that H_p^6 is true? It is 0, with current DNA technology for single donor full profiles, we consider it impossible to obtain a "no match" if the claimant is the origin of the blood. What's the probability of E_2 assuming that H_d^6 is true? It is 1 or near to 1. To rephrase this, the evidence E_2 is only possible if H_d^6 were true (ignoring mistakes in handling evidence), and so the evidence has extremely high probative value in favour of H_d^6 compared to H_p^6.

What's the probability of E_3 assuming that H_p^6 is true? It is 1. This is because, given the current contextual information, H_p^6 has nothing to do with gun shot residue. It is reasonable to assume (although this might have to be proven in a real case) that they are statistically independent and so the probability of E_3 assuming that H_p^6 is true is the same as the probability of E_3 disregarding H_p^6, which is 1 because we observed it. Conditioning on an irrelevant proposition tells us nothing that we didn't already know. The same argument can be applied to H_d^6. The evidence E_3 cannot discriminate between H_p^6/H_d^6 and so it does not have probative value for these propositions.

The preceding analysis should motivate that propositions are the key elements which define the probative value of evidence. The same piece of evidence can have no probative value for one set of competing propositions but may be highly probative for another. We have shown that using the language of probability, we can quantify the uncertainty surrounding evidence conditional on propositions. The natural question is then, can we quantify probative value? Yes. This is the topic of the next Chapter, the Likelihood Ratio.

5.7 Changing propositions

It goes without saying that since propositions determine the probative value of a piece of evidence, then changes in propositions can affect probative value. This might seem like an obvious statement to some, but it is an important one to highlight because changes in propositions are common and sometimes subtle. The resulting change in probative value might not be as subtle.

As a case develops and investigations continue, new information comes to light and previous information might update. Changes to propositions which are

considered by the scientist reflect this, which is why there is always an explicit statement about it that features in written expert witness reports: "these conclusions are based on current information and may change if new information becomes available."

One common example of this is if the defendant changes their version of events or reveals new relevant activities of theirs. The source level glass propositions H_p^5/H_d^5 might be the default ones considered by the scientist with a lack of other relevant contextual information. If the suspect then states in interview that they have only worn the same clothing recently once before and it was when they were working on a construction site replacing windows, then H_d^5 should be revised to reflect this. The proposition might change to:

- $H_d^{5^*}$: The glass from the defendant's clothing originated from the windows on the construction site.

This might be considered a small alteration to the proposition but it could completely diminish the probative value of the glass match E_1. Suppose that the glass type of the smashed window at the crime scene is very rare, it is only used for house windows of a very specific style. That means that when compared to all other reasonable sources of glass, this type of glass is found rarely. The probability of E_1 assuming H_d^5 is low in this situation. But, suppose that the defendant was fitting windows of this same style. Now, it would not be a rare occurrence at all. The probability of E_1 assuming $H_d^{5^*}$ to be true would be very high, maybe even 1. The evidence has no probative value with the new contextual information and propositions $H_p^5/H_d^{5^*}$. This represents a change in the reference class, which we saw back in Chapter [...].

[should be explicit about information I]

5.8 Probabilities for propositions

In this Chapter we have been very careful to refer to **probabilities for evidence**. We reasoned that probative value is related to probabilities for evidence assuming, in turn, competing propositions to be true and comparing them to each other. We explicitly did not state **probabilities for propositions**, and particularly probabilities for competing propositions in light of the available evidence.

This is because it is not currently within the role of the expert witness (in the UK) to do so. The expert witness uses their scientific expertise to comment on the likelihood of observing particular pieces of evidence in light of proposed versions of events. It is then within the fact finder's remit to use this information to determine the truth of propositions themselves. In other words, it is up to the fact finder to reason about their own uncertainty for the propositions in light of the evidence they see put before the court.

If the fact finder's uncertainty is reduced to "beyond reasonable doubt", for what the fact finder decides are the key propositions, then they should return a verdict in favour of the prosecution. If the same uncertainty is not reduced beyond reasonable doubt, then the verdict should be in favour of the defense. Whether the fact finder decides to reason about this uncertainty quantitatively using probability theory or otherwise is up to them.

To hammer home the point, in this book: the **expert** considers **probabilities for evidence**, the **fact finder** considers **probabilities for propositions**. In the discussions following this Chapter, probabilistic statements that we claim come from an expert will only be about observed evidence.

5.9 More information

5.10 Exercises

Chapter 6

Likelihood ratio to evaluate uncertainty

This Chapter introduces the likelihood ratio

6.1 Relative support for competing propositions

Suppose that we have two competing propositions, A and B, as in the previous chapter. We observe an event E, and wish to know whether E was more likely assuming A or B. If E is more likely assuming A was true, then we say that E provides more support for A than B, and vice versa if E is more likely assuming B to be true. The likelihood ratio (LR) quantifies this support.

The LR consists of the probability of observing E conditioned on A being true, divided by the probability of observing E conditioned on B being true. As a formula it is written as

$$\text{LR} = \frac{\text{probability of E assuming A is true}}{\text{probability of E assuming B is true}}.$$

The trick here is that by assuming each proposition to be true in turn, we can see how much more likely E was to occur in A's version of events compared to B's.

Since each term in the LR is a probability, their value must lie between 0 and 1. This means that the LR itself must be between 0 and ∞. Values of the LR which are greater than 1 indicate relative support for A compared to B, since it means that the probability of E assuming that A is true is greater than the