A Student's Guide to
BAYESIAN STATISTICS

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

A Student's Guide to
BAYESIAN STATISTICS

Ben Lambert

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver’s Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Jai Seaman
Editorial assistant: Alysha Owen
Production editor: Ian Antcliff
Copyeditor: Sarah J. Duffy
Proofreader: Neville Hankins
Marketing manager: Susheel Gokarakonda
Cover design: Bhairvi Gudka
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

First published 2018

Library of Congress Control Number: 2017942214

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-4739-1635-7
ISBN 978-1-4739-1636-4 (pbk)

At SAGE we take sustainability seriously. Most of our products are printed in the UK using responsibly sourced papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

2

Chapter contents

THE SUBJECTIVE WORLDS OF FREQUENTIST AND BAYESIAN STATISTICS

2.1 CHAPTER MISSION STATEMENT

At the end of this chapter, the reader will understand the purpose of statistical inference, as well as recognise the similarities and differences between Frequentist and Bayesian inference. We also introduce the most important theorem in modern statistics: Bayes' rule.

2.2 CHAPTER GOALS

As data scientists, we aim to build predictive models to understand complex phenomena. As a first approximation, we typically disregard those parts of the system that are not directly of interest. This deliberate omission of information makes these models *statistical* rather than deterministic because there are some aspects of the system about which we are uncertain. There are two distinct approaches to statistical modelling: Frequentist (also known as Classical inference) and Bayesian inference. This chapter explains the similarities between these two approaches and, importantly, indicates where they differ substantively.

Usually, it is straightforward to calculate the probability of obtaining different data samples if we know the process that generated the data in the first place. For example, if we know that a coin is fair, then we can calculate the probability of it landing heads up (the probability equals 1/2). However, we typically do not have perfect knowledge of these processes, and it is the goal of statistical inference to derive estimates of the unknown characteristics, or *parameters*, of these mechanisms. In our coin example, we might want to determine its bias towards heads on the basis of the results of a few coin throws. Bayesian statistics allows us to go from what is known – the *data* (the results of the coin throw here) – and extrapolate backwards to make probabilistic statements about the parameters (the underlying bias of the coin) of the processes that were responsible for its generation. In Bayesian statistics, this inversion process is carried out by application of Bayes' rule, which is introduced in this chapter. It is important to have a good understanding of this rule, and we will spend some time throughout this chapter and Part II developing an understanding of the various constituent components of the formula.

2.3 BAYES' RULE – ALLOWING US TO GO FROM THE EFFECT BACK TO ITS CAUSE

Suppose that we know that a casino is crooked and uses a loaded die with a probability of rolling a 1, that is $\frac{1}{3} = 2 \times \frac{1}{6}$, twice its unbiased value. We could then calculate the probability that we roll two 1s in a row:

$$Pr(1,1\,|\,\text{crooked casino}) = \frac{1}{3} \times \frac{1}{3} \;\; = \frac{1}{9}. \tag{2.1}$$

Here we use *Pr* to denote a probability, with the comma here having the literal interpretation of *and*. Hence, *Pr*(1, 1) is the probability we obtain a 1 on the first roll *and* a 1 on the second. (Don't worry if you don't fully understand this calculation, as we will devote the entirety of the next chapter to working with probabilities.) In this case, we have presupposed a cause – the casino being crooked – to derive the probability of a particular effect – rolling two consecutive 1s. In other words, we have calculated $Pr(\text{effect} \mid \text{cause})$. The vertical line, $\mid$, here means *given* in probability, so $Pr(1, 1 \mid \text{crooked casino})$ is the probability of throwing two consecutive 1s given that the casino is crooked.

Until the latter half of the seventeenth century, probability theory was chiefly used as a method to calculate gambling odds, in a similar vein to our current example. It was viewed as a dirty subject, not worthy of the attention of the most esteemed mathematicians. This perspective began to change with the intervention of the English Reverend Thomas Bayes, and slightly later and more famously (at the time at least), with the work done by the French mathematician Pierre Simon Laplace (see 'Bayes' rule or the Bayes–Price–Laplace rule?' below for a short history of Bayes' rule). They realised that it is possible to move in the opposite direction – to go from effect back to cause:

$$Pr(\text{effect} \mid \text{cause}) \xrightarrow{\text{Bayes' theorem}} Pr(\text{cause} \mid \text{effect}). \tag{2.2}$$

In order to take this leap, however, it was necessary to discover a rule, which later became known as Bayes' rule or theorem. This can be written:

$$Pr(\text{cause} \mid \text{effect}) = \frac{Pr(\text{effect} \mid \text{cause}) \times Pr(\text{cause})}{Pr(\text{effect})}. \tag{2.3}$$

In the casino example, this formula tells us how to invert the original probability $Pr(1, 1 \mid \text{crooked casino})$ to obtain a more useful quantity as a patron of said casino –$Pr(\text{crooked casino} \mid 1, 1)$. In words, this is the probability that the casino is crooked *given* that we rolled two 1s. We do not show how to carry out this calculation now, and instead delay this until we learn about probability in Chapter 3. However, this process where we go from an effect back to a cause is the essence of inference. Bayes' rule is central to the Bayesian approach to statistical inference. Before we introduce Bayesian inference, though, we first describe the history of Bayes' rule.

Bayes' rule or the Bayes–Price–Laplace rule?

In 1748, the Scottish philosopher David Hume dealt a serious blow to a fundamental belief of Christianity by publishing an essay on the nature of cause and effect. In it, Hume argues that '*causes and effects are discoverable, not by reason, but by experience*'. In other words, we can never be certain about the cause of a given effect. For example, we know from experience that if we push a glass off the side of a table, it will fall and shatter, but this does not prove that the push caused the glass to shatter. It is possible that both the push and the shattering are merely correlated events, reflecting some third, and hitherto unknown, ultimate cause of both. Hume's argument was unsettling to Christianity because God was traditionally known as the First Cause of everything. The mere fact that the world exists was seen as evidence of a divine creator that caused it to come into existence. Hume's argument meant that we can never deal with *absolute* causes; rather, we must make do with *probable* causes. This weakened the link between a divine creator and the world that we witness and, hence, undermined a core belief of Christianity.

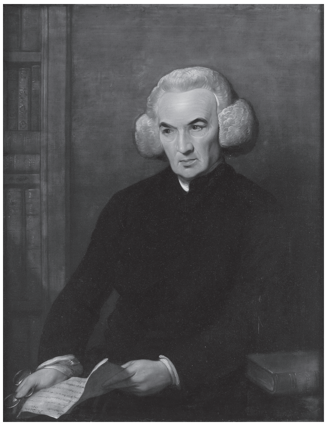
(Continued)

Around this time the Reverend Thomas Bayes of Tunbridge Wells (where this book's author grew up!) began to ponder whether there might be a mathematical approach to cause and effect.

Thomas Bayes was born around 1701 to a Presbyterian minister, Joshua Bayes, who oversaw a chapel in London. The Presbyterian Church at the time was a religious denomination persecuted for not conforming to the governance and doctrine of the Church of England. Being a *non-conformist*, the young Bayes was not permitted to study for a university degree in England and so enrolled at the University of Edinburgh, where he studied theology. After university, Bayes was ordained as a minister of the Presbyterian Church by his clergyman father and began work as an assistant in his father's ministry in London. Around 1734, Bayes moved south of London to the wealthy spa resort town of Tunbridge Wells and became minister of the Mount Sion chapel there.

Bayes: c.1701–1761

Around this time, Bayes began to think about how to apply mathematics, specifically probability theory, to the study of cause and effect (perhaps invigorated by the minerals in the spa town's cold water). Specifically, Bayes wanted a mathematical way to go from an effect back to its cause. To develop his theory, he proposed a thought experiment: he imagined attempting to guess the position of a ball on a table. Not perhaps the most enthralling of thought experiments, but sometimes clear thinking is boring. Bayes imagined that he had his back turned to the table, and asks a friend to throw a cue ball onto its surface (imagine the table is big enough that we needn't worry about its edges). He then asks his friend to throw a second ball, and report to Bayes whether it landed to the left or right of the first. If the ball landed to the right of the first, then Bayes reasoned that the cue ball is more likely to be on the left-hand side of the table, and vice versa if it landed to its left. Bayes and his friend continue this process where, each time, his friend throws subsequent balls and reports which side of the cue ball his throw lands. Bayes' brilliant idea was that, by assuming all positions on the table were equally likely a priori, and using the results of the subsequent throws, he could narrow down the likely position of the cue ball on the table. For example, if all throws landed to the left of the cue ball, it was likely that the cue ball would be on the far right of the table. And, as more data (the result of the throws) was collected, he became more and more confident of the cue ball's position. He had gone from an effect (the result of the throws) back to a probable cause (the cue ball's position)!



Price: 1723–1791

Bayes' idea was discussed by members of the Royal Society, but it seems that Bayes himself perhaps was not so keen on it, and never published this work. When Bayes died in 1761 his discovery was still languishing between unimportant memoranda, where he had filed it. It took the arrival of another, much more famous, clergyman to popularise his discovery.

Richard Price was a Welsh minister of the Presbyterian Church, but was also a famous political pamphleteer, active in liberal causes of the time such as the American Revolution. He had considerable fans in America and communicated regularly with Benjamin Franklin, John Adams and Thomas Jefferson. Indeed, his fame and adoration in the United States reached such levels that in 1781, when Yale University conveyed two degrees, it gave one to George Washington and the other to Price. Yet today, Price is primarily known for the help that he gave his friend Bayes.

When Bayes died, his family asked his young friend Richard Price to examine his mathematical papers. When Price read Bayes' work on cause and effect he saw it as a way to counter Hume's attack on causation (using an argument not dissimilar to the Intelligent Design hypothesis of today), and realised it was worth publishing. He spent two years working on the manuscript – correcting some mistakes and adding references – and eventually sent it to the Royal Society with a cover letter of religious bent. Bayes for his (posthumous) part of the paper did not mention religion. The Royal Society eventually published the manuscript with the secular title, 'An Essay towards solving a Problem in the Doctrine of Chances'. Sharon McGrayne – a historian of Bayes – argues that, by modern standards, Bayes' rule should be known as the Bayes–Price rule, since Price discovered Bayes' work, corrected it, realised its importance and published it.

Given Bayes' current notoriety, it is worth noting what he did not accomplish in his work. He did not actually develop the modern version of Bayes' rule that we use today. He just used Newton's notation for geometry to add and remove areas of the table. Unlike Price, he did not use the rule as proof for God, and was clearly not convinced by his own work since he failed to publish his papers. Indeed, it took the work of another, more notable, mathematician to improve on Bayes' first step, and to elevate the status of inverse probability (as it was known at the time).

Laplace: 1749–1827

Pierre Simon Laplace was born in 1749 in Normandy, France, into a house of respected dignitaries. His father, Pierre, owned and farmed the estates of Maarquis, and was Syndic (an officer of the local government) of the town of Beaumont. The young Laplace (like Bayes) studied theology for his degree at the University of Caen. There, his mathematical brilliance was quickly recognised by others, and Laplace realised that maths was his true calling, not the priesthood. Throughout his life, Laplace did important work in many fields including analysis, differential equations, planetary orbits and potential theory. He may also have even been the first person to posit the existence of black holes – celestial bodies whose gravity is so great that even light can't escape. However, here, we are most interested in the work he did on inverse probability theory.

Independently of Bayes, Laplace had already begun to work on a probabilistic way to go from effect back to cause, and in 1774 published 'Mémoire sur la probabilité des causes par les évènemens', in which he stated the principle':

> *Si un évènement peut être produit par un nombre n de causes différentes, les probabilités de lexistence de ces causes prises de évènement, sont entre elses comes les probabilités de l'évènement prises de ces causes, et la probabilité de lexistence de chacune delles, est égale á la probabilité de l'évènement prise de cette cause, diviseé par la somme de toutes les probabilités de l'évènement prises de chacune de ces causes.*

This translates as (Laplace (1986)):

> If an event can be produced by a number n of different causes, then the probabilities of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given the cause, divided by the sum of all the probabilities of the event given each of these causes.

(Continued)

This statement of inverse probability is only valid when the causes are all equally likely. It was not until later than Laplace generalised this result to handle causes with different prior weights.

In 1781, Price visited Paris and told the Secretary of the French Royal Academy of Sciences, the Marquis of Condorcet, about Bayes' discovery. This information eventually reached Laplace and gave him confidence to pursue his ideas in inverse probability. The trouble with his theory for going from an effect back to a cause was that it required an enormous number of calculations to be done to arrive at an answer. Laplace was not afraid of a challenge, however, and invented a number of incredibly useful techniques (for example, generating functions and transforms) to find an approximate answer. Laplace still needed an example application of his method that was easy enough for him to calculate, yet interesting enough to garner attention. His chosen data sample was composed of babies. Specifically, his sample comprised the numbers of males and females born in Paris from 1745 to 1770. This data was easy to work with because the outcome was binary – the child was recorded as being born a boy or girl – and was large enough to be able to draw conclusions from it. In the sample, a total of 241,945 girls and 251,527 boys were born. Laplace used this sample and his theory of inverse probability to estimate that there was a probability of approximately 10^{-42} that the sex ratio favoured girls rather than boys. On the basis of this tiny probability, he concluded that he was as 'certain as any other moral truth' that boys were born more frequently than girls. This was the first practical application of Bayesian inference as we know it now. Laplace went from an effect – the data in the birth records – to determine a probable cause – the ratio of male to female births.

Later in his life, Laplace also wrote down the first modern version of Bayes' mathematical rule that is used today, where causes could be given different prior probabilities. He published it in his "Théorie analytique des probabilités" in 1820 (although he probably derived the rule around 1810–1814):

$$P = \frac{Hp}{S.Hp};$$

> ce qui donne les probabilités des diverses causes, lorsqu'elles ne sont
> pas toutes, également possible *á priori*.

On the left-hand side, P denotes the posterior probability of a given cause given an observed event. In the numerator on the right-hand side, H is the probability of an event occurring given that cause, p, is the a priori probability of that cause. In the denominator, $S.$ denotes summation (the modern equivalent of this is Σ) over all possible causes, and H and p now represent the corresponding quantities to those in the numerator, but for each possible cause. Laplace actually presented two versions of the rule – one for discrete random variables (as we show above) and another for continuous variables. The typesetting he used for the continuous case, however, did not allow him to write limits on integrals, meaning that the numerator and denominator look the same.

History has been unfair to Laplace and Price. If they were alive today, the theory would, no doubt, be known as the Bayes–Price–Laplace rule. We hope by including this short biographical section that this will encourage you, in your own work, to give credit to others where it is due. We, in particular, would like to thank Sharon McGrayne for her excellent book, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*, that served as an invaluable reference to this section, and we encourage others to read it to learn of the tempestuous history of Bayesian inference [26].

2.4 THE PURPOSE OF STATISTICAL INFERENCE

How much does a particular drug affect a patient's condition? What can an average student earn after obtaining a college education? Will the Democrats win the next US presidential election? In life, we develop theories and use these to make predictions, but testing those theories is not easy. Life is complicated, and it is often impossible to exactly isolate the parts of a system which we want to examine. The outcome of history is determined by a complex nexus of interacting elements, each of which contributes to the reality that we witness. In the case of a drug trial, we may not be able to control the diets of participants and are certainly unable to control for their idiosyncratic metabolisms, both of which could impact the results we observe. There are a range of factors which affect the wage that an individual ultimately earns, of which education is only one. The outcome of the next US presidential election depends on party politics, the performance of the incumbent government and the media's portrayal of the candidates.

In life, noise obfuscates the signal. What we see often appears as an incoherent mess that lacks any appearance of logic. This is why it is difficult to make predictions and test theories about the world. It is like trying to listen to a classical orchestra which is playing on the side of a busy motorway, while we fly overhead in a plane. Statistical inference allows us to focus on the music by separating the signal from the noise. We will hear 'Nessun Dorma' played!

Statistical inference is the logical framework which we can use to trial our beliefs about the noisy world against *data*. We formalise our beliefs in models of *probability*. The models are probabilistic because we are ignorant of many of the interacting parts of a system, meaning we cannot say with certainty whether something will, or will not, occur. Suppose that we are evaluating the efficacy of a drug in a trial. Before we carry out the trial, we might believe that the drug will cure 10% of people with a particular ailment. We cannot say which 10% of people will be cured because we do not know enough about the disease or individual patient biology to say exactly whom. Statistical inference allows us to test this belief against the data we obtain in a clinical trial.

There are two predominant schools of thought for carrying out this process of inference: Frequentist and Bayesian. Although this book is devoted to the latter, we will now spend some time comparing the two approaches so that the reader is aware of the different paths taken to their shared goal.

2.5 THE WORLD ACCORDING TO FREQUENTISTS

In Frequentist (or Classical) statistics, we suppose that our sample of data is the result of one of an infinite number of exactly repeated experiments. The sample we see in this context is assumed to be the outcome of some probabilistic process. Any conclusions we draw from this approach are based on the supposition that events occur with probabilities, which represent the long-run frequencies with which those events occur in an infinite series of experimental repetitions. For example, if we flip a coin, we take the proportion of heads observed in an infinite number of throws as defining the probability of obtaining heads. Frequentists suppose that this probability actually exists, and is fixed for each set of coin throws that we carry out. The sample of coin flips we obtain for a fixed and finite number of throws is generated as if it were part of a longer (that is, infinite) series of repeated coin flips (see the left-hand panel of Figure 2.1).

In Frequentist statistics the data are assumed to be *random* and results from *sampling* from a fixed and defined *population* distribution. For a Frequentist the noise that obscures the true signal of the real population process is attributable to *sampling variation* – the fact that each sample we pick is slightly different and not exactly representative of the population.

We may flip our coin 10 times, obtaining 7 heads even if the long-run proportion of heads is $\frac{1}{2}$. To a Frequentist, this is because we have picked a slightly odd sample from the population of infinitely many repeated throws. If we flip the coin another 10 times, we will likely get a different result because we then pick a different sample.

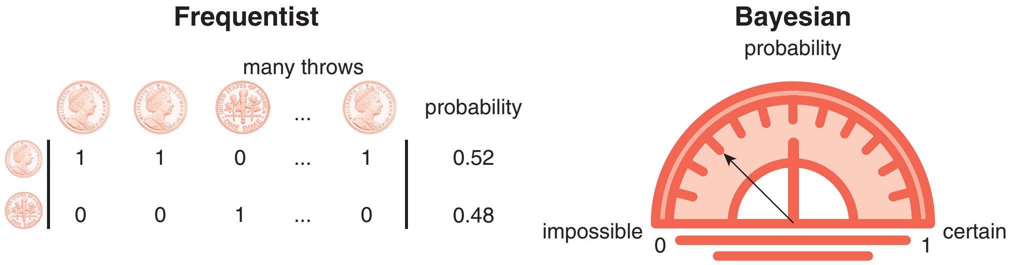


Figure 2.1 The Frequentist (left) and Bayesian (right) approaches to probability.

2.6 THE WORLD ACCORDING TO BAYESIANS

Bayesians do not imagine repetitions of an experiment in order to define and specify a probability. A probability is merely taken as a measure of certainty in a particular belief. For Bayesians the probability of throwing a 'heads' measures and quantifies our underlying belief that before we flip the coin it will land this way.

In this sense, Bayesians do not view probabilities as underlying laws of cause and effect. They are merely abstractions which we use to help express our uncertainty. In this frame of reference, it is unnecessary for events to be repeatable in order to define a probability. We are thus equally able to say, 'The probability of a heads is 0.5' or 'The probability of the Democrats winning the 2020 US presidential election is 0.75'. Probability is merely seen as a scale from 0, where we are certain an event will not happen, to 1, where we are certain it will (see the right-hand panel of Figure 2.1).

A statement such as 'The probability of the Democrats winning the 2020 US presidential election is 0.75' is hard to explain using the Frequentist definition of a probability. There is only ever one possible sample – the history that we witness – and what would we actually mean by the 'population of all possible US elections which happen in the year 2020'?

For Bayesians, probabilities are seen as an expression of subjective beliefs, meaning that they can be updated in light of new data. The formula invented by the Reverend Thomas Bayes provides the only logical manner in which to carry out this updating process. Bayes' rule is central to Bayesian inference whereby we use probabilities to express our uncertainty in parameter values after we observe data.

Bayesians assume that, since we are witness to the data, it is *fixed*, and therefore does not vary. We do not need to imagine that there are an infinite number of possible samples, or that our data are the undetermined outcome of some random process of sampling. We never perfectly know the value of an unknown parameter (for example, the probability that a coin lands heads up). This epistemic uncertainty (namely, that relating to our lack of knowledge) means that in Bayesian inference the parameter is viewed as a quantity that is probabilistic in nature. We can interpret this in one of two ways. On the one hand, we can view the unknown parameter as truly being fixed in some absolute sense, but our beliefs are uncertain, and thus we express this uncertainty using probability. In this perspective, we view the sample as a noisy representation of the signal and hence obtain different results for each set of coin throws. On the other hand, we can suppose that there is not some definitive true, immutable probability of obtaining a heads, and so for each sample we take, we unwittingly get a slightly different parameter. Here we get different results from each round of coin flipping because each time we subject our system to a slightly different probability of its landing heads up. This could be because we altered our throwing technique or started with the coin in a different position. Although these two descriptions are different philosophically, they are not different mathematically, meaning we can apply the same analysis to both.

2.7 DO PARAMETERS ACTUALLY EXIST AND HAVE A POINT VALUE?

For Bayesians, the parameters of the system are taken to vary, whereas the known part of the system – the data – is taken as given. Frequentist statisticians, on the other hand, view the unseen part of the system – the parameters of the probability model – as being fixed and the known parts of the system – the data – as varying. Which of these views you prefer comes down to how you interpret the parameters of a statistical model.

In the Bayesian approach, parameters can be viewed from two perspectives. Either we view the parameters as truly *varying*, or we view our knowledge about the parameters as imperfect. The fact that we obtain different estimates of parameters from different studies can be taken to reflect either of these two views.

In the first case, we understand the parameters of interest as varying – taking on different values in each of the samples we pick (see the top panel of Figure 2.2). For example, suppose that we conduct a blood test on an individual in two consecutive weeks, and represent the correlation between the red and white cell count as a parameter of our statistical model. Due to the many factors that affect the body's metabolism, the count of each cell type will vary somewhat randomly, and hence the parameter value may vary over time. In the second case, we view our uncertainty over a parameter's value as the reason we estimate slightly different values in different samples. This uncertainty should, however, decrease as we collect more data (see the middle panel of Figure 2.2). Bayesians are more at ease in using parameters as a means to an end – taking them not as real immutable constants, but as tools to help make inferences about a given situation.

The Frequentist perspective is less flexible and assumes that these parameters are constant, or represent the average of a long run – typically an infinite number – of identical experiments. There are occasions when we might think that this is a reasonable assumption. For example, if our parameter represented the probability that an individual taken at random from the UK population has dyslexia, it is reasonable to assume that there is a *true*, or fixed, *population* value of the parameter in question. While the Frequentist view may be reasonable here, the Bayesian view can also handle this situation. In Bayesian statistics these parameters can be assumed fixed, but that we are uncertain of their value (here the true prevalence of dyslexia) before we measure them, and use a probability distribution to reflect this uncertainty.

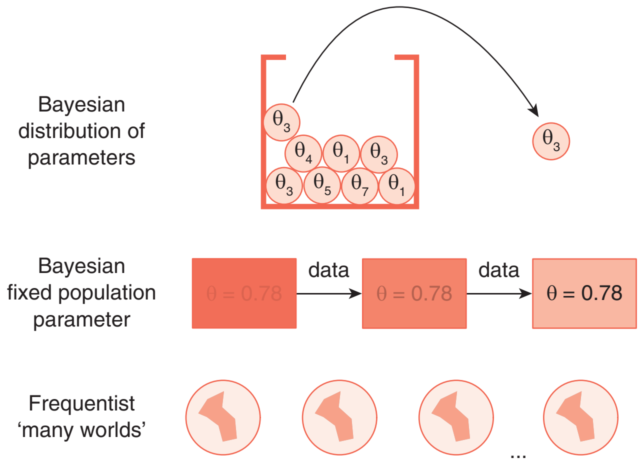

Bayesian distribution of parameters
θ3
θ4 θ1 θ3
θ3 θ5 θ7 θ1
θ3
Bayesian fixed population parameter
θ = 0.78
data
θ = 0.78
data
θ = 0.78
Frequentist 'many worlds'
...

Figure 2.2 The Bayesian (top and middle) and Frequentist perspectives on parameters. In the top panel, the urn holds a large number of parameter values – a population distribution – that we sample from each time we pick a new sample. These parameters, in turn, determine the data that we obtain in our sample. The middle panel shows the Bayesian view where the uncertainty about a parameter's true value (shown in the box) decreases as we collect more data. The bottom panel represents the Frequentist view where parameters represent averages across an infinite number of exactly repeated experiments (represented by the many worlds).

But there are circumstances when the Frequentist view runs into trouble. When we are estimating parameters of a complex distribution, we typically do not view them as actually existing. Unless you view the Universe as being built from mathematical building blocks,[1] then it seems incorrect to assert that a given parameter has any deeper existence than that with which we endow it. The less restrictive Bayesian perspective here seems more reasonable.

The Frequentist view of parameters as a limiting value of an average across an infinity of identically repeated experiments (see the bottom panel of Figure 2.2) also runs into difficulty when we think about one-off events. For example, the probability that the Democrat candidate wins in the 2020 US election cannot be justified in this way, since elections are never rerun under the exact same conditions.

2.8 FREQUENTIST AND BAYESIAN INFERENCE

The Bayesian inference process is the only logical and consistent way to modify our beliefs to account for new data. Before we collect data we have a probabilistic description of our beliefs, which we call a *prior*. We then collect data, and together with a model describing our theory, Bayes' formula allows us to calculate our post-data or *posterior* belief:

$$\text{prior} + \text{data} \xrightarrow{\textit{model}} \text{posterior}. \tag{2.4}$$

For example, suppose that we have a prior belief that a coin is fair, meaning that the probability of it landing heads up is ½. We then throw it 10 times and find that it lands heads up every time; this is our data. Bayes' rule tells us how to combine the prior with the data to result in our updated belief that the coin is fair. Ignore for the moment that we have not explained the meaning of this mysterious prior, as we shall introduce this element properly in Section 2.9.2.

[1] See [37] for an interesting argument for this hypothesis.

In inference, we want to draw conclusions based purely on the rules of probability. If we wish to summarise our evidence for a particular hypothesis, we describe this using the language of probability, as the 'probability of the hypothesis given the data obtained'. The difficulty is that when we choose a probability model to describe a situation, it enables us to calculate the 'probability of obtaining our data given our hypothesis being true' – the opposite of what we want. This probability is calculated by accounting for all the possible samples that could have been obtained from the population, if the hypothesis were true. The issue of statistical inference, common to both Frequentists and Bayesians, is how to invert this probability to get the desired result.

Frequentists stop here, using this inverse probability as evidence for a given hypothesis. They assume a hypothesis is true and on this basis calculate the probability of obtaining the observed data sample. If this probability is small, then it is assumed that it is unlikely that the hypothesis is true, and we reject it. In our coin example, if we throw the coin 10 times and it always lands heads up (our data), the probability of this data occurring given that the coin is fair (our hypothesis) is small. In this case, Frequentists would reject the hypothesis that the coin is fair. Essentially, this amounts to setting $Pr(\text{hypothesis} \mid \text{data}) = 0$. However, if this probability is not below some arbitrary threshold, then we do not reject the hypothesis. But Frequentist inference is then unclear about what probability we should ascribe to the hypothesis. Surely it is non-zero, but exactly how confident are we in it? In Frequentist inference we do not get an accumulation of evidence for a particular hypothesis, unlike in Bayesian statistics.

In reality, Frequentist inference is slightly different to what we described. Since the probability of obtaining any one specific data sample is very small, we calculate the probability of obtaining a range of possible samples to obtain a more usable probability. In particular, Frequentists calculate the probability of obtaining a sample as extreme as, or more extreme than, the one actually obtained, assuming a certain hypothesis to be true. For example, imagine we have a hypothesis that people's heights are normally distributed with a mean of 1.55m and a standard deviation of 0.3m. Then suppose we collect a sample of one individual with a height of 2.5m. To test the validity of the hypothesis, Frequentists calculate the probability of obtaining a height greater than, or equal to, 2.5m, assuming the hypothesis to be true. However, we did not actually witness an individual with a height greater than 2.5m. In Frequentist inference we must invent fictitious samples to test a hypothesis!

Bayes' formula allows us to circumvent these difficulties by inverting the Frequentist probability to get the 'probability of the hypothesis given the *actual* data we obtained'. In our heights example, this would be the probability that the mean population height is 1.55m and has a standard deviation of 0.3m given that our data consists of a single individual of height 2.5m. In Bayesian inference, there is no need for an arbitrary threshold in the probability in order to validate the hypothesis. All information is summarised in this (posterior) probability and there is no need for explicit hypothesis testing. However, to use Bayes' rule for inference, we must supply a prior – an additional element compared to Frequentist statistics. The prior is a probability distribution that describes our beliefs in a hypothesis before we collect and analyse the data. In Bayesian inference, we then update this belief to produce something known as a posterior, which represents our post-analysis belief in the hypothesis.

The next few, albeit silly, examples illustrate a difference in methodology but also, perhaps more significantly, in philosophy between the two different approaches.

2.8.1 The Frequentist and Bayesian murder trials

Assume you find yourself in the unfortunate situation where you are (hopefully falsely) accused of murder, and face a trial by jury. A complication in the tale is that you personally have a choice over the method used by the jury to assign guilt: either Frequentist or Bayesian. Another unfortunate twist is that the legal system of the country starts by presuming guilt rather than innocence.

Let's assume that security camera footage indicates you were in the same house as the victim – Sally – on the night of her demise.

If you choose the Frequentist trial, your jurors start by specifying a model based on previous trials, which assigns a probability of your being seen by the security camera if you were guilty. They use this to make the statement that 'If you did commit the murder, then 30% of the time you would have been seen by the security camera' based on a hypothetical infinity of repetitions of the same conditions. Since $Pr(\text{you were seen by the camera} \mid \text{guilt})$ is not sufficiently unlikely (the p value is not below 5%), the jurors cannot reject the null hypothesis of guilt, and you are sentenced to life in prison.

In a Bayesian trial, the jury is first introduced to an array of evidence, which suggests that you neither knew Sally nor had any previous record of violent conduct, being otherwise a perfectly respectable citizen. Furthermore, Sally's ex-boyfriend is a multiple offending-violent convict on the run from prison after being sentenced by a judge on the basis of Sally's own witness testimony. Using this information, the jury sets a prior probability of the hypothesis that you are guilty equal to $\frac{1}{1000}$ (don't worry about what is meant by a 'prior' as we devote all of Chapter 5 to this purpose). The jury then uses the same model as the Frequentists which indicates that 30% of the time you would have been seen by the camera if you were guilty. However, the jury then coolly uses Bayes' rule and concludes that the probability of your committing the crime is $\frac{1}{1000}$ (see Section 2.13.1 for a full description of this calculation). Based on this evidence, the jury acquits you, and you go home to your family.

2.8.2 Radio control towers

In a hypothetical war, two radio control workers, Mr Pearson (from the county of Frequentland) and Mr Laplace (from the county of Bayesdom), sit side by side and are tasked with finding an enemy plane that has been spotted over the country's borders. They will each feed this information to the nearest air force base(s), which will respond by sending up planes of their own. There are, however, two different air forces – one for each county. Although the air forces of Frequentland and Bayesdom share airbases, they are distinct, and only respond to Mr Pearson's and Mr Laplace's advice, respectively. The ongoing war, though short, has been costly to both allies, and they each want to avoid needless expenditure while still defending their territory.

Mr Pearson starts by inputting the plane's radar information into a computer program that uses a model of a plane's position which has been calibrated against historical enemy plane data. The result comes out instantly:

> The plane is most likely 5 miles North of the town of Tunbridge Wells.

Without another moment's thought, Mr Pearson radios the base of Tunbridge Wells, telling them to scramble all 10 available Frequentist fighter jets immediately. He then gets up and makes himself a well-earned coffee.

Mr Laplace knows from experience that the enemy has used three different flight paths to attack in the past. Accordingly, he gives these regions a high probability density in his prior for the plane's current location and feeds this into the same computer program used by Mr Pearson. The output this time is different. By using the optional input, the program now outputs a map with the most likely regions indicated, rather than a single location. The highest posterior density is over the region near Tunbridge Wells, where Mr Pearson radioed, although the map suggests there are two other towns which might also be victims of the plane's bombing. Accordingly, Mr Laplace radios to Tunbridge Wells, asking them to send up four jets, and to the other two towns, asking them to send up two jets each. At the end of all this, Mr Laplace remains seated, tired but contented that he has done his best for his own.

The enemy bomber turned out to be approaching Berkstad, one of the towns which Mr Laplace radioed. The Bayesdom jets intercept the encroaching plane and escort it out of allied airspace. Mr Laplace is awarded a medal in honour of his efforts. Pearson looks on jealously.

2.9 BAYESIAN INFERENCE VIA BAYES' RULE

Bayes' rule tells us how to update our prior beliefs in order to derive better, more informed, beliefs about a situation in light of new data. In Bayesian inference, we test hypotheses about the real world using these posterior beliefs. As part of this process, we estimate characteristics that interest us, which we call *parameters*, that are then used to test such hypotheses. From this point onwards we will use θ to represent the unknown parameter(s) which we want to estimate.

The Bayesian inference process uses Bayes' rule to estimate a probability distribution for those unknown parameters after we observe the data. (Don't worry if you don't know what is meant by a *probability distribution* since we shall devote the entirety of Chapter 3 to this purpose.) However, it is sufficient for now to think of probability distributions as a way to represent uncertainty for unknown quantities.

Bayes' rule as used in statistical inference is of the form:

$$p(\theta \mid data) = \frac{p(data \mid \theta) \times p(\theta)}{p(data)}, \tag{2.5}$$

where we use p to indicate a probability distribution which may represent either probabilities or, more usually, probability densities (see Section 3.3.2 for a description of their distinction). We shall now spend the next few sections describing, in short, the various elements of expression (2.5). This will only be a partial introduction since we spend the entirety of Part II on an extensive discussion of each of the constituent components.

2.9.1 Likelihoods

Starting with the numerator on the right-hand side of expression (2.5), we come across the term $p(data \mid \theta)$, which we call the *likelihood*, which is common to both Frequentist and Bayesian analyses. This tells us the probability of generating the particular sample of data if the parameters in our statistical model were equal to θ. When we choose a statistical model, we can usually calculate the probability of particular outcomes, so this is easily obtained. Imagine that we have a coin that we believe is fair. By *fair*, we mean that the probability of the coin landing heads up is $\theta = \frac{1}{2}$. If we flip the coin twice, we might suppose that the outcomes are independent events (see Section 3.4), and hence can calculate the probabilities of the four possible outcomes by multiplying the probabilities of the individual outcomes:

$$\begin{aligned}
Pr(H,H \mid \theta = \tfrac{1}{2}) &= Pr(H \mid \theta = \tfrac{1}{2}) \times Pr(H \mid \theta = \tfrac{1}{2}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
Pr(H,T \mid \theta = \tfrac{1}{2}) &= Pr(H \mid \theta = \tfrac{1}{2}) \times Pr(T \mid \theta = \tfrac{1}{2}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
Pr(T,H \mid \theta = \tfrac{1}{2}) &= Pr(T \mid \theta = \tfrac{1}{2}) \times Pr(H \mid \theta = \tfrac{1}{2}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
Pr(T,T \mid \theta = \tfrac{1}{2}) &= Pr(T \mid \theta = \tfrac{1}{2}) \times Pr(T \mid \theta = \tfrac{1}{2}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.
\end{aligned} \tag{2.6}$$

(Don't worry if you don't understand the logic in the above, as we devote the whole of Chapter 4 to understanding likelihoods.)

2.9.2 Priors

The next term in the numerator of expression (2.5) $p(\theta)$, is the most controversial part of the Bayesian formula, which we call the prior distribution of θ. It is a probability distribution which represents our pre-data beliefs across different values of the parameters in our model, θ. This appears, at first, to be counterintuitive, particularly if you are familiar with the world of Frequentist statistics, which does not require us to state our beliefs explicitly (although we always do implicitly, as we explain in Section 2.10). Continuing the coin example, we might assume that we do not know whether the coin is fair or biased beforehand, so suppose all possible values of $\theta \in [0, 1]$ – which represents the probability of the coin falling heads up – are equally likely. We can represent these beliefs by a continuous uniform probability density on this interval (see the black line in Figure 2.3). More sensibly, however, we might believe that coins are manufactured in a way such that their weight distribution is fairly evenly distributed, meaning that we expect that the majority of coins are reasonably fair. These beliefs would be more adequately represented by a prior similar to the one shown by the red line in Figure 2.3.

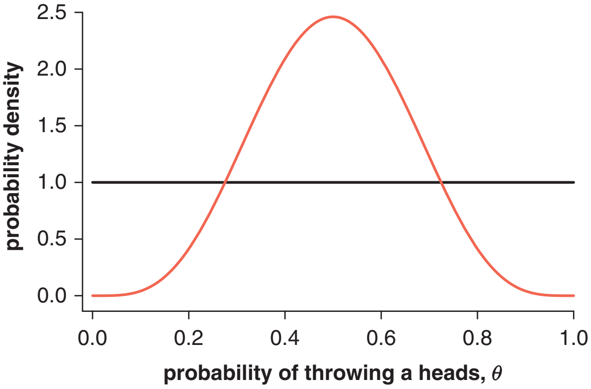


Figure 2.3 Two different prior distributions: a uniform prior, where we believe all values of θ (corresponding to the probability of throwing a heads) are equally likely (black line), and another where we believe that the coin is most likely fair before we throw it (red line).

The concept of priors will be covered in detail in Chapter 5.

2.9.3 The denominator

The final term on the right-hand side of expression (2.5) in the denominator is $p(data)$. This represents the probability of obtaining our particular sample of data if we assume a particular model and prior. We will mostly postpone discussion of this term until Chapter 6 when we understand better the significance of likelihoods and priors. However, for our purposes here it suffices to say that the denominator is fully determined by our choice of prior and likelihood function. While it appears simple, this is deceptive, and it is partly the difficulty with calculating this term that leads to the introduction of computational methods that we discuss in Part IV.

The concept of the denominator will be covered in detail in Chapter 6.

2.9.4 Posteriors: the goal of Bayesian inference

The posterior probability distribution $p(\theta \mid data)$ is the main goal of Bayesian inference. For example, we might want to compute the probability distribution representing our post-experimental beliefs of the inherent bias, θ, of a coin, given that it was flipped 10 times and it landed heads up 7 times. If we use Bayes' rule, assuming the likelihood model specified in Section 2.9.1, and the uniform prior shown in Figure 2.3 (black line), then the result is the posterior distribution shown as the grey line in Figure 2.4. Here, the peak of the distribution occurs at $\theta = 0.7$, which corresponds exactly with the percentage of 'heads' obtained in the experiment.

The posterior distribution summarises our uncertainty over the value of a parameter. If the distribution is narrower, then this indicates that we have greater confidence in our estimates of the parameter's value. More narrow posterior distributions can be obtained by collecting more data. In Figure 2.4, we compare the posterior distribution for the previous case where 7 out of 10 times the coin landed heads up with a new, larger, sample where 70 out of 100 times the same coin comes up heads. In both cases, we obtained the same ratio of heads to tails, resulting in the same peak value at $\theta = 0.7$. However, in the latter case, since we have more evidence to support our claim, we end up with greater certainty about the parameter value after the experiment.

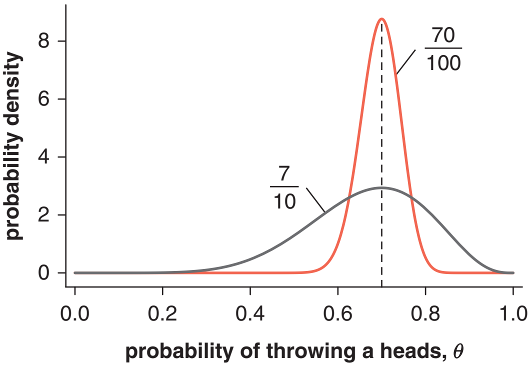


Figure 2.4 Posterior distributions for θ – the probability that a coin landing heads up when flipped. The grey line represents the posterior probability distribution function (PDF) resulting from a data sample where 7 out of 10 times the coin came up heads. The red line is the posterior probability distribution function for the case where 70 out of 100 times the coin came up heads. Both of the posteriors assume a binomial likelihood and uniform prior (don't worry if these mean nothing to you as we will introduce these concepts in Chapters 4 and 5).

The posterior distribution is also used to predict future outcomes of an experiment and for model testing. However, we leave discussion of these until Chapter 7.

2.10 IMPLICIT VERSUS EXPLICIT SUBJECTIVITY

One of the major arguments levied against Bayesian statistics is that it is *subjective* due to its dependence on the analyst specifying their pre-experimental beliefs through priors. This experimenter prejudice towards certain outcomes is said to bias the results away from the types of fair, objective outcomes resultant from a Frequentist analysis.

We argue that *all* analyses involve a degree of subjectivity, which is either explicitly stated or, more often, implicitly assumed. In a Frequentist analysis, the statistician typically selects a model of probability which depends on a range of assumptions. These assumptions are often justified explicitly, revealing their suggestive nature. For example, the simple *linear regression model* is often used, without justification, in applied Frequentist analyses. This model makes assumptions about the relationships between the dependent and independent variables that may, or may not, be true. In a Bayesian approach, we more typically build our models from the ground up, meaning that we are more aware of the assumptions inherent in the approach.

In applied research, there is a tendency among scientists to choose data to include in an analysis to suit one's needs, although this practice should really be discouraged (see [20]). The choice of which data points to include is subjective, and the underlying logic behind this choice is more often than not kept opaque from the reader.

A further source of subjectivity is the way in which models are checked and tested. In analyses, both Frequentist and Bayesian, there is a need to exercise (subjective) judgement in suggesting a methodology which will be used in this process. We would argue that Bayesian analysis allows greater flexibility and a more suitable methodology for this process because it accounts for the inherent uncertainty in our estimates.

In contrast to the examples of subjectivity mentioned above, Bayesian priors are explicitly stated. This makes this part of the analysis openly available to the reader, meaning it can be interrogated and debated. This transparent nature of Bayesian statistics has led some to suggest that it is honest. While Frequentist analyses hide behind a fake veil of objectivity, Bayesian equivalents explicitly acknowledge the subjective nature of knowledge.

Furthermore, the more data that is collected, (in general) the less impact the prior exerts on posterior distributions. In any case, if a slight modification of priors results in a different conclusion being reached, it must be reported by the researcher.

Finally, comparing the Frequentist and Bayesian approaches to the pursuit of knowledge, we find that both approaches require a subjective judgement to be made. In each case, we want to obtain $p(\theta \mid data)$ – the probability of the parameter or hypothesis under investigation, given the data set which has been observed. In Frequentist hypothesis testing we do not calculate this quantity directly, but use a rule of thumb. We calculate the probability that the data set would, in fact, have been more extreme than those we actually obtained assuming a null (the given, default) hypothesis is true. If the probability is sufficiently small, typically less than a cut-off of 5% or 1%, then we reject the null. This choice of threshold probability – known as a statistical test's *size* – is completely arbitrary, and subjective. In Bayesian statistics, we instead use a subjective prior to invert the likelihood from $p(data \mid \theta) \rightarrow p(\theta \mid data)$. There is no need to accept or reject a null hypothesis and consider an alternative since all the information is neatly summarised in the posterior. In this way we see a symmetry in the choice of Frequentist test size and Bayesian priors; they are both required to invert the likelihood to obtain a posterior.

2.11 CHAPTER SUMMARY

This chapter has focused on the philosophy of statistical inference. Statistical inference is the process of inversion required to go from an effect (the data) back to a cause (the process or parameters). The trouble with this inversion is that it is generally much easier to do things the other way round: to go from a cause to an effect. Frequentists and Bayesians start by defining a forward probability model that can generate data (the effect) from a given set of parameters (the cause). The method that they each use to run this model in reverse and determine the probability for a cause is different. Frequentists assume that if the probability of generating the data (actually data as extreme as or more extreme than that obtained) from a particular cause is small, then the cause is rejected; the probability of that cause is concluded to be zero. The set of all non-rejected causes then forms a confidence interval that contains the actual cause with some measure of certainty. Bayesians instead carry out the inversion formally using Bayes' rule. This results in an accumulation of evidence for each cause, rather than a binary 'yes' or 'no' as for the Frequentist case.

Frequentists and Bayesians also differ in their view on probabilities. Frequentists view probabilities as the frequency at which an event occurs in an infinite series of experimental repetitions. In this sense Frequentists view probabilities as fixed laws that actually exist independent of the individual analyst. Because they are fixed, it does not make sense to update them. Similarly, in the Frequentist viewpoint, it does not make sense to define probabilities for one-off events, where an infinite series of experimental reproductions is not possible. Bayesians take a more general view on probabilities. They see probabilities as measuring the strength of an individual's underlying belief in the likelihood of some outcome. For Bayesians probabilities are only defined in relation to a particular analyst and are hence, by their very nature, subjective. Since probabilities measure beliefs, they can be updated in light of new data. The only correct way to update probabilities is through Bayes' rule, which Bayesians use to do statistical inference. Because Bayesian probabilities measure a subjective belief in an outcome, they can be used for all categories of events, from those that could in some way be infinitely repeated (for example, coin flips) or one-off events (for example, the outcome of the 2020 US presidential election).

One argument that is often levied against Bayesian approaches to inference is that they are subjective, in contrast to the objectivity of Frequentism. We argued that all analytical approaches to inference are inherently subjective at some level. Beginning with the data selection process, the analyst often makes a subjective judgement of which data to include. The choice of a specific probability model is also inherently subjective and is typically justified by making assumptions about the data-generating process. In Frequentist inference the choice of the threshold probability for null hypothesis testing is also arbitrary and inherently depends on the analyst. Bayesian inference has priors, which should always be explicitly stated in an analysis. That priors are explicitly stated means that they can be debated and interrogated in a transparent fashion. While priors are inherently subjective, this does not mean they cannot be informed by data. In fact, in analyses that are repeated at different points in time, it often makes sense to use the posterior of a previous analysis as a prior for a new one (see Chapter 7).

In this chapter, we also introduced Bayes' rule for inference and discussed briefly its constituent parts. The Bayesian formula is the central dogma of Bayesian inference. However, in order to use this rule for statistical analyses, it is necessary to understand and, more importantly, be able to manipulate probability distributions. The next chapter is devoted to this cause.

2.12 CHAPTER OUTCOMES

The reader should now be familiar with the following concepts:

- the goals of statistical inference
- the difference in interpretation of probabilities for Frequentists versus Bayesians
- the differences in the Frequentist and Bayesian approaches to inference

2.13 APPENDIX

2.13.1 The Frequentist and Bayesian murder trials

In the Bayesian trial the probability of guilt if you are seen by the security camera on the night of the murder is:

$$\begin{aligned} p(\textit{guilt} \mid \textit{security camera footage}) &= \frac{p(\textit{security camera footage} \mid \textit{guilt}) \times p(\textit{guilt})}{p(\textit{security camera footage})} \\ &= \frac{\frac{30}{100} \times \frac{1}{1000}}{\frac{30}{100} \times \frac{999}{1000} + \frac{30}{100} \times \frac{1}{1000}} \\ &= \frac{1}{1000}. \end{aligned} \tag{2.7}$$

In the above equation we assume that the security camera is hidden, and hence a murderer does not change their behaviour to avoid being seen, meaning that the probability of being seen by the security camera in each case is 30%. We have also assumed that the footage is itself uninformative about the motivations of an individual; it is merely indicative of a person's location at a given time. In other words, we are supposing that criminals and innocents cannot be differentiated by their actions on the video.

2.14 PROBLEM SETS

Problem 2.1 The deterministic nature of random coin throwing

Suppose that, in an idealised world, the ultimate fate of a thrown coin – heads or tails – is deterministically given by the angle at which you throw the coin and its height above a table. Also in this ideal world, the heights and angles are discrete. However, the system is chaotic[2] (highly sensitive to initial conditions), and the results of throwing a coin at a given angle and height are shown in Table P2.1.

[2] The authors of the following paper actually experimentally tested this and found it to be the case, "The three-dimensional dynamics of the die throw", Chaos, Kapitaniak et al. (2012).

Problem 2.1.1 Suppose that all combinations of angles and heights are equally likely to be chosen. What is the probability that the coin lands heads up?

Problem 2.1.2 Now suppose that some combinations of angles and heights are more likely to be chosen than others, with the probabilities shown in Table P2.2. What are the new probabilities that the coin lands heads up?

Problem 2.1.3 We force the coin-thrower to throw the coin at an angle of 45 degrees. What is the probability that the coin lands heads up?

Problem 2.1.4 We force the coin-thrower to throw the coin at a height of 0.2m. What is the probability that the coin lands heads up?

Problem 2.1.5 If we constrained the angle and height to be fixed, what would happen in repetitions of the same experiment?

Problem 2.1.6 In light of the previous question, comment on the Frequentist assumption of exact repetitions of a given experiment.

Table P2.1 The results of a coin throw from a given angle and height above a table.

	Height above table (m)				
Angle (degrees)	**0.2**	**0.4**	**0.6**	**0.8**	**1**
0	T	H	T	T	H
45	H	T	T	T	T
90	H	H	T	T	H
135	H	H	T	H	T
180	H	H	T	H	H
225	H	T	H	T	T
270	H	T	T	T	H
315	T	H	H	T	T

Table P2.2 The probability that a given person throws a coin at a particular angle and at a certain height above a table.

	Height above table (m)				
Angle (degrees)	**0.2**	**0.4**	**0.6**	**0.8**	**1**
0	0.05	0.03	0.02	0.04	0.04
45	0.03	0.02	0.01	0.05	0.02
90	0.05	0.03	0.01	0.03	0.02
135	0.02	0.03	0.04	0.00	0.04
180	0.03	0.02	0.02	0.00	0.03
225	0.00	0.01	0.04	0.03	0.02
270	0.03	0.00	0.03	0.01	0.04
315	0.02	0.03	0.03	0.02	0.01

Problem 2.2 Objections to Bayesianism

The following criticisms of Bayesian statistics are raised in an article by Gelman [4]. Provide a response to each of these.

Problem 2.2.1 'As scientists we should be concerned with objective knowledge rather than subjective belief.'

Problem 2.2.2 'Subjective prior distributions don't transfer well from person to person.'

Problem 2.2.3 'There's no good objective principle for choosing a noninformative prior … Where do prior distributions come from, anyway?'

Problem 2.2.4 A student in a class of mine: 'If we have prior expectations of a donkey and our dataset is a horse then Bayesians estimate a mule.'

Problem 2.2.5 'Bayesian methods seem to quickly move to elaborate computation.'

Problem 2.3 Model choice

Suppose that you have been given the data contained in `subjective_overfitShort.csv` and are asked to find a 'good' statistical model to fit the (x, y) data.

Problem 2.3.1 Fit a linear regression model using least squares. How reasonable is the fit?

Problem 2.3.2 Fit a quintic (powers up to the fifth) model to the data. How does its fit compare to that of the linear model?

Problem 2.3.3 You are now given new data contained within `subjective_overfitLong.csv`. This contains data on 1000 replications of the same experiment, where the x values are held fixed. Using the least squares fits from the first part of this question, compare the performance of the linear regression model with that of the quintic model.

Problem 2.3.4 Which of the two models do you prefer, and why?