Bayesian book

Ben Lambert

September 30, 2015

Contents

Chapter 1

How to best use this book

1.1 The purpose of this book

This book aims to be friendlier introduction to Bayesian analysis than other books out there. Whenever we introduce new concepts, we hope to keep the mathematics to a minimum and instead focus on the intuition behind the theory. However, we hope not to sacrifice on content for the sake of simplicity; aiming to cover everything from the basics up to the bleeding edge of the field. Overall, this book seeks to plug a gap in the existent literature (see figure 1.1).

To help the reader on the way, we have developed a number of interactive elements which are accessible through the book's website, as well as example code, for the reader to peruse, and if willing, *rerun*. We also supplement key ideas by the presence of videos, which approach topics through different angles, and examples, to those present in the text.

At the end of each chapter there are problem sets, which allow the student to build up practical experience of Bayesian analysis. Whenever appropriate these problem sets will also be supplemented with video material.

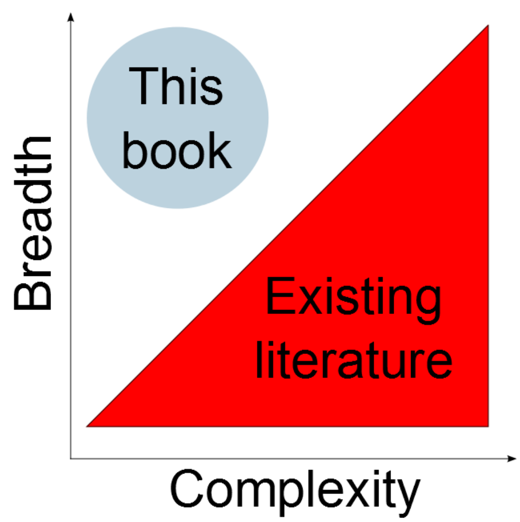


Figure 1.1: The aim of this book.

1.2 Who is this book for?

This book is for anyone who has ever tried, and **failed** at statistics, and particularly, Bayesian statistics.

The text is aimed at anyone who has completed high school mathematics, and wants to conduct Bayesian inference on real world data. We assume no previous knowledge of probability[1], and devote the entirety of chapter 3 to this purpose. We do not require that the student is versed in classical statistics, as we aim to build an alternative, and complementary path to a shared goal. As such, after chapter 2, we will not give too much attention to comparisons between these two approaches.

Whilst we start at the very beginning of statistical inference, we aim to provide a guide which will be of practical-use for a large proportion of analyses that are likely to be encountered in real life.

1.3 Pre-requisites

Knowledge of the following is strongly recommended, in order to allow the reader to grasp all that is contained within this text:

- **Algebra:** manipulation of symbolic expressions is widespread throughout the text.
- **Summations:** mainly used for writing down likelihood functions.
- **Calculus:** mainly integration, although there are sprinklings of differentiation in places.

The only other pre-requisite is more to do with the practical application of Bayesian analysis: a knowledge of the statistical software *R*[14] would be *useful*. We do not classify this item with those above, partly because we aim to only use very basic functioning in this language, and partly because we will explain those examples thoroughly. This language is widely-used for statistical analysis, and because of its popularity there are many excellent online introductions to this language, which are freely available:

[1] Which is central to Bayesian analysis.

- `www.coursera.org` - there are a number of great lecture courses with associated problem sets available for learning R here. In particular, the courses by Roger Peng at John Hopkins are very worthwhile.
- `http://tryr.codeschool.com` - a short interactive introductory lesson on the basics of R.
- `www.datacamp.com/courses/free-introduction-to-r` - 4 hours of interactive lectures on the basics of R.
- `http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf` - a nice written guide to R.

Whilst none of these are essential, if you have difficulty in following the examples in this text, then we recommend that you have a look at the above resources.

1.4 Book outline

We have written this text to make each chapter as self-contained as is possible. Whilst, at times, the reader may feel that this focus on modularity means that some elements reoccur, this is intended not to make reading repetitious, or worse *boring*! It is for two purposes: to aid the aforementioned compartmentalisation; and because we believe that some ideas are worth encountering, and re-encountering, at different points in your pursuit of knowledge.

The book is divided into five *Parts*:

- **Part** I**: An introduction to Bayesian inference.**
- **Part** II**: Understanding the Bayesian formula.**
- **Part** III**: Analytic Bayesian methods.**
- **Part** IV **: A practical guide to doing real life Bayesian analysis: Computational Bayes.**
- **Part** V **: Regression analysis and hierarchical models.**

Part I firstly provides an introduction to the purpose of statistical inference, then goes on to compare and contrast the Bayesian and classical approaches to this problem. Bayesian inference is based on probability. Hence it is imperative to understand how to manipulate these types of mathematical object. The latter half of this part is thus devoted to exactly this purpose. Part II introduces the reader to the constituent elements of the Bayesian inference formula, and in doing so provides an all-round introduction to the practicalities of doing Bayesian inference. Part III aims to equip the user with the knowledge of the most practically-relevant probability distributions for Bayesian inference. These objects come under two categories in general[2]: prior distributions, and likelihood distributions. Knowledge of these distributions is essential for understanding existent research papers, and books which use Bayesian statistics, as well as to conduct Bayesian inference in the first place. The rest of the part is concerned with introducing the reader to 'nice' combinations of distributions which allow for a pen-and-paper deduction of quantities of interest. This is important not only as a stepping stone to computational methods, but also because these places are often a good place to start before implementing more nuanced models. Part IV introduces the reader to the modern methods of undertaking Bayesian analysis; through computational Markov Chain Monte Carlo. This part aims to provide an intuitive introduction to some of the most important algorithmic tools used in computational methods. It also importantly introduces the reader to the statistical languages covered in this text: *Stan* and *JAGS*. This part is essential reading for anyone wanting to conduct serious real world Bayesian analysis of data. Part V builds on the previous part, and introduces the reader to the most important aspect of Bayesian modelling; hierarchical models. It also provides an in-depth introduction to doing Bayesian regression modelling.

Each chapter has two introductory summaries: the *chapter mission statement*, and *chapter goals*. The former is usually a one or two sentence gist of the material to be covered in the following chapter, as well as the learning outcomes of the section. The *chapter goals* section is more detailed, and links together material encountered in previous chapters, and provides more of a narrative on the position of the particular chapter in question in a journey to understand Bayesian statistics.

The *chapter summary* sections at the end of each chapter provide the reader with a description of the skills acquired within, as well as some perspective

[2] although some distributions fall into both

on the material's position within the book's overall goals. The *chapter outcomes* section provides a check-list of the skills acquired within the chapter.

1.5 Route planner - suggested journeys through Bayesland

Ben note to self: **Change this to make of the same tick and cross table format as that of the Bayesian machine learning book!**

In the style of most good guide books, we provide itineraries through the area in question. These journeys are meant to be shortish paths towards gaining a better understanding of particular elements of Bayesian statistics, although as most trips are, they are not as all-encompassing, as a more prolonged stay. We provide the following trips through Bayesland, based on time, goals, and pre-requisite knowledge:

- **The long-weekender (introductory):** chapter 2 introduces you to the theory behind statistical inference, as well as providing a gentle introductory comparison between Bayesian and classical approaches. If you have extra time, and some knowledge of handling probability distributions, then skip ahead to chapter 4.

- **The two week basic trip (introductory):** Part I and Part II provide a full introduction to Bayesian statistics, from the ground up.

- **The two week refresher (medium):** Read chapter 2 to get your bearings. Dependent on your knowledge of the Bayesian formula Part II can either be read, or left behind. Part III should be read almost in full, as this will get you up to speed with many of the tools necessary to understand many research papers. However, you may be able to 'get away' without reading chapter 11 on Objective Bayes.

- **The two week full practical swing (medium-master):** If you are happy with your knowledge of the Bayesian inference formula, as well as the distributions used in Bayesian analysis, then you may want to skip ahead to Part IV, which introduces computational methods. This introduces the reader to both *Stan* and *JAGS* which are the statistical languages used in this text to carry out Bayesian analyses. If you have time, and are willing, then you may want to progress to Part V.

- **The 'I need to do Bayesian analysis' now three day leg (medium-master):** This is a tailor-made trip suited to those practitioners who need to carry out Bayesian data analysis *fast*. The most likely people here are those in research: either academic or corporate; who have an existing knowledge of Bayesian statistics. Skip ahead to Part IV, and read the chapter on *Stan* (if your analysis only contains a few discrete variables), or *JAGS* (if your analysis contains a number of discrete parameters). You may then want to use the index to look for the most relevant sections, although Part V will be your main point of reference.

- **A theoretic and practical journey through modern Bayesian analysis (medium-master):** You want to learn as much about applied Bayesian methods as time will allow, but you also want to gain some practice in practically doing Bayesian statistics. Read all of Part IV.

- **A two week modelling masterclass (master):** you know all the basics, and are acquainted with the use of either *Stan* or *JAGS*. You want to see these applied to carrying out real life data analysis. Read all of Part V.

1.6 Video

Whenever the reader sees the following signpost, there is a video available to supplement the main text:

$\boxed{\textit{Video}:}$

By following the web address indicated, the user can watch the video on YouTube.

The videos are not meant to replace the reading of the text. They are supplementary, and aim to address topics through alternative approaches, and with different examples.

Ben note to self:**It might be good to provide a link to a real video here.**

1.7 Interactive elements

Whenever the reader sees the following signpost, there is an interactive element available to supplement the main text:

Interactive :

By following the web address indicated, the user can dynamically interact with these examples, by clicking, entering numbers, or moving slider bars. It is hoped that these elements in particular, will allow the user to build up an intuition behind Bayesian theory, which is essential for doing real world analysis.

These elements have been created using Mathematica, although there is no need to own the software in order to play.

1.8 Interactive problem sets

The reader can practice their knowledge using the problem sets at the end of each chapter. The problem sets are **interactive**, and are of the form used in many online education systems. The student can submit their answers to a selection of the questions, and will receive a mark out of 100 for each problem set.

The problems available aim to cover mainly practical application of Bayesian data analysis, although there are also sections which are more theoretic in nature.

For teachers of the subject, there are parts of each problem set which are not interactive, and can be used as a part of a taught course.

1.9 Code

Whenever appropriate, particularly in Part IV onwards, we will include snippets of code in Stan, JAGS and R. These are commented thoroughly, which should be self-explanatory. They aim to be as self-contained as space allows.

1.10 R, Stan and JAGS

Modern Bayesian data analysis uses computers. Luckily, for the student of Bayesian statistics, the most up-to-date and useful software packages are open source; meaning they are freely available to use. In this book we solely use this type of software, so prevent the student needing to spend more on acquiring them.

The most recent, and powerful software to emerge is that developed by Andrew Gelman et al.[17] and is called *Stan*. The language of this software is not difficult to understand, and easier to write and debug than its competition. Stan allows a user to fit complex models to data sets, without having to wait an age for the results. With updates planned to this software, which will make it even more powerful, and user-friendly, this is without question the most appropriate software on which to conduct Bayesian analysis.

Ben note to self: Dependent on when the book is released, we may remove the following section, since Stan may allow for discrete sampling directly by then.

The only snag with Stan is that because of the particular type of engine under the hood, it is not possible to directly use it to conduct analysis for discrete parameters. As we shall see in Part IV, this issue can be sidestepped in Stan, but for some circumstances it is easier to write the model using another piece of software called *JAGS*[13]. This language is slightly different to write, and in nature to that of Stan. We will explain its functioning in detail in Part IV, and walk though a number of introductory examples, however we advocate using this software only in circumstances where it is simpler to implement than a corresponding model in Stan.

The two aforementioned statistical languages are usually run through another piece of 'helper' software. Whilst there are a number of alternatives here, particularly for Stan, we choose to use *R* here. This is because this software is both open source, and widely used. The former means that anyone with a modern computer should be able to get their hands dirty in Bayesian analysis; the latter is important since the code base is well-maintained and tested.

1.11 Why don't more people use Bayesian statistics?

Many are discouraged from using Bayesian approaches to analysis due to its supposed difficulty, and dependence on mathematics. However, we would argue that this is, in part, a weakness of the existent literature on the subject, which this book looks to address. It also highlights how many books on classical statistics sweep their inherent complexity and assumptions under the carpet, resulting in texts which are easy to digest; meaning that for many the path of least resistance is to forge ahead with classical tools.

By its dependence on the logic of probability, this means on first glances, Bayesian statistics appears more mathematically-complex. However, what is often lost in introductory texts on Bayesian theory, is the intuitive explanations behind the mathematical formulae. In this text instead, we shift the emphasis towards the latter; choosing to focus on graphical and illustrative explanations rather than getting lost in the details of the mathematics, which to be honest, is not necessary for much of modern Bayesian analysis. We hope that by doing so, we shall lose fewer casualties to mathematical complexity, and redress the imbalance between classical and Bayesian analysis applications.

Again, on first appearances, the concept of the *prior* no doubt leads many to 'abandon ship' early on the path to understanding better Bayesian methodologies. However, we will cover this concept in detail in chapter 6 which is fully-devoted to this subject, we hope to banish this particular thorn in the side of would-be Bayesian statisticians.

The reliance on computing, in particular simulation, is also seen to add to the complexity of Bayesian approaches. Whilst, this is true, we argue that the modern algorithms used for simulation are straightforward to understand, and with modern software, easy to implement. Furthermore, the added complexity of simulation methods is more than compensated by the straightforward extension of Bayesian models to handle arbitrarily complex situations. Like most things worth learning, there is a slight learning curve to become acquainted with the languages used to write modern Bayesian simulations. However, we hope to make this curve sufficiently shallow by incremental introduction of elements used in these computational applications.

1.12 What are the tangible (non-academic) benefits of Bayesian statistics?

In Bayesian textbooks much discourse is devoted to advocating the academic reasons for choosing to use a Bayesian analysis over classical approaches. However, often authors neglect to promote the more tangible, everyday benefits of the former. Here, we list the following *real* benefits of a Bayesian approach:

- **Simple and intuitive model testing and comparison**. The prior- and posterior-predictive distributions allow for in-depth testing of any particular aspect of a model, by comparing it with the same aspects from the data collected. The Bayesian approach also provides a logical framework in which to compare different models.

- **Straightforward interpretation of results.** In classical analyses, the *confidence interval* is often taken to be a measure of uncertainty for a particular parameter. As we shall see in section 4.3.4, this is not the case, and interpretation of this concept is not straightforward. By contrast Bayesian *credible intervals*, can be taken to be a measure of uncertainty in a parameter, as they are obtained directly from probability distributions.

- **Full model flexibility**. Modern Bayesian analyses use computational simulation in order to carry out analyses. Whilst this might appear excessive when compared to classical approaches, an additional benefit is the straightforward extension to almost arbitrarily complex models when using Bayesian approaches. This means that Bayesian models can be extended to encompass any complexity of data process. This is in contrast to classical approaches, where the intrinsic difficulty of analysis scales with the complexity of the model chosen.

- **The best predictions.** Leading figures both inside and outside of academia use Bayesian approaches for prediction. An example being Nate Silver's correct prediction of the 2008 US Presidential election results [16].

1.13 Suggested further reading

A good book should leave the reader wanting more. Due to the finiteness of this text, we recommend the following books, articles and websites. These aren't necessarily all on Bayesian statistics, but fall under the wider categories of statistical inference and learning. We also provide a score of the complexity of reading these texts, which may guide your choice:

- **Bayesian Data Analysis (medium-master):** A masterpiece produced by the master statisticians Andrew Gelman, and Donald Rubin amongst others. This is the most all-encompassing, and up-to-date text available on applied Bayesian data analysis. There are plenty of examples of Bayesian analysis applied to real world data, and are well-explained[5].
- **Data Analysis Using Regression and Multilevel/Hierarchical Models (medium):** Another belter from Andrew Gelman, along with his co-author Jennifer Hill. This takes the reader through numerous examples of regression modelling, and hierarchical analysis. The text is not solely constrained to Bayesian analysis, and covers classical methods as well.
- **Mastering metrics (introductory):** A great, back-to-basics, book on causal inference by the masters of econometrics Josh Angrist and Jrn-Steffen Pischke. This is an exhibition of the five main methods used to conduct causal inference in the social sciences: regression, matching, instrumental variables, differences-in-differences, and regression discontinuity design. This is a very readable text, and suitable for anyone wanting to learn about policy evaluation.
- **Mostly Harmless Econometrics (master-of-metrics):** Again by Josh Angrist and Jrn-Steffen Pischke. A thorough, and mathematically detailed text which takes the reader through most of those methods used in causal inference today. Its small size is deceptive, and is not one to read over a single weekend. However, it is well worth persisting with this book, as the nuggets that await the determined reader are worth their weight in gold. Also see Gelman's review of this book, which provides an interesting critique of the text.

Part I

An introduction to Bayesian inference

1.14 Part mission statement

The purpose of this Part is twofold: firstly to introduce the reader to the principles of inference; secondly to familiarise them with knowledge of how to manipulate probability distributions, which is essential to Bayesian inference.

1.15 Part goals

Both Frequentist and Bayesian inferential processes aim to obtain a probability of a hypothesis given the data that was obtained. However, it is much easier to obtain the inverse - the probability of the data given a hypothesis. Thus this process of *inversion* is central to all statistical inference. However, there are two distinct ways to go about inversion: Frequentists use a rule of thumb, which has historically been agreed upon as a useful way of carrying out this process; by contrast Bayesians use Bayes' rule - the only method consistent with the logical of probability. Chapter 2 fully introduces the reader to the aims of statistical inference, along with the differences in philosophy and approach taken by Frequentists and Bayesians to this shared goal.

One of the main differences in approach is the Bayesian insistence on describing everything of interest using probability distributions. The resultant theory is more elegant, as well as being more practically useful, than that proposed by Frequentists. However, in order to appreciate this elegance, it is necessary to have a good working knowledge of probability distributions, and their manipulations. Hence chapter 3 provides a fully introductory course on probability distributions.

Chapter 2

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; importantly recognising the similarities and differences between Frequentist and Bayesian inference. We then go on to introduce the most important theorem in modern statistics: *Bayes' rule*.

2.2 Chapter goals

In life, we are often tasked with building predictive models to understand complex phenomena. As a first approximation, we often disregard parts of the system, which are not directly of interest; making the models *statistical* rather than deterministic. There are two distinct approaches to statistical modelling: Frequentist or Classical inference, and Bayesian. This chapter will explain the similarities between these two approaches, and importantly, indicate where they differ substantively. It is typically fairly straightforward to calculate the probability of obtaining different data samples, if we assume that we *know* the process that is responsible for generating these data in the first place. However, we normally do not know these processes with certainty, and it is the goal of statistical inference to derive estimates of the unknown characteristics, or *parameters*, of these mechanisms. Bayesian statistics allows us to go from what is known - the *data* - to extrapolate backwards in order to make probabilistic statements about the overriding parameters which were responsible for its generation. This inversion process is carried out in Bayesian statistics by application of Bayes' rule, which will be 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 the various constituent components of the formula.

2.3 Bayes' rule - allowing us to go from the effect back to its cause

Suppose we knew that a casino was crooked, and uses a die with a probability of rolling a 1 that is twice that of its unbiased value. We could then calculate the probability that we roll two 1s in a row:

$$\begin{aligned} Pr(1,1) &= \frac{1}{3} \times \frac{1}{3} \\ &= \frac{1}{9} \end{aligned} \tag{2.1}$$

Here we use 'Pr' to denote a probability. The probability of $\frac{1}{3}$ is twice that of the unbiased probability: $Pr(1) = 2 \times \frac{1}{6} = \frac{1}{3}$. Do not worry if you don't understand fully 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 - in order to derive the probability of a particular effect - rolling two consecutive 1s. In other words we have calculated $Pr(effect|cause)$. The vertical line, |, here means *given* in probability[1].

Until the latter half of the 17th Century, probability was most frequently used as a method to calculate gambling odds; similar in nature to that

[1] Again do not worry if you don't understand this notation, as we will devote the entirety

shown in (2.1). It was viewed as a dirty subject, not worthy of the attention of the most esteemed mathematicians.

However, the arrival of the Reverend Thomas Bayes, and slightly later and

of chapter 3 to developing an understanding of probability and its notations.

more famously, Laplace[2] started a change in this perspective. They realised that it is possible to move in the opposite direction - to go from effect back to cause:

$$Pr(effect|cause) \xrightarrow{Bayes'\ theorem} Pr(cause|effect) \tag{2.2}$$

In order to take this leap however, it was necessary to create/discover a rule,

[2]See [12] for an interesting history of Thomas Bayes and Laplace, as well as a history of

the use of Bayes' theorem.

which later became known as Bayes' rule[3]. This can be written:

$$Pr(cause|effect) = \frac{Pr(effect|cause) \times Pr(cause)}{Pr(effect)} \tag{2.3}$$

In the casino example, this formula tells us how to 'invert' the original probability $Pr(1, 1|crooked\ casino)$ to get the thing in which we are probably more interested as a patron of said casino $Pr(crooked\ casino|1, 1)$. In words, what is the probability that the casino is crooked, *given* that we rolled two 1s. We will not show how to carry out this calculation in practice now; delaying this until we have learned about probability in chapter 3.

Bayes' rule is central to the Bayesian approach to statistical inference. However, before we introduce Bayesian inference, we first need to explain the purpose of inference!

2.4 The purpose of statistical inference

How much does a particular drug contribute to treatment success? What can an average student earn after obtaining a college education? Will the Democrats win the next US Presidential election? In life we often want to test theories, then go on to draw conclusions.

However, it is often impossible to exactly isolate the parts of a system which we want to examine. The outcome of history is hence determined by a 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 see. Evaluating the return to a college education - there are a range of factors which affect the wage which an individual ultimately commands, of which education is only one element. The outcome of the next US Presidential election depends on party politics, the performance of the incumbent government, as well as the media's portrayal of the candidates.

In life noise obfuscates the signal. The wind, rain, snow and sleet make it difficult to forge a path to where we want to go.

[3] We use the terms "Bayes' theorem" and "Bayes' rule" interchangeably here.

Statistical inference allows us to draw conclusions in this blustery landscape; separating the signal from the noise. It is the logical framework which we can use to trial our beliefs against *data*.

In statistics, we formalise our beliefs in models of *probability*. The models are probabilistic because we assume we are ignorant to many of the multitude of interacting parts of a system, meaning we cannot say with certainty whether something will, or will not, occur.

Suppose we are evaluating the efficacy of a drug in a trial. We might suppose that *on average*, the drug might have a given probability of working as desired. However, before we carry out any trials, we are unaware of its exact treatment success rate. Fortunately, statistical inference allows us to estimate this unknown characteristic, or *parameter*, from the data we are given.

There are two schools of thought for carrying out this process of inference: *Frequentist* and *Bayesian*. Although this book is devoted to the latter, we now spend some time comparing the two approaches, so that a 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 hence assumed to be the outcome of some probabilistic process. Any conclusions that we draw from this approach are based on the supposition that events occur with probabilities, which represent long-run frequencies.

For example, if we flip a coin, we might assume that any sequence of outcomes we obtain is indicative of results we'd get if we were to conduct the experiment an infinite number of times. Further, we might take the proportion of heads observed in this infinite set of throws as defining the probability of obtaining a 'heads'. We suppose that this probability actually exists, and is fixed for each set of coin-throws that we carry out (see figure 2.1).

In general, in Frequentist statistics assume that the data is *random* and results from *sampling*, from a fixed and defined *population* distribution. For

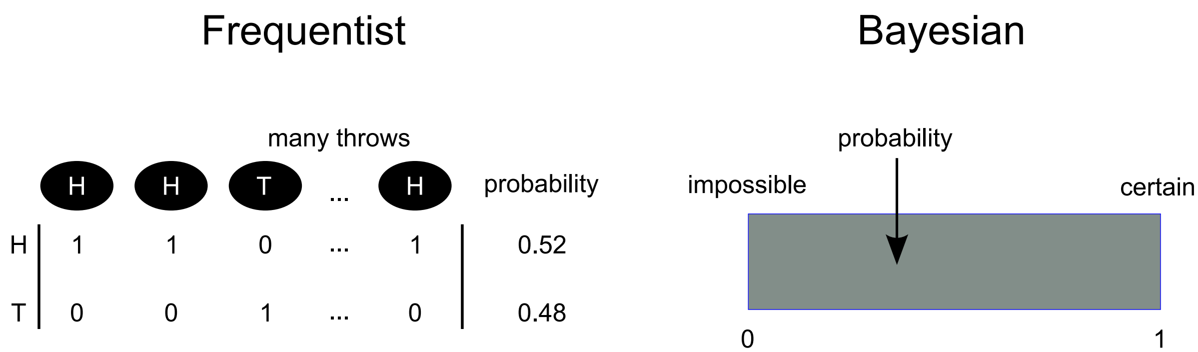


Figure 2.1: The Frequentist and Bayesian approaches to probability.

a Frequentist the noise that obscures the true signal of the population relationship in which we are interested is due to *sampling variation*; the fact that the sample that we pick will each time be 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. Further, if we flip the coin another 10 times, we will likely get a different result, because we have picked a different sample.

2.6 The world according to Bayesians

Bayesians do not imagine repetitions of an experiment in order to define and specify a probability. It is merely taken as a measure of certainty of a particular belief. From this viewpoint, the probability of us 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 concrete entities that actually exist. They are merely abstractions which we can use to help express our uncertainty. In this frame of reference there is no necessity 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 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 a 'population of all possible US elections which happen in the year 2020'?

Probabilities are therefore seen as an expression of subjective beliefs, meaning 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 process, and is central to Bayesian inference, where we aim to express probabilistically our uncertainty in parameters after we have seen the *data*.

Bayesians assume, since we are witness to the data, that it is *fixed*, and therefore does not vary. We do not need to imagine that there are an infinite number of possible samples. We 'see' our data, and hence do not need to view it as the outcome of some random process.

In contrast, we do not ever learn exactly the value of an unknown parameter. This epistemic uncertainty means that in Bayesian inference we choose to view the parameter as a quantity that is probabilistic in nature. We can view this in one of two perspectives. Either we view the unknown parameter as truly being *fixed* in some absolute sense, but our beliefs are uncertain, and thus probabilistic. Alternatively, we can take the view that there isn't some definitive *population* process, and for each sample we take, we get a slightly different parameter.

In the latter perspective we get different results from the coin flipping because each time we are subjecting our system to a slightly different probability of it landing 'heads' up. This could be because we mildly altered our throwing technique, or started with the coin in a different position. In the former perspective, we view the sample as a noisy representation of the signal, and hence we get different results for each set of throws. Although these two descriptions are different philosophically, they are not mathematically, meaning we can apply the same analysis to both.

2.7 Frequentist and Bayesian inference

The Bayesian inference process is the only logical way to modify our beliefs to take into account new data. We start out before we collect data with 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 for probability allows us to calculate our post-data or *posterior* belief:

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

Ignore for the moment that we have not explained was is meant by this mysterious *prior*, as we shall introduce this element properly in section 2.8.2.

In Bayesian inference, we want to draw conclusions based on purely probabilistic descriptions of phenomena. If we wish to summarise our evidence for a particular hypothesis, we describe this probabilistically, as the 'probability of the hypothesis *given* the data obtained'.

The difficulty in obtaining this type of conclusion is that when we write down a probability model describing the situation under examination, we can only use it to compute the 'probability of obtaining our data *given* our hypothesis being true'; the inverse of that which we desire. This probability is calculated by taking into account all the possible samples that we could have obtained from the population, if we assume the hypothesis is true. The issue of statistical inference, common to both Frequentists and Bayesians, is to how to invert this probability to get the desired quantity of interest.

Frequentists stop here, using this probability as evidence for a particular hypothesis in question. If the probability of obtaining the data, or a data sample more extreme, given a hypothesis is small, then it is assumed that it is unlikely that the hypothesis is true, and is rejected. Note however if reject a hypothesis, we have no way of telling whether it is true, and we just witnessed a weird sample, or that it is actually false. Also, note that in this inference process, we have also had to imagine obtaining a sample more extreme than our result, in order to get a usable probability.

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'. There is no need for an arbitrary cut-off in the probability in order to validate the hypothesis; all information is summarised in this probability, and hence can be viewed as the end point of an analysis in itself.

The next few, albeit silly, and (more than) somewhat contrived examples, illustrate a difference both in methodology, but perhaps more significantly,

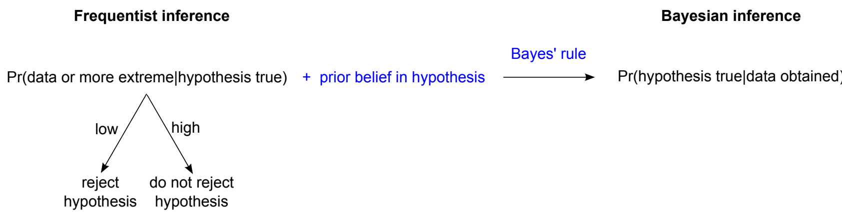


Figure 2.2: Frequentist and Bayesian inference.

in philosophy, between the two different approaches.

2.7.1 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 variation in the usual 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 you have been shown by a security camera to have definitely been in the same house as the victim - Sally - on the night of her demise.

If you choose the Frequentist trial, your jurors start by coming up with a model based on previous trials, which assigns a probability of you being seen by the security camera if you were guilty. They then 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 this 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 are first introduced to an array of evidence, that suggests that you neither knew Sally, nor had any previous record of violent conduct; being otherwise a perfectly respectable citizen. Furthermore, the ex-boyfriend of Sally is a multiple-violent-offending convict on the run from prison after being sentenced by a judge on the basis of witness testimony by Sally. On this basis, the jury determine a *prior* probability in the hypothesis that you are guilty that is $\frac{1}{1000}$[4]. They then use the same model as the Frequentists, to determine that the probability of you being seen by the security camera given your guilt is 30%. However, they then coolly use

[4]Do not worry that we are yet to explain how one can construct priors, since we shall

devote the entirety of chapter 6 to this purpose.

Bayes' rule[5], and conclude that the probability of you 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.

2.7.2 Radio control towers: example

In a hypothetical war two radio control workers sit side-by-side, Mr Pearson (from frequentland), and Mr Laplace (from the county of Bayesdom), 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 airforce base(s) which will respond by sending up aircraft of their own. There are however, administratively two different airforces, which correspond to the two different counties. Although the airforces of frequentland and Bayesdom share airbases, they are distinct, and only respond to Mr Pearson and Mr Laplace's advice respectively. The war, although short, has been costly to both allies, and they each want to avoid needless expenditure, as well as the unwarranted scaring of the local populace by sending up jets.

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

"...The plane is most likely 5 miles from 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 to get himself a well-earned coffee.

Laplace knows from experience there are three different flight paths that the enemy has used to attack previously. 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 that Pearson used. The output this time is different. By using the optional input, the program now outputs a map with the most likely regions shown via a colour shading. There is the highest posterior density over the region near Tunbridge Wells, where Pearson radioed, although the map suggests there are two other towns which might be likely victims of the plane's bombing. Accordingly, Laplace radios to Tunbridge Wells, asking them to send up four jets, and to the other

[5]We will introduce this concept in section 2.3.

two towns, asking them to send up two jets each. At the end of this all, 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 Laplace radioed. The Bayesdom jets intercept the encroaching aircraft, and escort it out of allied airspace. Laplace is awarded a medal in honour of his efforts. Pearson looks on jealously.

2.8 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. As was explained in section 2.4, statistical inference is concerned with estimating characteristics of interest, which we call *parameters*, from a dataset that we have to hand. From this point onwards we will use, θ, to represent the unknown parameter(s) which we are interested in estimating.

The Bayesian inference process uses Bayes' rule to estimate a probability distribution over those unknown parameters, after we witness the data. Do not worry if you do not 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 describe probability distributions as a way of representing uncertainty over unknown quantities.

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

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

Here we use p to indicate a probability distribution which may either represent probabilities, or more usually, probability densities[6]. We shall now devote the next few sections to describing, in short, the various elements of (2.5). Note these only provide a partial introduction, since we will spend the entirety of Part II to an extensive discussion of each of its constituent components.

[6]See section 3.3.2 for a description of their distinction.

2.8.1 Likelihoods

Starting with the numerator on the right hand side of (2.5), we come across the term $p(data|\theta)$, which we call the *likelihood*. 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 write down a statistical model, we can generally calculate the probability of particular outcomes, so this is easily obtained. Imagine that we have a coin that we believe to be fair. By *fair*, we typically mean that the probability of the coin falling 'heads-up' is $\theta = \frac{1}{2}$. If we flip the coin twice, we might suppose that it is reasonable to model the outcomes as independent (see section 3.5), and hence we can calculate the probabilities of the four possible outcomes, by multiplying the probabilities of the individual outcomes together:

$$\begin{aligned}
p(HH) = p(H) \times p(H) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
p(HT) = p(H) \times p(T) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
p(TH) = p(T) \times p(H) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
p(TT) = p(H) \times p(H) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}
\end{aligned} \tag{2.6}$$

Hence, if we obtained a *sample* of two heads, we could write down the corresponding likelihood, $p(HH|\theta = \frac{1}{2}) = \frac{1}{4}$.

Do not worry if you do not fully understand this concept, as we will be devoting an entire chapter to likelihoods in chapter 5.

2.8.2 Priors

The next term in the numerator of the right hand side, $p(\theta)$ is the most controversial[7] part of the Bayesian formula, which we call the *prior* distribution of θ. It is a probability distribution which represents our pre-data beliefs as to the likely values of the parameters in our model, θ. This appears at first to be slightly counter-intuitive, particularly if you are used to the world of classical statistics, which does not require us to state our beliefs

[7] Although this controversy is unwarranted, as we explain in section 2.9.

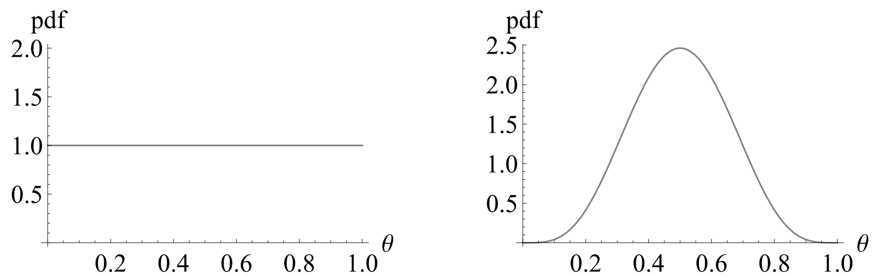


Figure 2.3: Left: All values of a the bias of a coin are equally likely. Right: It is believed that the coin is most likely fair.

explicitly[8]. Continuing the coin example, we might assume that we do not know whether the coin is fair or biased beforehand, so we might think that all possible values of $\theta \in [0, 1]$ - which represents the probability of the coin falling 'heads-up' - are equally likely. We might represent these beliefs by a continuous uniform probability density on this interval (see the left-hand graph of figure 2.3). Normally however, we might think that coins are manufactured such that the weight distribution is fairly symmetrical on either face; meaning that we expect that the majority of coins are reasonably fair. These latter beliefs of unbiasedness might be fairly well represented by the right-hand graph of figure 2.3.

The concept of *priors* will be covered in detail in chapter 6.

2.8.3 The denominator

The final term on the right hand side, on 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 further discussion of this term until chapter 7, when we understand better the significance of *likelihoods* and *priors* respectively. However, for our purposes here it suffices to say that the denominator is fully determined by choice of prior and likelihood function. Whilst it appears simple, this is deceptive, and it is partly the difficulty with calculating this term directly that leads to the introduction of computational methods of the form we will encounter in Part IV.

[8] Although, we always do *implicitly*, as we explain in section 2.9.

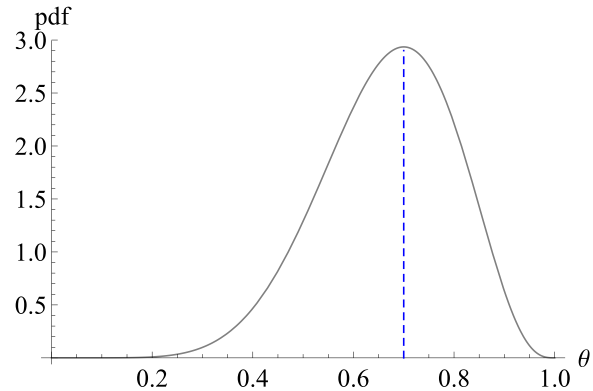


Figure 2.4: The posterior distribution for, θ, the bias of a coin when flipped, assuming a flat uniform prior and Bernoulli likelihood. We assume that 7/10 times the coin came up 'heads'.

The concept of *the denominator* will be covered in detail in chapter 7.

2.8.4 Posteriors: the goal of Bayesian inference

The posterior probability distribution $p(\theta|data)$ is often the main goal of Bayesian inference. For example, we might want to derive a probability distribution representing our post-experimental beliefs of the inherent bias, θ, of a coin, *given* that we flipped it 10 times, and it came up 'heads' 7 times. If we use (2.5), assuming the likelihood model specified in section 2.8.1 and the flat uniform prior shown in figure 2.3, then we would end up with a posterior distribution shown in figure 2.4. Notice that the peak of the distribution occurs at $\theta = 0.7$, which corresponds exactly with the percentage of 'heads' seen in the experiment[9].

The posterior distribution summarises our uncertainty regarding the value of a parameter. If the distribution is more peaked, then this emphasises that there is a greater degree of certainty with a particular value for a parameter. This increased certainty over a parameter value is frequently obtained by collecting more data. In figure 2.5, we compare the posterior distribution for the previous case of 7/10 times a coin appearing 'heads' up, with a new, larger, sample where 70/100 times the same coin comes up 'heads'. In both cases the same ratio of 'heads' to 'tails' appeared, resulting in the same peak

[9] Note that if we chose a non-uniform prior, this peak would most likely shift.

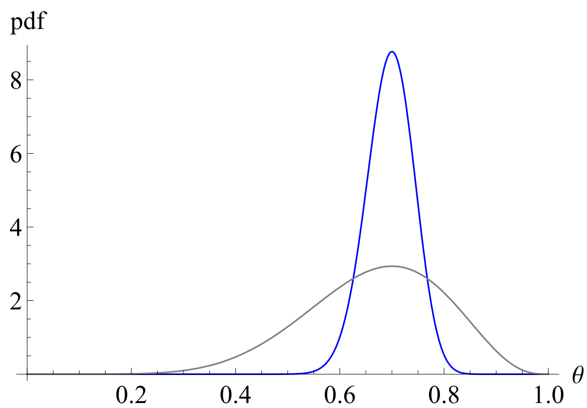


Figure 2.5: Posterior distributions for, θ, the bias of a coin when flipped, assuming a flat uniform prior and Bernoulli likelihood. The grey line assumes that 7/10 times the coin came up 'heads'. The blue line is for the case where 70/100 times the coin came up 'heads'.

value of $\theta = 0.7$. However, in the latter case, since we have more evidence to support our claim, we end up with greater certainty over the parameter value after the experiment.

The posterior distribution is also used as a starting point for prediction of future outcomes of an experiment, as well as for model testing. However, we will leave discussion of these until chapter 4.

2.9 Implicit vs Explicit subjectivity

One of the major arguments levied against Bayesian statistics is that it is by its nature *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 classical analysis.

We argue that *all* analyses involve a degree of subjectivity, which is *implicitly* assumed. In a classical analysis, the statistician typically states a model for probability which depends on a range of assumptions, should be justified explicitly. This process of justification is indicative of the subjective nature of the assumptions on which most analyses rest. For example, the choice to use a simple *linear regression model* in many applied classical analyses assumes that the response of a dependent variable is linear in the model's parameters. This choice of model architecture is generally arbitrary, and used mostly to simplify the analysis.

In science, there is a tendency amongst scientists to use data to suit one's needs, although this practice should really be discouraged (see [9]). This choice as to which data points to include is subjective, and will remain independent of the type of analysis applied.

A further source of subjectivity is in the way in which models are checked and tested. In analyses, both classical 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 a Bayesian analysis allows greater flexibility, and suitable methodologies for these processes, since the prior- and posterior- predictive distributions are straightforwardly manipulated to suit most situations. A Bayesian methodology also allows different models to be compared in a logically-coherent manner, whilst classical analysis relies on fairly arbitrary criteria [10] to do so.

In contrast to the examples of *subjectivity* which we have mentioned above, Bayesian *priors* are *explicitly* stated. This makes this part of the analysis openly available to the reader, allowing it to be as thoroughly interrogated and debated, as any part of an argument. This transparent nature of Bayesian statistics has lead many to suggest that it is *honest*; whilst classical analyses hide behind a fake veil of *objectivity*, Bayesian equivalents explicitly acknowledge the subjective nature of knowledge.

Furthermore, the more data that is collected, the less impact the prior exerts on posterior distributions. In any case, if slight modification to priors results in a different conclusion being reached, it is the job of the researcher to report this sensitivity. In fact, in contrast to classical analyses, a Bayesian analysis allows for a range of models/priors to be stated, which can then be used to test the sensitivity of conclusions to any subjective assumptions made.

Finally, comparing the classical and Bayesian approach to pursuit of knowledge, we find two different solutions; both of which require a subjective judgement to be made. In both cases we would like to have access $p(\theta|data)$ - the probability of the parameter/hypothesis of interest after we have obtained a given data set. In classical hypothesis testing we do not calculate this quantity directly, but use a rule of thumb: we calculate the probability

[10] $\bar{R}^2$, AIC and BIC are examples

that the data would have been more extreme than that which we obtained under a 'null hypothesis'. If the probability is sufficiently small, typically less than a cut-off of 5% or 1%, then we reject the null. Note that this choice of threshold probability - known as a statistical test's *size* - is completely *arbitrary*, and subjective. In Bayesian statistics, we instead use a prior to invert the likelihood from $p(data|\theta) \rightarrow p(\theta|data)$. There is no need to have a null hypothesis and an alternative, since all information is summarised neatly in the posterior. In this way we see a symmetry in the choice of classical *size* and Bayesian priors; they are both attempts to invert the likelihood to get a posterior.

2.10 Chapter summary

This chapter has focused on the philosophy of inference processes in general, and in particular on the philosophical differences between Bayesian and Frequentist inference. We then introduced the Bayesian formula, and provided a short introduction to 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.

2.11 Chapter outcomes

The reader should now be familiar with the following concepts:

1. The goals of statistical inference.

2. The difference in interpretation of probabilities for Frequentists vs Bayesians.

3. The differences in inferential approaches for Frequentists vs Bayesians.

2.12 Problem set

2.12.1 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 the height above a table. Also, in this ideal world, the heights and angles are discrete. However, the system is chaotic[11], and the results of throwing a coin at a given angle and height are shown in table 2.1.

	Height above table (m)				
Angle	**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 2.1: The results of a coin throw from a given angle and height above a table.

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

Now suppose that the some combinations of angles and heights are more likely to be chosen than others, with the probabilities seen in table 2.2.

[11] Highly sensitive to initial conditions.

What are the new probabilities that the coin lands heads-up?

Suppose 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?

Suppose 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?

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

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

Angle	**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	0.00	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

Table 2.2: The probability that a given person throws a coin at a particular angle, and at a certain height above a table.

2.12.2 Model choice

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

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

Fit a quintic (powers up to the 5th) model to the data. How does its fit compare to that of the linear model?

You are now given new data contained within "Intro_PS_overfitLong.csv". This contains data on 1000 replications of the same experiment, where the x values are held fixed.

Fit a linear regression to each of the data sets, and similarly for the quintic model. Which of these performs best?

Using the fits from the first part of this question, compare the performance of the linear regression model, with that of the quintic model.

Hint: do not re-estimate the model with new datasets.

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

If you then found out that the data were years of education (x), and salary in $000s (y). Which model would you favour?

2.13 Appendix

2.13.1 The Frequentist and Bayesian murder trial

In the Bayesian trial the probability that you are guilty given being seen by the security camera on the night of the murder is:

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

In (2.7) we have implicitly assumed that the security camera is hidden, and hence the murderer does not alter his behaviour to avoid being seen; meaning that the probability of being seen by the security camera in each case is 30%.

Chapter 3

Probability - the nuts and bolts of Bayesian inference

3.1 Chapter mission statement

Bayesian statistics formulates models in terms of entities called *probability distributions*. This chapter provides an introduction to all things related to probability; starting at interpretation, and allowing the reader to gain an understanding of how to manipulate these distributions.

3.2 Chapter goals

The logical way to express uncertainty in one's beliefs is through probability distributions, and hence Bayesians aim to derive these as goals of the inferential process. Bayesian inference begins with the prior probability distribution expressing our pre-study beliefs. Bayes' rule then tells us how to update these beliefs in light of data, to produce an updated set of beliefs that we call a *posterior* probability distribution. All information of interest is contained within this latter probability distribution, and are the starting point for making decisions based on the statistical analysis. This chapter takes a step away from Bayesian inference to focus on probability distributions; assuming no previous knowledge of them. In order to understand these abstract objects, we shall first devote some time to explicitly define what is meant by probability distributions. This exercise is also useful since Bayesian inference is based on attempts to invert the *likelihood* to get a proper probability distribution. We will also discuss why the distinction between a likelihoods and probabilities is important. We then explain how to manipulate probability distributions in order to derive quantities of interest; starting with simple 1-dimensional distributions, and working up to more adventurous examples, typical of the variety encountered in Bayesian inference. We finish with a derivation of the Bayesian formula from the law of conditional probability.

3.3 Probability distributions: helping us explicitly state our ignorance

Before we look out the window in the morning, before we get our exam results, before the cards are dealt, we are uncertain of the world that lies in wait of us. In order to plan, as well as make sense of things, we often predict the relative likelihood of different outcomes. However, in order to allow interrogation of thought, with a view to transparency and self-improvement, we sometimes would like to state our pre-conceptions *explicitly*, using a suitable framework.

The mathematical theory of probability provides a logic and language which is suitable to describe the majority of cases in which we are uncertain. Imagine that we enter a lottery, where we select a number from 1-100, to have a chance of winning $1000. We suppose that in the lottery only one ball is drawn, and it is fair with all numbers being equally likely to win. Although we haven't stated this world-view in mathematical notation, we have without realising it, formulated a valid probability distribution for the number drawn in the lottery (see 3.1).

3.3.1 What make a probability distribution *valid*?

The lottery example given in section 3.3 refers to discrete probability distribution, since the variable we were measuring - the winning number - is confined to take on finite set of values. However, we could similarly define a probability distribution where our variable is able to take on an infinity of values across a spectrum. Imagine that before test drive a second-hand car

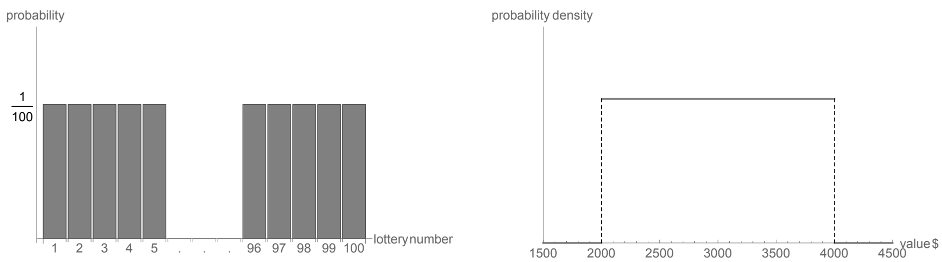


Figure 3.1: Probability distributions representing **left:** the chance of winning a lottery, and **right:** the value of a second-hand car.

we are uncertain about its value. We might think that from seeing pictures of the car, that it could be worth anywhere from \$2000 to \$4000, with all values being equally likely (see 3.1).

The aforementioned examples are both examples of valid/proper probability distributions. So, what are their defining properties?

- All values of the distribution must be real, and non-negative.
- The sum (integral) across all possible values of the discrete (continuous) random variable must be 1.

In the lottery case, this is satisfied since $p(X) = \frac{1}{100} \geq 0$, and:

$$\frac{1}{100} + \frac{1}{100} + ... + \frac{1}{100} = \sum_{i=1}^{100} \frac{1}{100} = 1 \tag{3.1}$$

For the continuous case of the probability of a heads when flipping a coin, the probability distribution is always $\frac{1}{2000} \geq 0$, and when we do the continuous analogue of summing - integrating - we find that:

$$
\begin{aligned}
\int_{2000}^{4000} p(v)\mathrm{d}v &= \int_{2000}^{4000} \frac{1}{2000}\mathrm{d}v \\
&= \left[\frac{1}{2000}v\right]_{2000}^{4000} \\
&= \frac{1}{2000}(4000 - 2000) \\
&= 1
\end{aligned}
\tag{3.2}
$$

Although, it may seem that this definition is relatively arbitrary, and perhaps well-trodden-territory for some readers, it is of *central* importance to Bayesian statistics. This is because Bayesians like to work with, and produce *valid* probability distributions. The pursuit of this ideal underlies the majority of *all* methods in applied Bayesian statistics - analytic and computational - and hence its importance cannot be overstated!

3.3.2 Probabilities vs probability density : interpreting discrete and continuous probability distributions

The discrete probability distribution for the lottery shown on the left hand side in figure 3.1, is straightforward to interpret. To calculate the probability that the winning number, X, is 3, we simply read off the probability from the graph corresponding to the height of the leftmost bar, and find that:

$$p(X = 3) = \frac{1}{100} \tag{3.3}$$

In the discrete case, if we want to calculate the probability that a random variable takes on a range of values, then we simply need to sum the individual probabilities corresponding to each specific event. In the die example, if we want to calculate the probability that the winning number is 10 or less, we just add together the probabilities of it being {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}:

$$
\begin{aligned}
p(X \leq 10) &= p(X = 1) + p(X = 2) + p(X = 3) + ... + p(X = 9) + p(X = 10) \\
&= \frac{1}{100} + \frac{1}{100} + \frac{1}{100} + ... + \frac{1}{100} + \frac{1}{100} \\
&= \frac{1}{10}
\end{aligned}
\tag{3.4}
$$

How can we use the continuous probability distribution such as the one shown on the right hand side of figure 3.1? If we want to calculate the probability that the value of the second-hand car is \$2,500, then we could simply draw a vertical line from this point on the *value* axis up to the line of the distribution; concluding that $p(value = \$2,500) = \frac{1}{2000}$! However, under this logic, we could also deduce that the probability of the value of the car being {\$2,500,\$2,500.10,\$2,500.01,\$2,500.001} are all $\frac{1}{2000}$. Furthermore, we could generate an infinity of these test values of *value*, meaning that if we summed them all together we could get a total probability of ∞.

There is evidently something wrong with our method for interpreting continuous densities. If we reconsider the test values $\{\$2,500, \$2,500.10, \$2,500.01, \$2,500.001\}$, we reason that these are all equally unlikely, and part of a set of an infinity of potential values we could draw. This means for a continuous random variable, we always have $p(\theta = number) = 0$. Hence, when we write $p(\theta)$ for a continuous random variable, we should be careful to interpret the value of it at a particular value as a probability *density*, *not* a probability.

However, we can use a continuous probability distribution to calculate the probability that a random variable lies between two bounds. To do this we use the continuous analogue of a sum, an *integral*. For the car example, we can calculate, $\$2,500 \leq \theta \leq \$3,000$:

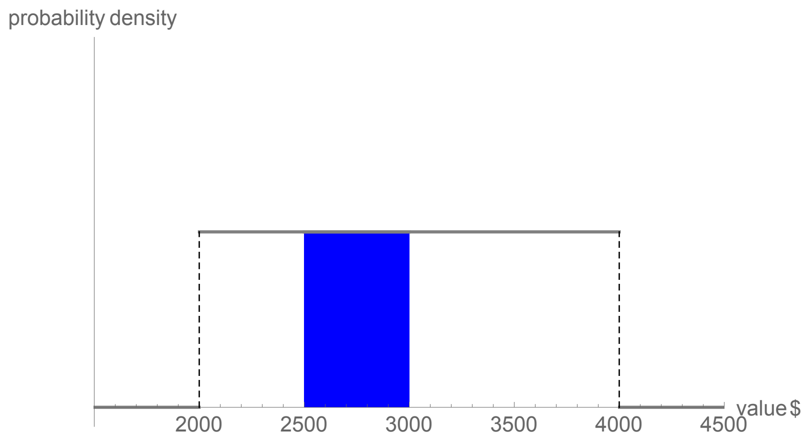


Figure 3.2: The probability that a second-hand car's value lies between $2,500 and $3,000.

$$
\begin{aligned}
Pr(2500 \leq value \leq 3000) &= \int_{2500}^{3000} p(v)\mathrm{d}v \\
&= \int_{2500}^{3000} \frac{1}{2000}\mathrm{d}v \\
&= \left[\frac{1}{2000}v\right]_{2500}^{3000} \\
&= \frac{1}{2000}(3000 - 2500) = 0.25
\end{aligned}
\tag{3.5}
$$

In (3.5), we have used Pr to explicitly state that the result is a *probability*, whereas $p(\theta)$ is a probability density. Of course, the calculation carried out in (3.5), is equivalent to working out the area under the graph within those limits (see figure 3.2).

$\boxed{\textit{Interactive}:}$ see the interactive tool to allow you to dynamically manipulate figure 3.2.

Analogy: stepping stones vs containers

Imagine you wish to cross a fast-flowing river to reach friends on the other side. Fortunately, a helpful person has laid out stepping stones across the river's width. Wanting to demonstrate his mathematical skill, he has made the height of each stone, part of a valid probability distribution. In this discrete environment, the height of each stone exactly determines the probability (see figure 3.3).

By contrast, imagine a filled container of water. The mass of the entire container is set to equal 1. If you were to calculate the probability that a single water molecule lies within a given volume of the container, all we need to do is calculate the mass of that element. We do this by first calculating the element's volume, then multiplying through by the density (in this example, the density is constant). Note here, we can't use probability density directly; we need to calculate a corresponding volume in order to get a probability.

A quick note on terminology: often theorists use probability *mass* to handle discrete distributions, where the distribution's values are directly interpretable as probabilities, and probability *densities* to handle continuous distributions. The latter need to be integrated to yield a probability. We eschew the 'mass' terminology as we find it counter-productive to differentiate between the two types of distributions, since Bayes' rule handles them in the same way (see below).

$\boxed{Video:}$ see the video XXX which explains more about the difference between a probability and a probability density.

The good news: Bayes' rule doesn't distinguish between probabilities and probability densities

Whilst it is important to understand that probabilities and probability densities are *not* the same type of entity, the good news for us is that Bayes' rule is the same for each. So we can readily write:

$$Pr(\theta = 1|X = 1) = \frac{Pr(X = 1|\theta = 1)Pr(\theta = 1)}{Pr(X = 1)} \tag{3.6}$$

for the case of when the data, X, and the parameter θ are discrete, and hence

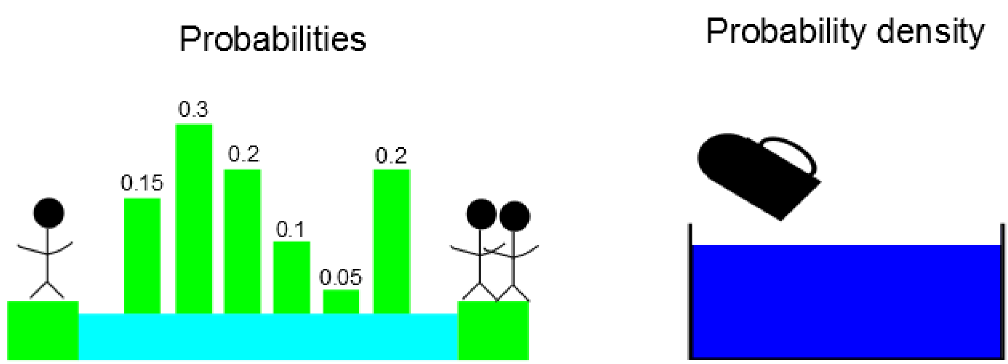


Figure 3.3: Left: a discrete probability distribution, and corresponding probabilities. Right: a probability density.

Pr denotes a probability.

Alternatively, we can write Bayes' rule as:

$$p(\theta = 1|X = 1) = \frac{p(X = 1|\theta = 1)p(\theta = 1)}{p(X = 1)} \tag{3.7}$$

for the case when the data and parameter are continuous, and p denotes a probability density.

We will more commonly use the latter representation since for the majority of cases, the parameters will be continuous.

3.3.3 Mean and variance of distributions

A popular way of summarising a distribution is via its *mean*, which is one measure of central tendency of a distribution. More intuitively, a mean, or *expected value*, of a distribution represents the long-run average value that would be obtained if we sampled from that particular distribution in question, an infinite number of times.

The way in which we calculate the *mean* of a distribution depends on whether it is *discrete* or *continuous* in nature. However, the concept is essentially the same in both cases. The mean is calculated as a weighted sum of the values taken on by the random variable in question, where the weights are provided by the probability distribution. This results in the following forms for the mean of a discrete and continuous variable respectively:

$$\mathbb{E}(X) = \sum_{All\ \alpha} \alpha Pr(X = \alpha) \tag{3.8}$$

$$\mathbb{E}(X) = \int_{All\ \alpha} \alpha p(\alpha) \mathrm{d}\alpha \tag{3.9}$$

In (3.8) and (3.9), α represents the multitude, or continuum of *values* taken on by the random variable X respectively. We have chosen to use *Pr* in (3.8), and p in (3.9), to illustrate that these represent probabilities and probability *densities* respectively.

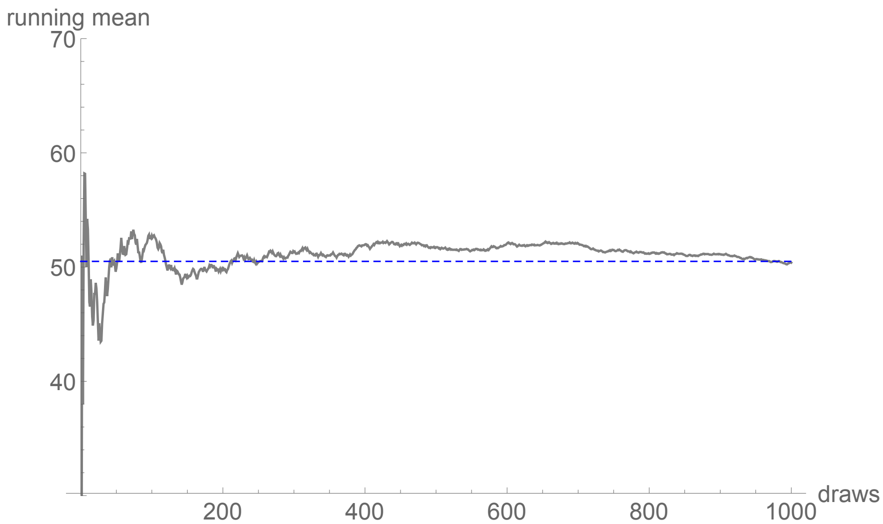


Figure 3.4: Playing a computational lottery. We see the approach of the running mean of repeatedly playing the lottery to the long-run mean of $50\frac{1}{2}$, as the number of plays increases.

We can now apply (3.8) to allow us to calculate the mean winning number from the lottery:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{\alpha=1}^{100} \alpha Pr(X=\alpha) \\ &= 1\times\frac{1}{100} + 2\times\frac{1}{100} + 3\times\frac{1}{100} + ... + 99\times\frac{1}{100} + 100\times\frac{1}{100} \\ &= 50\tfrac{1}{2}\end{aligned} \tag{3.10}$$

We can also demonstrate the *long-run* nature of the mean value of $50\frac{1}{2}$ found in (3.10) by simulating a number of rolls of a fair die computationally (see figure 3.4). As the number of rolls increases, the running mean tends towards this value.

We can also apply (3.9) to calculate our expectation of the second-hand car value:

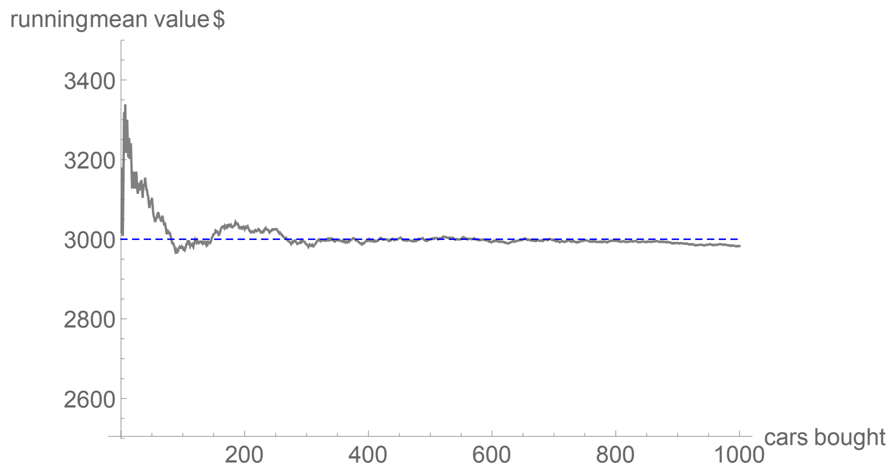


Figure 3.5: Career second-hand car sales. We can see the approach of the sample mean towards the long-run mean of $3,500.

$$
\begin{aligned}
\mathbb{E}(value) &= \int_{2000}^{4000} v p(v) \mathrm{d}v \\
&= \frac{1}{2000} \int_{2000}^{4000} v \mathrm{d}v \\
&= \frac{1}{2000} \left[\frac{v^2}{2} \right]_{2000}^{4000} \\
&= \$3,000
\end{aligned}
\tag{3.11}
$$

If we were to have a business buying (and selling) second-hand cars, we might keep tabs on the values of cars we buy over time. If all cars came from the same uniform distribution that we have proposed, then we would see the sample average value approaching the above long-run mean of $3,000, as the number of cars we buy gets large[1] (see figure 3.5).

If you can grasp the process undertaken to produce figures 3.4 and 3.5 respectively, then you already understand the basis behind modern computational Bayesian statistics! If you need a bit more explanation of the

[1]Technically, tends to infinity.

theory of Bayesian computational, then fear not, we devote an entire Part of the book for this purpose (see Part IV).

Whilst the *mean* of a distribution is a measure of central tendency for a particular distribution, we do not yet have a way of summarising the width of the range of the values of the random variable which are most likely. This motivates the introduction of the concept of a *variance* of a distribution:

$$var(X) = \mathbb{E}\left([X - \mathbb{E}(X)]^2\right) \tag{3.12}$$

To apply (3.12) to discrete and continuous distributions respectively, we straightforwardly replace the α on the right-hand side of (3.8) and (3.9) respectively, by $(\alpha - \mathbb{E}(X))^2$ in each case[2]:

$$var(X) = \sum_{All\ \alpha} (\alpha - \mathbb{E}(X))^2 Pr(X = \alpha) \tag{3.14}$$

$$var(X) = \int_{All\ \alpha} (\alpha - \mathbb{E}(X))^2 p(\alpha) \mathrm{d}\alpha \tag{3.15}$$

If the equations are starting to overwhelm, then fret not, we really only wanted to include them for completeness. What is more important is their significance. Essentially, a *variance* measures the width of the distribution of values obtained around its mean. A wider variance therefore signifies a greater variety of values away from the mean. In figure 3.6 we compare the variability of the fair lottery, with one heavily biased to take on the values between 40 and 60. We see that the variability of the running mean for the biased lottery is smaller than that of the fair one, particularly as the number of rolls increases. This is due to the fact that the fair lottery has a variance of:

$$\begin{aligned} var(X) &= \sum_{\alpha=1}^{100} (\alpha - 50\tfrac{1}{2})^2 \times Pr(X = \alpha) \\ &= 833\tfrac{1}{4} \end{aligned} \tag{3.16}$$

whereas similar calculations for the loaded die distribution shown in the middle of figure 3.6, yield a value of approximately 265. To be concrete,

[2]This is a specific example of the general rule, that to calculate the mean value of some

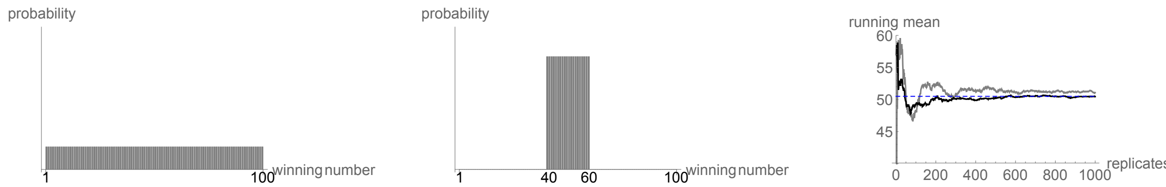


Figure 3.6: Comparing the variance of two lotteries, left: a lottery where all values between 0 and 100 are equally likely. Middle: a lottery where only values between 40 and 60 have a positive probability. Right: comparing the variability of these distributions about their common mean.

the variance of a distribution is an indicator of the long-run average square distance of values away from the mean.

3.3.4 Generalising probability distributions to two dimensions

Life is often more complex than the examples of section 3.3. Often we are tasked with formulating opinions on a range of different outcomes; each of which may influence or shed light on the other results. We begin by considering the outcome of two measurements, in order to introduce the reader to the mechanics of probability. The great thing is that these rules do not become any more complex when we generalise to higher dimensional problems, meaning that if the reader is comfortable with the following examples, then they should be able to handle the vast majority of probability distribution operations encountered. In Bayesian statistics, being comfortable manipulating probability distributions is essential, since the output of the Bayesian formula - the posterior probability distribution - is used to derive all post-experiment conclusions. As such, it is important to devote some time to introduce two examples which we will use to describe and explain the manipulations of 2-dimensional probability distributions.

---

opposed to a sum.

Horses for courses: a 2-dimensional discrete probability example

Imagine that you are a horse racing aficionado, and are interested in quantifying the uncertainty regarding the outcome of two (fictitious) races for two thoroughbreds in a particular stable. From their historical performance over 100 races you notice that both horses tend to react in the same way to the racing conditions. When horse A wins, it is more likely that, later in the day, horse B will win, and vice versa. Similarly regarding the losses; when horse A finds the going tough, so does horse B. Wanting to flex your statistical muscle, you choose to represent this information by the two-dimensional probability distribution shown in table 3.1

		horse A	
		0	**1**
horse B	**0**	$\frac{30}{100}$	$\frac{10}{100}$
	1	$\frac{10}{100}$	$\frac{50}{100}$

Table 3.1: A probability distribution showing the historical performance of two horses, A and B, in two separate races. $\{0,1\}$ refers to each horse losing or winning in their respective races.

How can we check whether this distribution satisfies the requirements for a valid probability distribution? We simply apply the rules described in section 3.3.1. Firstly, all the values of the distribution are real and non-negative; satisfying our first requirement. For the second rule rather than summing over the values of one random variable, we now have to sum over the outcome of two:

$$\sum_{X_A=0}^{1}\sum_{X_B=0}^{1} Pr(X_A, X_B) = \frac{30}{100} + \frac{10}{100} + \frac{10}{100} + \frac{50}{100} = 1 \tag{3.17}$$

function $f(X)$, where X is governed by a particular distribution, we do:

$$\mathbb{E}(f(X)) = \sum_{All\ \alpha} f(\alpha) Pr(X = \alpha) \tag{3.13}$$

for a discrete distribution, and analogously for the continuous case, but using an integral

A probability distribution describing the foot size and scores on a literacy test for an individual within our sample. Left) Represented as a 3-dimensional plot, and Right) Contour lines specify isolines of probability.**Left out figure as it was too big, and was causing latex to crash!**

In (3.17), X_A and X_B are random variables[3] which refer to the outcome of the races for horse A and horse B respectively. Notice that since we are now considering a situation with the outcome of two random variables, we are now required to index the probability, $Pr(X_A, X_B)$, by both. Due to the probability now being a function of two variables, we say that the probability distribution is 2-dimensional.

How can we interpret the probability distribution shown in table 3.1? The probability that both horses lose (and hence both their random variables take on the value of 0), is simply read off from the top-left entry in the table, meaning $Pr(X_A = 0, X_B = 0) = \frac{30}{100}$. We ascribe a smaller likelihood to heterogeneous outcomes, $Pr(X_A = 0, X_B = 1) = \frac{10}{100}$ or $Pr(X_A = 1, X_B = 0) = \frac{10}{100}$, since we believe that the horses are more likely to react similarly to the racing conditions. We believe that the most likely outcome is that both horses win, since historically the horses have done well on this particular racing course, and hence ascribe the highest probability to this result, with $Pr(X_A, X_B) = \frac{50}{100}$.

Foot length and literacy: a 2-dimensional continuous probability example

We suppose that we have a sample of individuals, and we measure their foot size, as well as how well they score on a literacy test. Both of these variables can reasonably be assumed to be continuous, meaning that we are now required to represent our strength of belief, by specifying a probability distribution across a continuum of values (see figure **??**).

We could verify that the distribution shown in figure **??** is in fact valid, by showing that the volume underneath the left hand plot is 1, via integration. However, since we don't want to overcomplicate things now, you will have to take our word for it.

Notice that there to be a degree of correlation between foot size and scores

[3] A function which associates a unique numerical value with each outcome of an experi-

ment. In this case the function gives value 0 if the result is a tails, and 1 if it is heads.

on the literacy test. Why might this be the case[4]?

3.3.5 Marginal distributions

We may be interested in simplifying the preceding analysis, by stating the distribution of one variable, completely *unconditional* of the other. In our horses example, we might be interested in say, only the result of the first horse race, which involves horse A. Alternatively, we might want to remove the dependence on foot size, in our literacy score example, and what remains would then be an *unconditional* probability distribution for literacy score.

In order to do this, we essentially need to *average* out the dependence of the other variable. In our horses example, if we are only interested in the result of horse A, we can sum down the column values for horse B, obtaining the *marginal* distribution of horse A, shown at the bottom of table 3.2.

		horse A		
		0	**1**	$Pr(X_B)$
horse B	**0**	$\frac{30}{100}$	$\frac{10}{100}$	$\frac{40}{100}$
	1	$\frac{10}{100}$	$\frac{50}{100}$	$\frac{60}{100}$
	$Pr(X_A)$	$\frac{40}{100}$	$\frac{60}{100}$	

Table 3.2: The marginal distribution of horses A and B, achieved by summing the values in each column or row respectively.

Hence, we have that the *marginal* probability of horse A winning is 0.6. This value is composed out of the two possible ways in which this *single* event can occur:

$$Pr(X_A = 1) = Pr(X_A = 1, X_B = 0) + Pr(X_A = 1, X_B = 1) \tag{3.18}$$

In (3.18), we see that A can win with B losing, or alternatively both horses can win.

Thus, in order to calculate the probability of a single event, we simply need to sum across all possible occurrences of it, allowing the other variable to take on its possible values. Mathematically, we can summarise this rule by the following for the case of two discrete random variables:

[4]Our sample of individuals here is a sample of children of various ages. Age is correlated

$$Pr(A=\alpha)=\sum_{\beta} Pr(A=\alpha, B=\beta) \tag{3.19}$$

In (3.19), α and β refer to the specific values taken on by the random variables A and B.

We can use (3.19) for the horses example to calculate the probability that horse B loses:

$$\begin{aligned} Pr(X_B=0) &= \sum_{\alpha=0}^{1} Pr(X_B=0, X_A=\alpha) \\ &= Pr(X_B=0, X_A=0)+Pr(X_B=0, X_A=1) \\ &= \frac{30}{100}+\frac{10}{100}=\frac{40}{100} \end{aligned} \tag{3.20}$$

For continuous random variables we need the continuous analogue of a sum, an *integral*, in order to calculate the marginal distribution. Intuitively, this is because the other variable is now able to take on an continuum of values:

$$p_A(\alpha)=\int_{All\ \beta} p_{AB}(\alpha,\beta)\mathrm{d}\beta \tag{3.21}$$

In (3.21), $p_{AB}(\alpha,\beta)$ corresponds to the joint probability distribution of random variables A and B evaluated at $(A=\alpha, B=\beta)$. Similarly, $p_A(\alpha)$ refers to the marginal distribution of random variable A, evaluated at $A=\alpha$. Although it is somewhat of an abuse of notation, for simplicity, from now on we will now write $p_{AB}(\alpha,\beta)$ as $p(A,B)$, and $p_A(\alpha)$ as $p(A)$.

In the foot size/literacy example, we may not be interested in foot size; wanting only the distribution of literacy scores in our sample. We can obtain this by simply integrating out the dependence on foot size:

$$p(score)=\int_{0}^{30} p(score, FS)\mathrm{d}FS \tag{3.22}$$

---

with shoe size and literacy.

The result of carrying out the step in (3.22) is that we are left with the distribution shown on the right of figure 3.8. We have rotated this graph to emphasise that it is the result of essentially summing[5] across the joint density at each particular value of literacy scores.

Another way to think about marginal densities, is imagine that you are walking along the landscape of the joint density. The total distance walked - horizontally and vertically - from $FS = 0$ to $FS = 30$ along a line of constant literacy, gives the height of the marginal density for literacy score at that point. If the path is relatively flat, indicating a low value of joint density, then the corresponding marginal density is low. However, if the path encompasses a large hill, indicating a high value of joint density, then the marginal density will be relatively high.

Add a 3D version of the figure with the contours traced out on the landscape, and leading to the height of the marginals, perhaps with stick figures walking along lines of iso-literacy.

Venn diagrams

An alternative way of thinking about marginal distributions is provided by the Venn diagram shown in figure 3.9. In a Venn diagram, the area of a particular event indicates its probability, and the rectangular area represents all the events that can possibly happen, and so has an area of 1. We have chosen to specify the events of horse A winning, and B winning A as sub-areas the diagram, which overlap indicating a region of joint probability, $p(X_A = 1, X_B = 1)$. In this set-up it is straightforward to calculate the marginal probability of horse A winning, or horse B; we find the area of the elliptic shapes A or B respectively. Considering horse A, when we calculate the area of the entire ellipse, we are implicitly carrying out the sum of the form indicated in (3.19):

$$p(A) = p(A, B) + p(A, not\ B) \tag{3.23}$$

[5]We really mean integrating, but it is more intuitive to think about this in terms of discrete

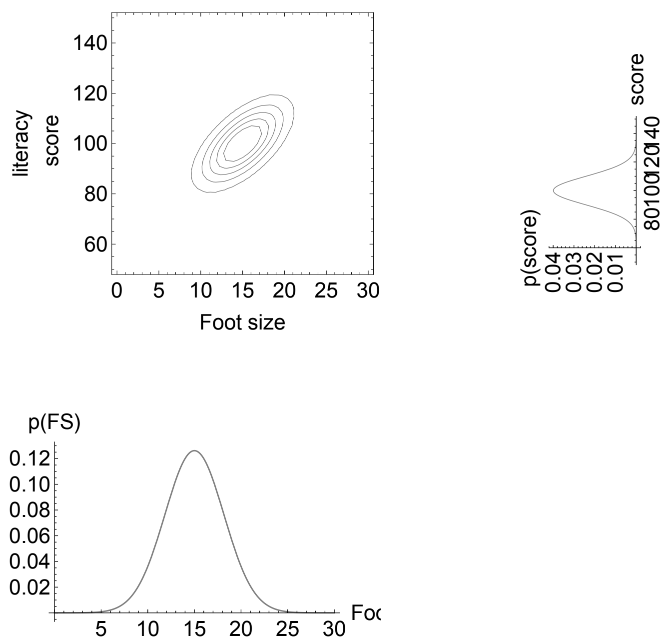


Figure 3.8: Top-left: the joint density of foot size and intelligence. Right: the marginal density of literacy scores. Bottom: the marginal density of foot size. **I want to add a line at a particular value of literacy scores, and at a particular value of FS, to illustrate the horizontal and vertical summing.**

summing.

and are working out the probability that A wins unconditionally[6].

In (3.23), the terms on the right hand side correspond to the overlap region, and the remaining part of A (where B does not occur) respectively.

3.3.6 Conditional distributions

We sometimes only receive partial information by observing part of the system in which we are interested. In horses example, we might only see the result of one horse race, and on this basis update our probabilities of the other horse winning. Alternatively, in the foot size - literacy example described before, we might measure an individual's shoe size, and then want to obtain the updated probability distribution for literacy scores.

In probability, when we observe one variable, and reformulate the probability distribution for the other variable, we say that we are deriving the *conditional* distribution of the latter. *Conditional* refers to the fact that we are deriving the probability distribution of one variable, *conditional* on the value of the other(s).

In each case, we have reduced some of the uncertainty in the system, by observing one of its characteristics. Hence in the two-dimensional examples described above the conditional distribution is only one-dimensional, because we are only now uncertain about one variable.

Luckily, there is a simple rule that we can use to obtain the probability of one variable, conditional on the value of the other:

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{3.24}$$

In (3.24), $p(A|B)$ refers to the probability of A occurring, given that B has occurred. In the right hand side of (3.24), $p(B)$ is the *marginal* distribution of B occurring, and $p(A,B)$ is the joint probability of A and B occurring.

We can use (3.24) for the horses example to calculate the probability that *given* that horse A wins, what is the probability of horse B also winning?

[6] Irrespective of what happens to B.

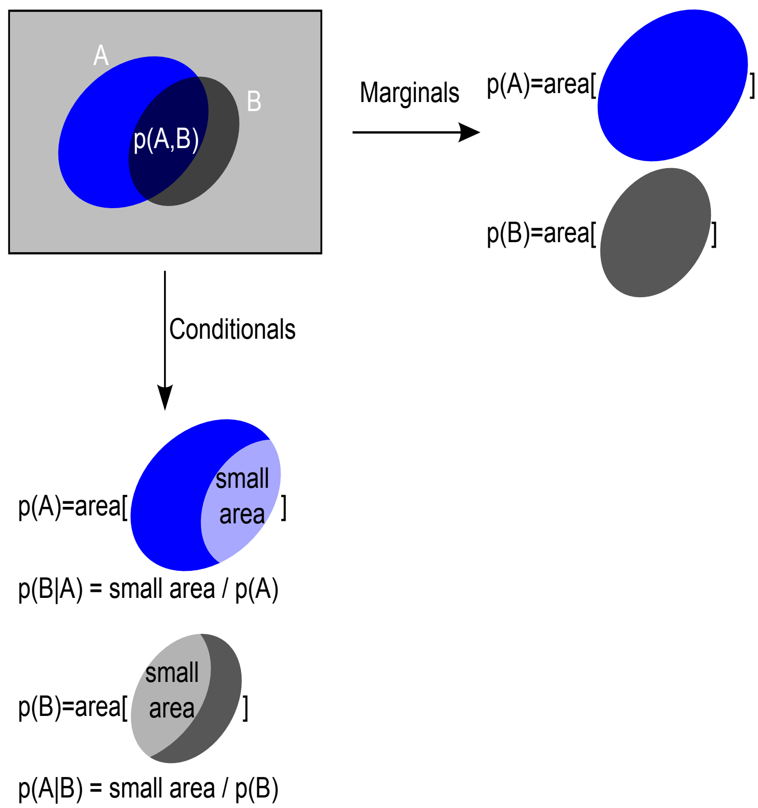


Figure 3.9: A Venn diagram showing one way of interpreting marginal and conditional distributions for the horse racing example.

$$
\begin{aligned}
Pr(X_B = 1|X_A = 1) &= \frac{Pr(X_A = 1, X_B = 1)}{Pr(X_A = 1)} \\
&= \frac{Pr(X_A = 1, X_B = 1)}{Pr(X_A = 1, X_B = 0) + Pr(X_A = 1, X_B = 1)} \\
&= \frac{\frac{50}{100}}{\frac{10}{100} + \frac{50}{100}} \\
&= \frac{5}{6}
\end{aligned}
\tag{3.25}
$$

In (3.25), we have used the rule we discussed earlier for calculating marginal probabilities, shown in (3.19), to calculate the denominator, $Pr(X_A = 1)$ (allowing us to move from line 1 to 2).

Another way to see the workings of this calculation is shown in table 3.3. When we see that horse A wins, we essentially reduce our solution space to only the central column (highlighted in blue). Therefore we need to renormalise the solution space such that it has a probability of 1, by dividing each of its entries through by its original total of probabilities, 0.6; yielding the conditional probabilities shown in the right hand column of table 3.3.

		horse A		
		0	**1**	$Pr(X_B \vert X_A = 1)$
horse B	**0**	$\frac{30}{100}$	$\frac{10}{100}$	$= \frac{10}{100} / \frac{60}{100} = \frac{1}{6}$
	1	$\frac{10}{100}$	$\frac{50}{100}$	$= \frac{50}{100} / \frac{60}{100} = \frac{5}{6}$
	$Pr(X_A)$	$\frac{40}{100}$	$\frac{60}{100}$	

Table 3.3: The highlighted region indicates the new solution space, since we know that horse A has won.

The Venn diagram in figure 3.9 shows another way of interpreting conditional distributions. If we are told that horse B wins, then our event space collapses to only the area specified by B. The conditional probability, $p(X_A = 1|X_B = 1)$ is then simply given by the ratio of the area of overlap between A and B to the total area of B. This makes intuitive sense, since this is the only way that horse A can win, given that B has already won.

We can also use (3.24) to allow us to calculate the conditional distribution of literacy scores for individuals after we have measured their shoe size.

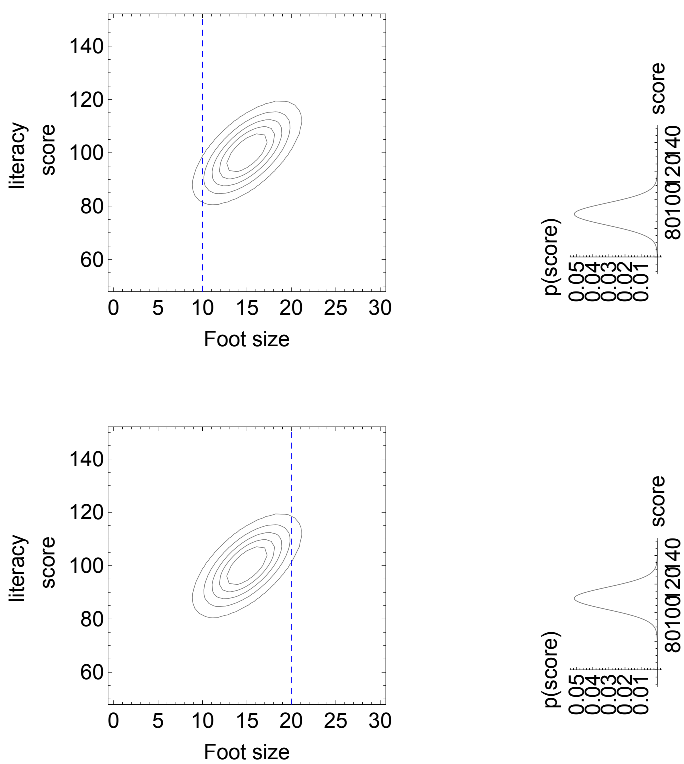


Figure 3.10: The dashed blue lines indicate the new event space in each case. The height walked following these lines is related to the magnitude of the conditional distributions shown on the right.

The only difference with the discrete example is that we now have to use an integral to work out the marginal probability for foot size; the denominator of (3.24). Figure 3.10 shows the conditional distributions traced out when we measure an individual's foot size to be 10cm and 20cm respectively. The blue dashed lines show the new event space, since we have lost our uncertainty over foot size in each of the cases. Therefore the heights traversed on the walk along these lines of constant foot size indicate the relative likelihood of different values of literacy scores.

Add a stick man diagram.

3.4 Higher dimensional probability densities: no harder than 2-D, just looks it!

Now that we are equipped with the tools to calculate marginal and conditional distributions in two dimensions, we can use these to work with probability distributions that depend on any number of variables. Although formulae appear more complex, this is really just a result of having to keep track of each individual variable.

Imagine that we are told that there is another horse C, which comes from the same stable as horses A and B, which will compete in a third race, on the same day as the other two. This probability distribution could be represented by a 3-dimensional array, or alternatively as two separate tables of the same form as table 3.1, one for each outcome for horse C's race. We could write this probability distribution as before, but with a third random variable $p(X_A, X_B, X_C)$. If we were only interested in the result of horses A and B, we can define a vector $\mathbf{Z} = (X_A, X_B)$, meaning that we could write our density as $p(\mathbf{Z}, X_C)$. We can then just apply the same rule as before (in (3.19)) to get the marginal density, because our notation makes it look '2-dimensional':

$$\begin{aligned} p(X_A, X_B) = p(\mathbf{Z}) &= \sum_{X_C=0}^{1} p(\mathbf{Z}, X_C) \\ &= p(\mathbf{Z}, 0) + (\mathbf{Z}, 1) \end{aligned} \tag{3.26}$$

If we wanted to work out the new probabilities of the two horses winning *conditional* on the fact that horse C has won, we can simply use our 'two-dimensional' notation, and equation (3.24):

$$\begin{aligned} p(X_A, X_B|X_C = 1) &= \frac{p(X_A, X_B, X_C = 1)}{p(X_C = 1)} \\ &= \frac{p(X_A, X_B, X_C = 1)}{p(1,1,X_C = 1) + p(1,0,X_C = 1) + p(0,1,X_C = 1) + p(0,0,X_C = 1)} \end{aligned} \tag{3.27}$$

Notice now however, that the denominator $p(X_C = 1)$; representing the marginal probability of horse C winning is actually obtained by a sum over all the four possible ways this can occur: A and B winning, A winning and B not, B winning and A not, and finally both A and B losing[7].

For another example, suppose that we have a continuous posterior density that is defined in terms of three person-specific variables, $p(IQ, FS, W)$; where W represents the weight of an individual. If we wish to determine the posterior solely as a function of IQ and FS, then we start by defining the parameter vector $\boldsymbol{\theta} = (IQ, FS)$, meaning we are left with $p(\boldsymbol{\theta}, W)$. Note that all we have done is defined a new composite variable, $\boldsymbol{\theta}$. Now our density is of the '2-dimensional' form shown in (3.21), and we can apply this relation:

$$\begin{aligned} p(IQ, FS) &= p(\boldsymbol{\theta}) \\ &= \int_{All\ W} p(\boldsymbol{\theta}, W)\mathrm{d}W \end{aligned} \tag{3.28}$$

If we wish to find the marginal distribution for *IQ only*, then all we do is integrate the resultant distribution in (3.28) with respect to *FS*.

We can use exactly the same trick to calculate the *conditional* density of *IQ* and *FS*, conditional on observing weight:

$$\begin{aligned} p(IQ, FS|W) &= p(\boldsymbol{\theta}|W) \\ &= \frac{p(\boldsymbol{\theta}, W)}{p(W)} \end{aligned} \tag{3.29}$$

In (3.29), we have used (3.24) to arrive at the second line.

[7] I have shortened the notation to allow all the outcomes to be shown on a single line, so

Finally, note that combining all parameters of interest into a single parameter vector $\boldsymbol{\theta}$ allows us to calculate *marginal* and *conditional* distributions for probability distributions that depend on an arbitrary number of parameters.

3.5 Independence

If we think that there is a relationship between two random variables, then we say that they are *dependent*. This does not necessarily mean *causal* dependence, as it is sometimes supposed, in that the behaviour of variable A affects the outcome of variable B. What it really means is that the value taken on by A is informative for predicting B.

An example of dependence might be the *colour* and *suit* of a playing card. If we are told that the colour of a playing card is *red*, this means that our other variable *suit* is constrained to be either *hearts* or *diamonds*. In this case, knowing with certainty the value of the first variable, *colour*, helps us to narrow down the list of outcomes of the second variable, *suit* (see figure 3.11).

Another example of dependent variables, is the weather outside and the

$p(X_A = 1, X_B = 1, X_C) = p(1, 1, X_C)$, for example.

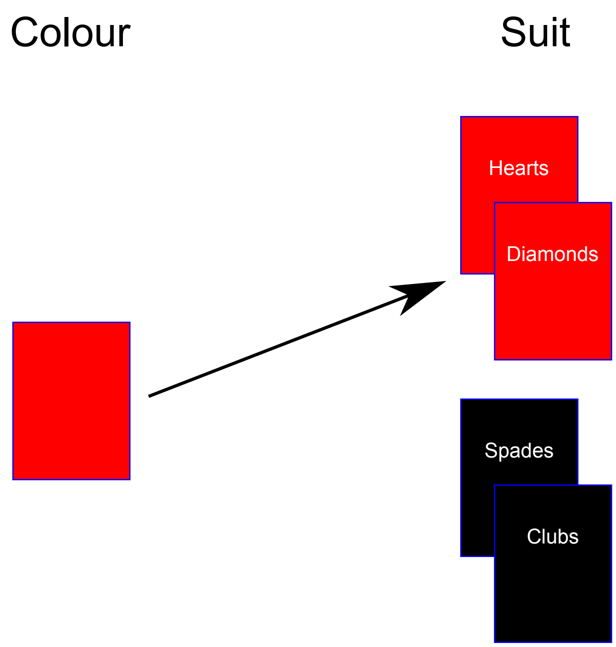


Figure 3.11: Knowledge of the colour of a card provides information about the suit of the card. The colour and suit of a card are *dependent*.

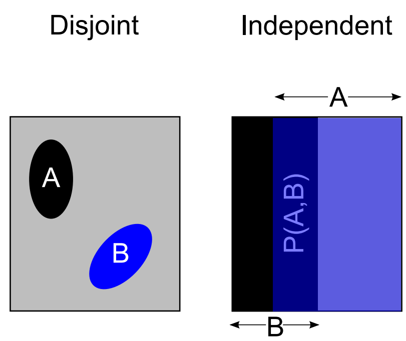


Figure 3.12: Venn diagram depictions of left: disjoint, and right: independent, events A and B.

suntan of a particular, vein, individual[8]. If it is sunny, then we assume it is more likely that an individual is tanned. Whereas, if the weather is cloudy, it is less so.

If two variables, A and B are *disjoint*, then if one occurs, then the other cannot. In this case, it is often mistakenly believed that the variables are *independent*, although this is very much not the case (see the left hand panel of figure 3.12). In this case, knowledge of variable A, provides significant information about variable B. If A occurs, then we know for *certain* that B cannot!

By contrast, if two events are *independent*, then knowledge of B provides no additional information on A. Mathematically, this means that the conditional probability of A is equal to the marginal:

$$p(A|B) = p(A) \tag{3.30}$$

[8]Discounting sunbeds.

Using our conditional probability rule given in (3.24), we can then use this to rewrite the above as:

$$\frac{p(A,B)}{p(B)} = p(A) \tag{3.31}$$

In words, the ratio of the area of overlap between A and B to the area of B, is the same as the overall probability of A (see the right hand panel of figure 3.12). This makes intuitive sense, since uncovering that B has occurred (being in B) should result in no change to the probability of A occurring (now $p(A|B)$).

Another way of stating independence that is commonly used, is obtained by multiplying (3.30) by its denominator:

$$p(A,B) = p(A) \times p(B) \tag{3.32}$$

To make this idea more concrete, we can think again of our horses example. Imagine that now are are considering two horses C and D, that come from separate stables, and race on different days. Using historical race results we have come up with the probability distribution shown in table 3.4.

We can use this table to test whether or not the results for the two horses are independent using (3.32). We should be able to get the joint probabilities of both C and D winning from multiplying together the individual marginal densities:

$$\begin{aligned} p(X_C = 1, X_D = 1) &= 0.3 \\ &= p(X_C = 1) \times p(X_D = 1) \quad = 0.6 \times 0.5 = 0.3 \end{aligned} \tag{3.33}$$

which we see is true. We could similarly also validate this by checking the other three joint outcomes in the table, and find that this is the case.

3.5.1 Conditional independence

In statistics we often come across situations where we suppose that observations are *conditionally* independent. This is often swept under the carpet, and in discussion authors (falsely) state that the observations are *independent*. Suppose that we work for the WHO, as an epidemiologist, and are

		horse C		
		0	**1**	$Pr(X_D)$
horse D	**0**	0.2	0.3	**0.5**
	1	0.2	0.3	**0.5**
	$Pr(X_C)$	**0.4**	**0.6**	

Table 3.4: The probability distribution for horses C and D. The marginal distribution of horses C and D are achieved by summing the values in each column or row respectively.

interested in estimating the proportion of obese individuals within a distinct urban area. Within this are, we can suppose that there exists a true proportion of this condition, θ. Suppose we select an individual uniformly at random from this population, and record their status, X_1, for this disorder. We then pick another individual, and record their status, calling it X_2. If we knew θ, then we might suppose that the outcomes of these two picks are independent *conditional* on this overall parameter. In other words, the outcome of the first sampling provides us *no* further information on the disorder status of the second.

This conditional independence property has frequently be represented in diagrammatic form of the type shown in figure 3.13. The arrows represent a dependency. When viewed in this light we see that it is incorrect to view the disease statuses of individuals 1 and 2 as *fully* independent, since they are linked through a joint dependence on θ. However, note that there is no direct arrow joining X_1 and X_2; this means that the random variables are *conditionally* independent, when we allow for a joint dependence on θ.

3.6 Central Limit Theorems

Life is rarely simple, and in stating statistical models of phenomena we are attempting to at best, *approximate* what is actually happening. In this environment, anything that brings a degree of certainty and coherence to our models is most welcome.

Imagine that we are tasked with coming up with a model of probability for the *mean* IQ test score in a particular school. Furthermore, as a simplification, we imagine that IQ is constrained to lie in the range $[0, 300]$, and

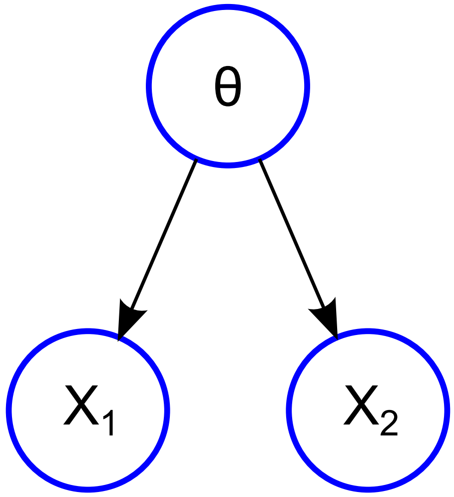


Figure 3.13: A diagrammatic depiction of conditional independence.

we believe that a reasonable probability distribution for an individual's test score is uniform on this range[9] (see figure 3.14). We also suppose that individuals' scores are independent of one another.

Before we consider a sample size of N, we imagine that we only have a sample of two individuals, and we are asked to describe the distribution for the mean of their scores. If we use our assumption of uniformity, we might then ask, whether any mean scores are more likely than others, or are *all* values equally likely, as per the individual case? We start by considering the extremes: there is only one way to achieve a mean test score of 300; both individuals would have to had scored 300. Similarly, to obtain a mean of 0, both individuals must score 0. However, consider obtaining a mean of 150.

[9]Albeit this isn't a particularly good model, but suspend your disbelief for this thought

experiment.

This could have been obtained with a number[10] of combinations of scores, for example $(score_A, score_B) =: (150, 150)\ ,\ (100, 200)\ ,\ (125, 175)$. Intuitively, there are many more ways of obtaining moderate values for the sample mean, that there are for the extremes.

This tendency towards the centre increases the more values we sum over, since in order to get extreme values we would require *more* individual scores to be simultaneously extreme; which is less likely. We can see this increasing central-tendency in figure 3.14, as the number of points being summed over increases.

However, we also see another tendency of the probability densities: *an increasingly good fit of the normal distribution to these sample averages.* This approximation, it turns out, becomes *exact* in the limit that we have an infinite sample, and is known as the *Central Limit Theorem*. Although for our purposes, it will often be practically-exact if the sample size is sufficiently large.

There are in fact a number of central limit theorems, of which we have only exposed you to the one for the average of independent, identically-

[10]Technically an infinity.

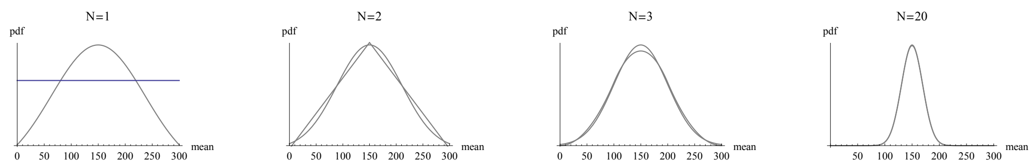


Figure 3.14: The convergence to a normal distribution for the mean of a sum of uniform distributions for IQ. The pdf for the average is shown in blue, with a normal distribution of the same mean and variance indicated in grey.

distributed random variables[11]. However, what is important for you to note is that, whenever we have a number of factors, which additively result in an outcome, then an assumption of normality may be reasonable.

For example, there is an argument which would suggest that an individual's intelligence is the result of a number of factors including: un-bringing, genetics, life-experience, and health amongst others. Hence, we might tentatively propose that an individual's test score picked at random from the population would actually be normally-distributed! This is why I earlier discussed that the assumption that individuals' test scores were uniform might not be reasonable.

3.7 The Bayesian formula

We first of all rewrite the conditional probability formula (3.24), regarding the probability of event A occurring, *given* that event B has occurred:

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{3.34}$$

However, we could also swap A and B around leading to the following for the probability of B *given* that A has already occurred:

$$p(B|A) = \frac{p(B,A)}{p(A)} \tag{3.35}$$

[11]Known as the Lindeberg-Levy CLT. **Note: need an accent on the e.**

We however reason from the Venn diagram in figure 3.9, that the overlap region of $p(A,B)$ really translated means, the probability of A *and* B occurring. This means that this is exactly the same as the reverse; the probability of B *and* A coinciding, $p(B,A)$. We can therefore rearrange (3.35) for this joint probability:

$$p(A,B) = p(B|A) \times p(A) \tag{3.36}$$

We can use (3.36) to break down the probability of both A and B occurring into two steps. Firstly, for this to happen we require that A *must* happen, with its corresponding probability $p(A)$. Then for both to occur, we straightforwardly require the probability of B occurring, *given* that A has already occurred, which is given by $p(B|A)$. This reasoning provides a little intuition as to the workings of the conditional probability law that we wrote down in (3.24).

We can finally substitute (3.36) into the numerator of the fraction in (3.34), to yield the famous Bayesian formula!

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{3.37}$$

The Bayesian formula importantly tells us how to correctly convert from $p(B|A)$ to its inverse $p(A|B)$, which is central to Bayesian statistics.

3.7.1 The intuition behind the formula

If we multiply both sides of (3.37) by $p(B)$, we arrive at the following alternative statement of Bayes' rule:

$$p(A|B) \times p(B) = p(B|A) \times p(A) \quad [= p(A,B)] \tag{3.38}$$

In (3.38), we have added the final part in square parentheses due to the reasoning of section 3.7, that both sides are equivalent to the joint probability of A and B.

What the relation (3.38) tells us however, is that there are two ways of arriving at this joint probability (see figure 3.15). The first way is given by the left hand side, and is due to B occurring, with probability $p(B)$,

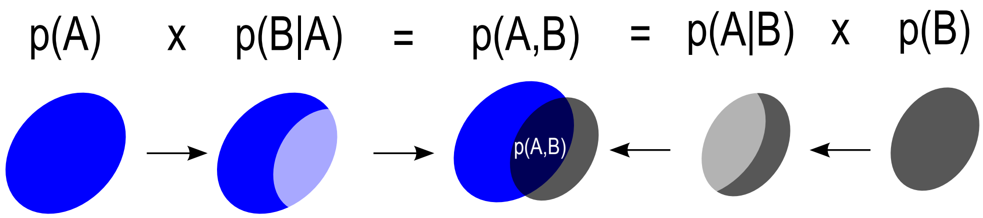


Figure 3.15: The two ways of arriving at the joint probability $p(A, B)$; providing some intuition behind Bayes' rule.

followed by A *given* that B has occurred, with probability $p(A|B)$. An exactly equivalent way route to both A and B occurring is given by the right hand side of the leftmost equals sign. Here we require that A occurs first, with probability $p(A)$, followed by B *given* that A has occurred, with probability $p(B|A)$.

3.7.2 Breast cancer screeing

Let's now make this discussion more concrete by means of an example. Let's imagine that we are an oncology doctor, working in breast cancer diagnosis. We suppose that out of all women aged forty who participate in screenings, about 1% of them will have breast cancer at the time of testing. We suppose that the screening process is relatively robust, and it is known that for women who have breast cancer, the tests will indicate a positive result 80% of the time. However, there is also the risk of false-positives, with 10% of women without breast cancer also testing positive[12].

We now suppose that we are in the position where a woman has tested positive. What we would like to do, is work out the probability that she has breast cancer.

In the language of probability we would like to work out the conditional probability: $p(cancer| + ve)$. In words, the probability that she has breast cancer, *given* that she has tested positive. However, summarising the information we currently have in probability language: $p(cancer) = 0.01$, $p(+ve|cancer) = 0.8$, and finally $p(+ve|no\ cancer) = 0.1$. How can we proceed? Bayes' formula to the rescue:

[12] I am not necessarily indicating clinically-up-to-date values, more I am using these example values to indicate the importance of reducing false-positives of any medical-test.

$$\begin{aligned} p(cancer|+ve) &= \frac{p(+ve|cancer) \times p(cancer)}{p(+ve)} \\ &= \frac{p(+ve|cancer) \times p(cancer)}{p(+ve, cancer) + p(+ve, no\ cancer)} \\ &= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} \approx 0.08 \end{aligned} \tag{3.39}$$

In (3.39), we got from the first line to the second by using (3.19) to work out the marginal probability of an individual testing positive. The fact that the probability is so small is due to the risk of false positives (see video XXX for a fully-intuitive explanation of this), which produce many more positive diagnoses than the real-positives (because the risk of cancer is much lower than the probability of not having cancer).

3.8 The Bayesian inference process from the Bayesian formula

In Bayesian statistics we aim to use probability distributions to describe all components of our system. Our starting point is Bayes' rule:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \tag{3.40}$$

In statistics we are typically looking to estimate a number of parameters, which we will from now on call θ, which represent a component of a statistical model which we build to represent a particular phenomenon. These parameters can be real (such as the proportion of individuals within a given population that have a disease), or mere abstractions (for example the scale parameter of a hyper-distribution).

In Bayesian statistics, we want to update our beliefs about values of a parameter *given* that we have obtained a particular sample of data. Being Bayesians, we would like to represent these beliefs via a probability distribution, which we write as $p(\theta|data)$. However, we can use (3.40), (if we associate A with θ, and B with the *data*) to write:

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{3.41}$$

Although we have straightforwardly made two substitutions to arrive at (3.41) from (3.40), what have we gained by doing so? Also, what exactly do the terms on the right hand side actually mean? Although we introduced these in chapter 2, we will spend the next part of this book examining each of these components in detail.

3.9 Chapter summary

The reader should now have developed a working understanding of probability distributions. Of particular importance in Bayesian statistics are the concepts of marginal and conditional distributions introduced in sections 3.3.5 and 3.3.6 respectively. If you do not feel fully confident with probability distributions, do not fret, since we will have ample opportunity to work with these mathematical objects in the next part of the book, where we provide an in-depth discussion of the various elements of the central formula of Bayesian inference: Bayes' rule.

3.10 Chapter outcomes

The reader should now be familiar with the following concepts:

1. The conditions under which a probability distribution is valid.
2. The difference between probability and probability density.
3. Summary distribution measures, and how to calculate them.
4. Two-or-more dimensional probability distributions.
5. Marginal distributions and how to calculate them.
6. Conditional distributions and how to calculate them.
7. Independence,

8. Central Limit Theorems.

9. The Bayesian formula.

3.11 Problem set

3.11.1 The expected returns of a derivative

The returns of a particular stock are thought to be well described by a log-normal distribution:

$$log(S_t) \sim N(\mu, \sigma^2) \tag{3.42}$$

What are the expected returns of the stock?

What are the variance in returns?

The performance of a particular type of derivative is given by the following formula:

$$D_t = S_t^2 \tag{3.43}$$

Would you expect the return to be equal to the mean of the original stock squared? If not, why not?

Note to myself: the convexity of the squared function means that it is greater.

What would be a fair price to pay for this derivative?

An alternative stock is probabilistically related to the performance of S_t:

$$R_t \sim N(S^2, 1) \tag{3.44}$$

What is a fair price to pay for this stock?

Which asset is more risky, D_t or R_t?

The data file XXX contains Stock return data for the stock S_t. A way to estimate the parameter is via a method known as *method of moments*. Here, we might estimate the parameters μ and σ by equating the sample mean and variance, to the *population* quantities we believe our process abides by.

Use *method of moments* to estimate μ and σ. Note: this requires use of a computer with mathematical software, as analytic solutions aren't possible.

Compare the estimated model with the data.

Is the model a reasonable approximation to the data generating process?

If not, suggest a better alternative.

3.11.2 The boy or girl paradox[13]

Suppose we are told the following information:

Mr Bayes has two children. The older child is a girl. What is the probability that both children are girls? Mr Laplace has two children. At least one of the children is a girl. What is the probability that both children are girls?

3.11.3 The Bayesian game show

A game show presents contestants with four doors: behind one of the doors is a car worth \$1000; behind another is a forfeit whereby the contestant must pay \$500 out of their winnings thus far on the show. Behind the other two doors there is nothing.

The order of the game is thus:

1. The contestant chooses one of four doors.

[13] First introduced by Martin Gardner in 1959.

2. The game show host opens one of the empty doors.
3. The contestant is given the option of changing their choice to one of the two remaining doors.
4. The contestant's final choice door is opened, either to their delight (a car!), dismay (a penalty), or indifference (nothing).

Assuming that:

1. The contestant wants to maximise the probability that they win the car.
2. The contestant is risk averse.

Find the optimal strategy for the contestant.

If you are unsure as to how to proceed, why not simulate the game using a computer. The program should only be a few lines, and might provide you with some intuition as to what is happening.

3.11.4 Blood doping

Suppose as a benign omniscient observer, we tally up the historical cases where professional cyclists used/didn't-use blood doping, and either won or lost a particular race. This results in the probability distribution shown in table 3.5.

What is the probability that a professional cyclist wins a race?

What is the probability that a cyclist wins a race given that they have cheated?

What is the probability that a cyclist is cheating given that he wins?

Now suppose that drug testing officials have a test which is relatively accurate in finding cheats. It accurately indicates a blood-doper 90% of the time. However, it incorrectly indicates a positive for clean athletes 5% of the time.

	Lost	**Won**
Clean	0.7	0.05
Doping	0.15	0.15

Table 3.5: The historical probabilities of behaviour and outcome for professional cyclists.

Should the officials test all the athletes or only the winners, for the cases when:

1. They care only about the proportion of people correctly identified as dopers.
2. They want only to minimise the proportion of falsely accused athletes.
3. They care only about the *number* of people correctly identified as dopers.
4. They care five times about the *number* of people who are falsely identified as they do about the *number* of people who are correctly identified as dopers.
5. What factor would make the officials choose the other group? (By factor, we mean the number 5 in the previous question.)

Part II

Understanding the Bayesian formula

3.12 Part mission statement

This part introduces the reader to the various elements of the Bayesian inference formula: the posterior, the likelihood, the prior and the denominator.

3.13 Part goals

The goal of Bayesian inference is to calculate the posterior distribution for parameters of interest. This distribution is the starting point for making decisions, and drawing conclusions about situations under examination. It is thus important to understand why this part of Bayes' formula is so useful, and hence we devote chapter 4 to this purpose. The first term on the right-hand-side of Bayes' formula that we encounter is the likelihood function, introduced in chapter 5. These functions/distributions allow for variance in the data, even if the parameters are fixed. This variance is what classical statisticians call sampling variation. Importantly, likelihood functions are not valid probability distributions, and are actually the inverse of what we desire. Bayes' rule essentially tells us how to invert these distributions in order to get our desired distributions - the posterior. However, in order to invert the likelihood, we need to specify the final piece of the puzzle - the prior distribution. This is a measure of our pre-data confidence across all the possible values of the parameter. We are required to specify this term, since Bayes' rule only tells us how to *update* our beliefs, not how to specify them in the first place. Priors are without doubt the most controversial part of Bayesian inference, although as we have seen in chapter 2, and we advocate in chapter 6, this criticism is unwarranted. The final part of the formula - the denominator - is determined fully by our choice of likelihood and prior. However, this predetermination does not make out lives simple, and we shall see that the difficulty in evaluating this term, along with similar expressions, motivates the computational methods introduced in part IV.

Chapter 4

The posterior - the goal of Bayesian inference

4.1 Chapter Mission statement

At the end of this chapter the reader will understand the central importance, and use of the posterior probability distribution in Bayesian statistics.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{4.1}$$

4.2 Chapter goals

Calculating the posterior distribution for a model's parameters is the focus of Bayesian analysis. This *probability distribution* which results from the application of Bayes' rule (see 4.1) can be used to infer the effects of given variables, to forecast, compare different models of phenomena, as well as test its own foundations! In order to do justice to the multitude of uses of the posterior distribution, it is necessary that the reader is familiar with the basics of probability distributions explained in chapter 2.

4.3 Expressing uncertainty through the posterior probability distribution

Unlike looking out the window, getting exam results, or playing a hand at blackjack, we frequently in inference never learn the *true*[1] state of nature. The uncertainty here is both in the future and *present*; the latter meaning we are unable to perfectly measure the state of the world today, and hence cannot hope to perfectly know the former.

A way of representing our ignorance, or uncertainty in a parameter's value is through probability distributions. For example, suppose that we wanted to know the proportion of individuals that would vote for the democrats in an upcoming election. We might, on the basis of past exit poll surveys,

[1]Whether a true value for a parameter actually exists we leave until section 4.3.3.

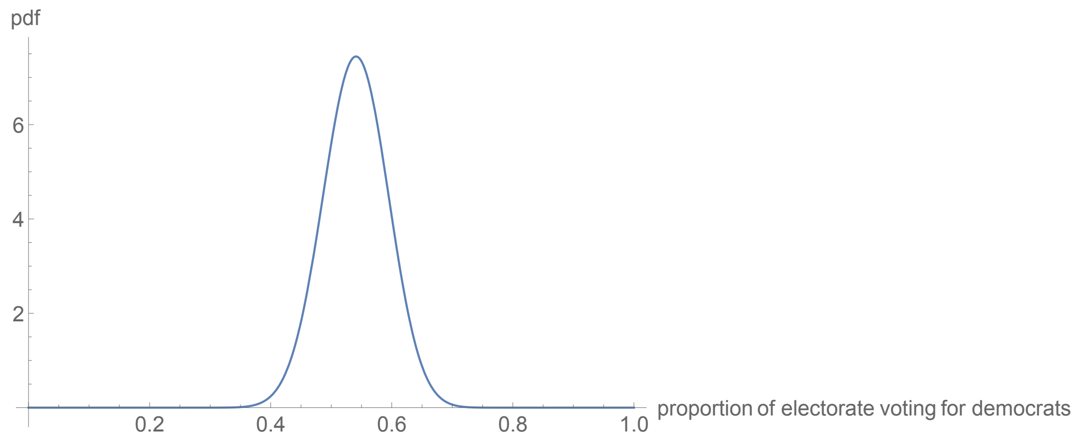


Figure 4.1: A probability distribution representing uncertainty over the proportion of the electorate that will vote for the Democrats in an upcoming election.

calculate a posterior[2] uncertainty which is represented by the probability distribution which is shown in figure 4.1.

How can we interpret the probability distribution shown in figure 4.1? And further, how can we use it to express our uncertainty to a non-mathematician?

Often we describe a distribution by its summary characteristics. These are aspects of the distribution that we commonly want to know. For example, we normally want to know the *mean* value of a parameter. This is a measure of central tendency of our estimates, that is essentially a weighted mean (where the weights are provided by the values of the probability density function). If we have the mathematical description of the distribution shown in figure 4.1, we can calculate this by simply finding its mean, by taking the expectation:

$$\mathbb{E}[\theta] = \int_0^1 p(\theta)\theta \mathrm{d}\theta = 54\% \tag{4.2}$$

This provides a point estimate of the proportion of individuals - 54% - that we expect to vote for the Democrats, which may be a useful piece of

[2]Via Bayes' rule - don't worry if you don't know how to do this, that is the goal of this

information to pass on to an interested party.

A point estimate of the proportion of individuals that we expect to vote for the Democrats is not useful in itself, (and quite dangerous to pass on), without some measure of our inherent confidence/uncertainty in this particular value. One measure of uncertainty in a parameter's value is its variance:

$$var(\theta) = \int_0^1 p(\theta)(\theta - (E[\theta])^2 \mathrm{d}\theta \tag{4.3}$$

In many cases, it is easier to understand the meaning of uncertainty if it is expressed in the same units as the mean, which we do by taking the square root of the variance, yielding 5.3% for the case of figure 4.1. A larger variance indicates that we view a wider range of outcomes as being feasible. In this case, a wider variance would mean that we would be less surprised if the electorate voted in the Republican party.

In sections 4.3.5 and 4.3.5 we will introduce other summary measures of distributions, that are also often presented in research articles, and books. However, the important thing to note is that all of these are derived from the posterior distribution for our parameters.

Another example of the use of posteriors can be illustrated by a regression example. Suppose that we are investigating the effect of military participa-

---

whole first part!

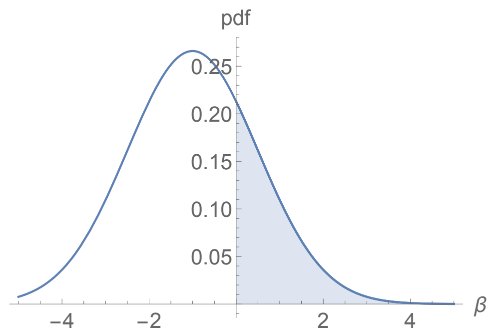


Figure 4.2: An estimated posterior probability distribution for the parameter β in (4.4). The shaded region represents the posterior probability that the parameter is positive.

tion on lifetime earnings[3]. We suppose that the effect is, on average, negative due in part to the psychological stresses of warfare. We also suppose that the effect is can be modelled as being linear, and hence our relationship of interest can be represented as:

$$LI_i = \alpha + \beta MP_i + \epsilon_i \tag{4.4}$$

where we expect that β is negative. Here ϵ_i represents the myriad of other factors that are important for determination of an individual's income.

If we formulate a Bayesian model, with a given prior and likelihood, then we might end up with a posterior density for β, that is shown in figure 4.2. We can use this density to help us to calculate the posterior probability that the given parameter is in fact non-negative, by finding the area under the curve corresponding to $\beta \geq 0$ (shown as the shaded region in figure 4.2), which we find in this case to be approximately 25%. This suggests that we are not all that confident in the fact that the parameter is negative (at 75%), and cannot reasonably go along with our hypothesis.

4.3.1 Bayesian coastguard: introducing the prior and the posterior

Imagine you are stationed in a radio control tower at the top of a steep cliff, in the midst of a stormy night. The tower received a distress call over radio

[3]See Angrist's fantastic 1990 article for a detailed study of this effect for Vietnam war

from a ship - The Frequentasy - which has got engine trouble, somewhere out in the bay. It is your job to direct a search helicopter to rescue the poor sailors.

When you first receive the weak crackled radio signal, you are not made aware of their location. However, you know that the boat must be somewhere in the bay, less that 25km away from the tower, since this is the maximum possible range of the radio. Accordingly, you represent these views via the prior shown in the left hand panel of figure 1 (fairly flat prior in a circle from the tower, perhaps going down to zero gradually at 25km). The search area represented by this prior is currently far too wide for a rescue crew to reach the flagging ship in time!

In an attempt to improve the odds, you radio to the ship, and ask that they switch on their emergency transmitter. After radioing a number of times, you receive a weak signal, which you feed into the computer, resulting in a posterior probability density for the ship's location shown in the central panel of figure 1. (This panel shows a broadly peaked density along a line from the ship to the station, near a particular location 15km away).

The trouble is, the search area inferred from the aforementioned posterior is still wide. Luckily for the crew however, another nearby radio station has also picked up the signal, and they share this information with you. Finally, feeding this information into the computer, you obtain a final estimation of the ship's location, shown in the right hand panel of figure 1.

Since there is only a small area of high density, you direct the rescue helicopter to search this area, and they find the find the captain and crew in time.

4.3.2 Bayesian statistics: updating our pre-analysis uncertainty

In Bayesian statistics the posterior distribution summarises the combination of our pre-study, and post-analysis knowledge about a given situation, and is used as the starting point for any further analysis or descriptions of our results. In order to calculate it, we need to choose a probability model, which in turn defines uncertain parameters, over which we place priors. The model also provides us with a *likelihood* of the data we obtained. The priors and likelihoods are then combined in a certain way - using Bayes'

veterans [2].

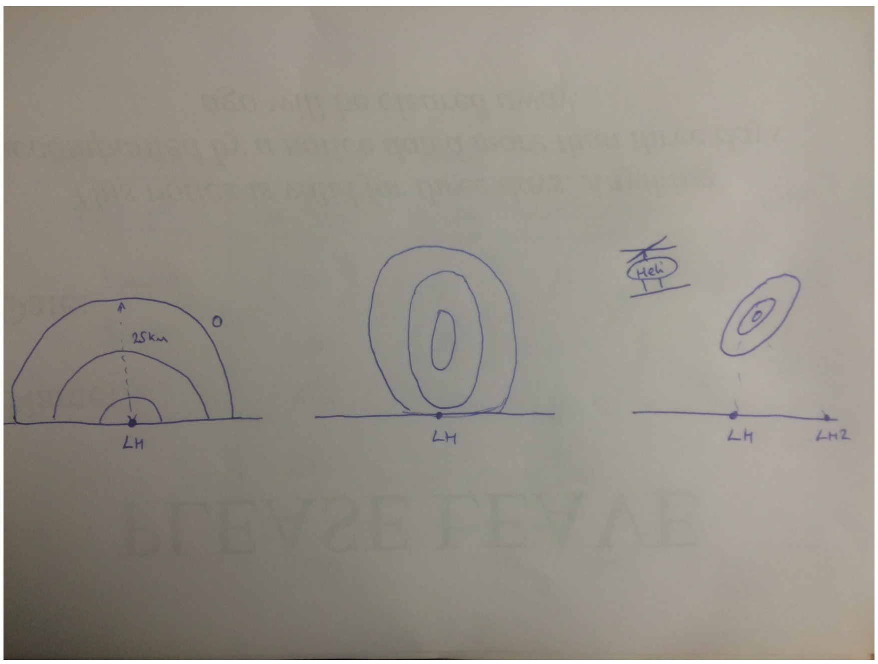


Figure 4.3: Three plots. Left hand plot is a contour plot of probability radiated symmetrically in a semi-circle away from the lighthouse, with the density declining to zero at 25km. The middle plot shows a contour plot of probability, with a higher density towards the centre of the diagram (here the densities are still relatively smooth, indicating high uncertainty). The final plot shows a definite peak in intensity around a particular point about 10km away from the coast, just right of centre. The different contours will be increasing shades of a particular colour.

rule - to yield the posterior distribution.

$$prior + data \rightarrow posterior \tag{4.5}$$

In the Bayesian lighthouse example, we started off with a fairly wide prior, since we were quite uncertain of the boat's location. We then fed the data from the ships' emergency transmitter, along with the prior, into the computer - which uses Bayes' rule - to provide an updated estimate of the ship's location. We actually went through this process twice, to emphasise the fact that Bayes' rule can be used iteratively to update knowledge about an uncertain situation.

Statistical inference is useful whenever there is uncertainty regarding a parameter of interest. Bayesians use the posterior distribution, and various summaries of it, in order to describe the degree of uncertainty regarding a parameter. Before we delve too deep though, it is useful to take a step back, and ask the somewhat philosophical question, 'Do parameters actually exist?'

4.3.3 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. Whereas Frequentist statisticians view the unseen part of the system - the parameters of the probability model - as being fixed, whereas the known parts of the system - the data - as varying. Whether you agree with one of these views more than the other mainly comes down to how you want to interpret the parameters of a given statistical model.

The Bayesian perspective on parameters can be viewed as having a duality of meaning. Either we view the parameters as truly *varying*, or we view our knowledge about the parameters as imperfect. The fact that we will get different estimates of parameters from different studies can be taken to reflect either of these two views. Either we view the parameters of interest as varying - taking on different values in each of the samples we pick (see the bottom panel of figure 4.4). Alternatively, we can view our uncertainty over a parameter's value as the reason we will estimate slightly different values in different samples. This uncertainty is thought of as decreasing as we collect more data (see the middle panel of figure 4.4). Bayesians are more at ease with using parameters as a means to an end - taking them not as real immutable constants, but as tools from which to make inferences about a given situation.

The Frequentist perspective is less flexible, and assumes that these parameters are constant; alternatively representing the average of a long run - typically 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 proportion of the electorate that voted for the Democrat party in the last election, or the probability that an individual taken at random from the UK population will have dyslexia. In both these examples, it is reasonable to assume that there is a *true*, or fixed *population* value of the parameter of question. Whilst the Frequentist view might be in some ways reasonable here, the Bayesian view easily extends here to encompass these two situations by assuming that we are uncertain about the value of these fixed parameters before we measure them, and using a probability distribution to represent this lack of perfect knowledge.

However, there are circumstances when the Frequentist view runs into trouble. When we are estimating parameters of an arbitrarily complicated distribution, we normally do not view them as actually existing. Unless you view the universe as being built from mathematical building blocks[4], 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 top panel of figure 4.4) also runs into difficulty when we think about one-off events, such as the 2016 US Presidential Election result. Any parameter we wish to estimate about such an event cannot easily be existentially justified on these grounds, since elections cannot be rerun.

4.3.4 Failings of the Frequentist confidence interval

A mainstay of the Frequentist estimation procedure is the *confidence interval*. In empirical research we often see these intervals as stated for a given parameter (where for now we assume that the parameters are unknown and fixed). For example,

[4]See [20] for an interesting argument for this hypothesis

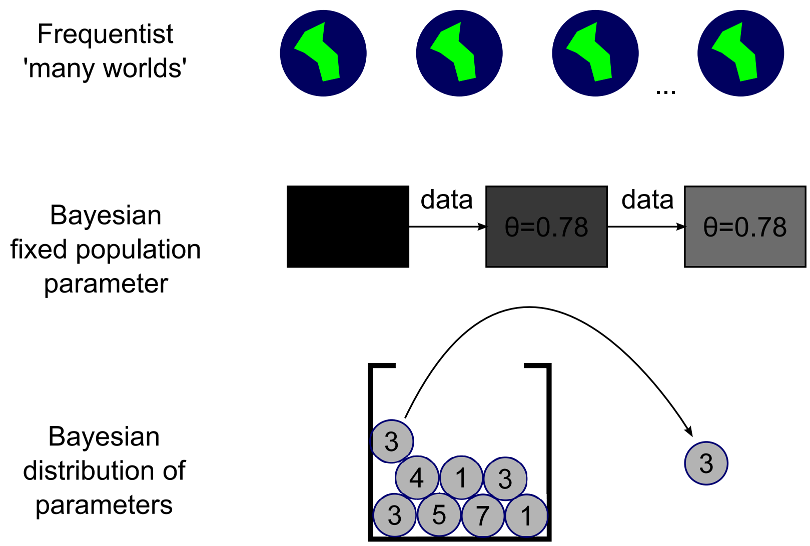


Figure 4.4: The Frequentist and Bayesian perspectives on parameters.

'From our research, we concluded that the percentage of penguins with red tails, RT, has a 95% confidence interval of $1\% \leq RT \leq 5\%$.'

This is often incorrectly taken as having an implicit meaning, 'We are 95% sure that the true percentage of penguins with red tails lies in the range of 1% to 5%.' However, what it actually captures is not uncertainty about the parameter in question, but about the interval we calculate.

In the Frequentist paradigm we imagine that we are taking repeated samples from a population of interest, and for each of the samples, we estimate a confidence interval (see figure 4.5). A 95% confidence interval means that across all of the intervals we calculate, the true value of the parameter will lie in this range 95% of the time.

However, what is important to note here is that in reality, we only draw one sample from the population, and we have no way of knowing whether the confidence interval we calculate contains the true parameter value. This means that although 95% of the time the confidence intervals we calculate will contain the true value of the parameter, and 5% of confidence intervals will be nonsense!

In general, a confidence interval represents the uncertainty about the interval we obtained, rather than a statement of probability about the parameter

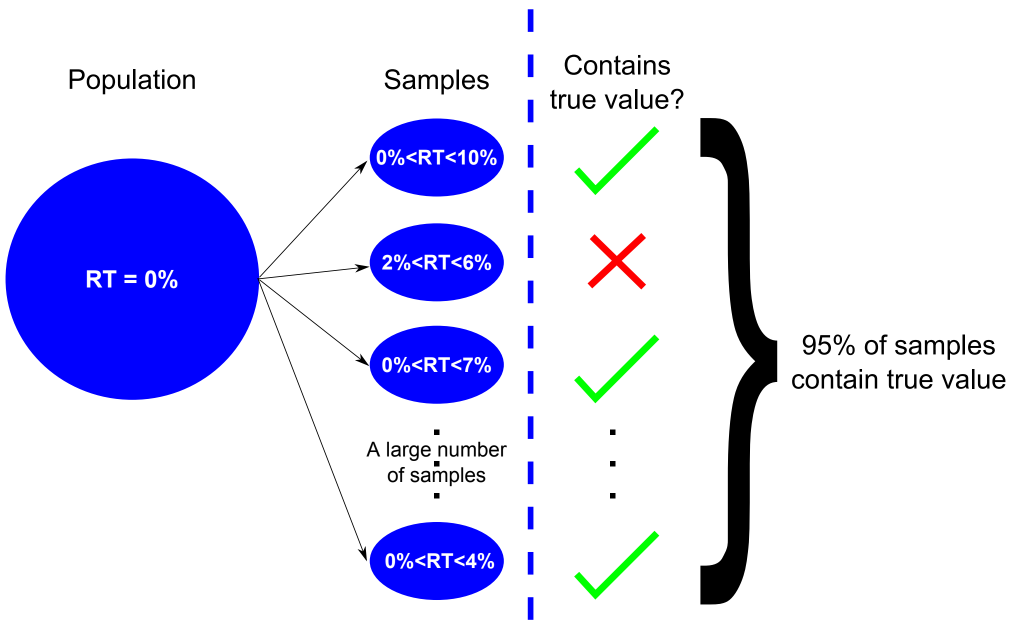


Figure 4.5: The classical confidence interval. In each sample, we can calculate a 95% confidence interval. Across repeated samples from a given population distribution, the classical confidence interval will contain the true parameter value 95% of time.

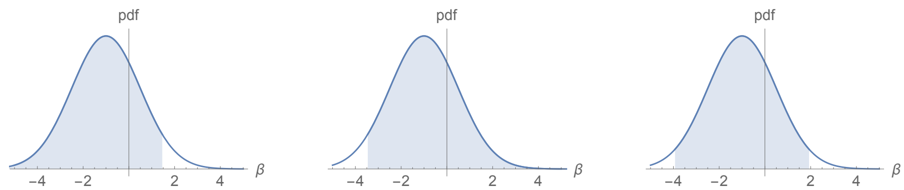


Figure 4.6: Three examples of a 95% credible interval for the regression parameter β of the example described in section 4.3.

of interest. The uncertainty is quantified in terms of all the samples we *could* have taken, not only the one we have in our hands.

4.3.5 Credible intervals

Credible intervals, in contrast to confidence intervals, describe our uncertainty in the location of the parameter values and thus can be interpreted as a probabilistic statement about the parameters. They are a Bayesian concept, that is calculated from the posterior density.

In particular, a 95% credible region satisfies the condition that 95% of the posterior density's area lies in this parameter range. The statement below,

'From our research, we concluded that the percentage of penguins with red tails, RT, has a 95% credible interval of $0\% \leq RT \leq 4\%$.'

can be interpreted as, 'From our research, we conclude that there is a 95% probability that the percentage of penguins with red tails, lies in the range $0\% \leq RT \leq 4\%$'.

In general an arbitrary credible interval of $X\%$ can be constructed from the posterior density, by finding a region whose area is equal to $\frac{X}{100}$.

In contrast to the classical confidence interval, a credible interval is more straightforward to understand. It is a probability statement of confidence in the location of a parameter. Also, in contrast to the classical confidence intervals, the uncertainty here refers to our inherent uncertainty in the value of the parameter, rather than counter-factual samples.

There are usually an infinite number of regions which satisfy this condition, as figure 4.6 indicates for the regression example used in section 4.3. All three of the examples shown in figure 4.6 satisfy the condition, that given our choice of model and prior, we conclude that there is a 95% probability that the parameter lies in this range.

In order to reduce the number of credible intervals down to one, there are 'industry standards' that are followed in most applied research. We introduce two of the most frequently used metrics now.

Treasure hunting: The central posterior and highest density intervals

Imagine you (as a pirate) were told by a fortune-teller that treasure of $1000 is buried somewhere along the seashore of an imagined island. Further, imagine that the mystic has gone to the trouble of using their past experience, and intuition, to arrive at a posterior density for the location of the treasure, along the x-axis seashore, that is shown in figure 4.7. The cost to hire a digger to dig up 1km of coastline is $100.

Suppose you want to find the gold with 95% certainty, and maximise your profit in doing so[5]. In order to reach this level of confidence in plundering the gold, you have the choice of the two 95% credible intervals shown in figure 4.7: the left-hand *central posterior interval*, and the right-hand *highest density interval*.

Both of these intervals have the same area, so we are equally likely to find the gold in either. So which one would be best to choose?

The central posterior interval spans a range of $0.25km - 9.75km$ along the beach. This would entail a cost of $9.5 \times \$100 = \$950$.

The highest density interval by contrast spans two non-contiguous regions, given by $0km - 2.5km$ and $7.5km - 10km$. Each has a cost of $2.5 \times \$100 = \$250$, meaning a total cost of $500. We pick the right-hand strategy, and cross our fingers!

Intuitively, we should avoid the left-hand central posterior interval since it this involves digging up more coastline which has a very low probability of containing the gold. The right-hand highest density interval avoids this misspent effort by directing our search efforts towards the areas most likely to contain the treasure.

If it was costly to drive a digger a given distance, without digging, then we might change our minds, and favour the contiguous region of the *central*

[5]This is really a Bayesian Decision Theory question, where you are choosing an *action*,

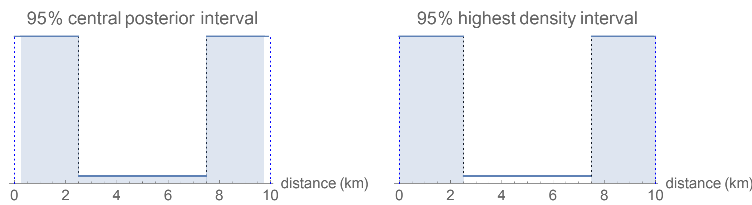


Figure 4.7: The posterior probability for treasure being found along the seashore (represented by a linear x-axis).

posterior interval over the *highest density interval*. However, in most practical (non-pirate) situations, the most sensible thing to do is to report the highest density interval.

To work out the upper and lower bounds of an X% central posterior interval, we find the $\frac{X}{2}$% and $(100 - \frac{X}{2})$% quantiles of the posterior distribution. This will result in an interval that is centred on the median parameter value.

To find the X% highest posterior interval, we find the set of values which contains this percentage of the posterior density area, with the property that the probability density in this set is never lower than outside.

For a unimodal, symmetric distribution the central posterior density and highest density intervals will be the same. However, for more complex distributions, this is no longer the case (compare the left and right hand panels of figure 4.7).

4.3.6 Reconciling the difference between confidence and credible intervals

It is easy to jump on the bandwagon, and dismiss classical confidence intervals as misleading; favouring the Bayesian alternative definitively. How-

which corresponds to an interval; specifying a constant cost function across the domain.

ever, in doing so we are somewhat guilty of zealotry[6]. The two concepts really just represent different measures of uncertainty. As we explained in section 2.5, Frequentists view data sets as one of an infinity of exactly repeated experiments, and hence design an interval which contains the true value X% of the time across all these repetitions. The Frequentist confidence interval states uncertainty in terms of the interval itself. By contrast, Bayesians view the data as fixed, and the parameter as coming from an overarching distribution. They correspondingly calculate an interval where X% of the parameter's estimated probability mass lies.

The problem with the classical confidence interval is more that it is often interpreted *incorrectly*, as a *credible* interval. It is not a problem with the concept itself. It just depends on your personal preference, and situation, as to which you find more useful.

The following (slightly silly) example hopefully makes this difference of viewpoint clearer.

The interval ENIGMA

Suppose that at the outbreak of war, you are employed as a code breaker in hut 8 at Bletchley Park. By monitoring enemy communications we are able to identify the *source* of the message, although not its *contents*. The source of the message is either submarine, boat, tank or aircraft. The content of the messages is details of the next domestic target of the enemy, and can either be dams, ports, towns or airfields.

Fortunately, previous code-breakers have managed to decode a significant proportion of messages, and have tallied up the proportions of communications from each source, which resulted in a particular attack destination (see figure 4.1). We also know from experience that the proportion of attacks on each destination are roughly similar.

Our job is to predict the next attack destination *given* that we have received the mode of communication used. Since there is uncertainty regarding the attack destination, we shall be making confidence intervals which consist

[6] Fanaticism.

	Attack destination			
Communication method	Dam	Port	Town	Airfield
Submarine	73%	50%	50%	13%
Boat	9%	25%	25%	16%
Tank	0%	25%	25%	66%
Aircraft	18%	0%	0%	5%
Total	100%	100%	100%	100%

Table 4.1: Historical communication frequencies resulting in an attack on a given location.

of groups of these entities. We are told to use the most narrow[7] intervals of width greater than or equal to 75% in all cases.

From this historical evidence we first put the data into a 'statistics-machine', turning the knob that says 'classical confidence intervals'. The result are the confidence intervals shown in table 4.2. In words, this is because the sum of the interval values contained in each column exceeds the threshold. So, for every attack destination, our intervals ensure that the true attack destination lies within the specified sets at least 75% of the time.

	Attack destination				
Communications method	Dam	Port	Town	Airfield	**Credibility**
Submarine	[73%	50%	50%]	13%	**93%**
Boat	[9%	25%	25%	16%]	**100%**
Tank	0%	25%	25%	[66%]	**57%**
Aircraft	18%	0%	0%	5%	**0%**
Coverage	**82%**	**75%**	**75%**	**82%**	

Table 4.2: Classical confidence intervals calculated from data shown in table 4.1. Confidence intervals greater than or equal to 75% are indicated in red, surrounded by parentheses.

We next turn the dial to 'Bayesian credible intervals', and obtain the results shown in table 4.3 (see section 4.11.1 for a full explanation). In this case, we are implicitly assuming that the choice of destination is a random variable,

[7]We suppose there is a cost to readying a destination against attack.

and that the enemy chooses amongst them uniformly[8]. In this case, since the sum of interval values in each row exceeds 75%, we have credible intervals. With these intervals, for each mode of communication, we will ensure that the true attack destination is contained within these destinations at least 75% of the time.

	Attack destination				
Communication method	Dam	Port	Town	Airfield	**Credibility**
Submarine	[73%	50%	50%]	13%	**93**%
Boat	[9%	25%	25%]	16%	**79**%
Tank	0%	25%	[25%	66%]	**78**%
Aircraft	[18%]	0%	0%	5%	**78%**
Coverage	**100%**	**75%**	**100%**	**66%**	

Table 4.3: Bayesian credible intervals calculated from data shown in table 4.1. Credible intervals greater than or equal to 75% are indicated in red, surrounded by parentheses. Note: 'credibility' is calculated by dividing the sum of interval values in each row by the row's total (see section 4.11.1 for a full explanation.)

The difference between these two measures is subtle. In fact, as is often the case, the intervals are actually similar, and overlap considerably. But which should we choose? Using the classical confidence intervals, we are assured that whatever destination the enemy chooses, our interval will contain the true attack destination at least 75% of the time. A Bayesian would criticise the confidence interval for the case of the *Aircraft* communication method, since this is the empty interval! This clearly is nonsensical, since we know that one of the locations is about to be attacked. This error could be particularly costly if attacks coordinated via Aircraft are particularly costly. A Frequentist would argue that since, *at most*, Aircraft communications happen 18% of the time (for dams), this isn't something to worry about.

A Bayesian would also criticise a classical confidence interval, since for a given communication mode, what is the use in worrying about all the other communication modes? We aren't uncertain about the communication mode!

A Frequentist would argue that for attacks on airfields, the Bayesian confidence intervals only correctly predict this as the attack destination 66%

[8]Which isn't unreasonable given that we know from experience that the enemy attacks

of the time. Again, if these types of attack are particularly costly, then this

---

each of these locations in similar proportions.

interval might not be ideal[9]. A Bayesian would argue that, assuming a uniform prior, this type of attack only happens 25% of the time, and so is not something to worry about. Further, for every mode of communication, our credible intervals are guaranteed to never be nonsense, in contrast to the classical confidence interval.

4.4 Point parameter estimates

Whilst it is beneficial to estimate the entire posterior distribution for a particular parameter, we are often required to give point estimates. This is sometimes for direct comparison with Frequentist approaches, but more often it is to allow people not versed in Bayesian statistics to make deci-

[9] If we were to assign costs to each of these attack events, then we could work out optimal

intervals, although this is the realm of Bayesian Decision Theory; not covered in this text.

sions[10]. We want to reiterate, that we are *not* advocating these estimators[11]

[10]The correct process in order to make decisions using posterior distributions is through decision theory, not covered in this text. A good review of this field is provided in [15].

as a replacement to providing the entire posterior distribution; only that they are useful summaries which supplement an analysis.

There are three main approaches here, although we argue that only two out of the three should be used. These estimators can be summarised as:

- The posterior mean.
- The posterior median.
- The maximum a posteriori estimator (MAP).

The posterior mean is simply defined for a univariate-continuous example as:

$$E[\theta|data] = \int_{\Theta} \theta p(\theta|data) \mathrm{d}\theta \tag{4.6}$$

For the discrete case, we replace the integrals above with summations (see section 3.3.3).

The posterior median is the point of a posterior distribution to which 50% of its area (the probability), lies to its left. The MAP estimator is simply the highest point in the posterior (see figure 4.8).

Whilst each of these three estimators can be optimal for a given loss func-

[11] Rules that take an input, and produce estimates.

tion[12], we believe that there is a clear hierarchy of these estimators. At the top of the hierarchy - denoting our preferred estimator - we have the posterior mean. This is our favourite for two reasons: firstly, it normally yields *sensible* estimates, which are fairly representative of the central position of the posterior distribution; secondly, and more mathematically, this estimator makes sense from a measure-theoretic perspective, since the quantity of interest is a probability not a density. The density depends on the particular parameterisation that we choose, the probability should not. Don't worry about this last point too much, we just wanted to mention it for completeness. Next down the hierarchy we have the posterior median. This is usually pretty close to the mean (see figure 4.8), and is often indicative of our posterior distribution. It is sometimes preferable to use a median if the mean is heavily-skewed by a slowly-changing density, although the choice between the two estimators depends on circumstance.

At the bottom of the hierarchy, we have the MAP estimator. Proponents argue that the simplicity of this estimator is a benefit. Since it maximises the posterior, across values of the parameter, we do not actually need to calculate the tricky denominator (see chapter 7), as this is independent of the parameter. This means that it is often feasible to find the MAP estimator exactly, even if this is not possible for the other two measures. However, its simplicity is misleading. Importantly, the mode of a distribution often lies away from the majority of the probability distribution's area, and is hence not a particularly indicative central measure of the posterior. Secondly, and mathematically, this estimator does not make sense because it relies on the density, which depends on the particular parameterisation in question.

The bottom line is that one should not use the MAP estimator unless you have a *very* good reason for doing so.

4.5 Prediction using predictive distributions

Parameters of a statistical distribution are typically only of interest insomuch as they influence *real variables*, for which we collected data in the first place. Be it income levels, disease cases, or electoral votes; the data is what drove us to conduct a statistical analysis. As such, it is frequently useful to compare models (which may have very different statistical formulations) using this common currency, of *data*.

[12] A decision theoretic concept, again.

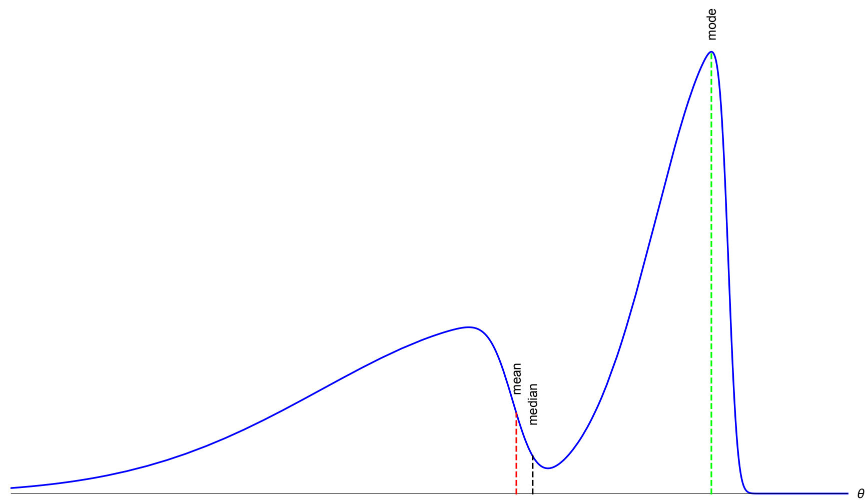


Figure 4.8: The mean, median and mode for a skewed, multi-modal distribution.

Before we have run our model, we only have our prior views of the likely values of parameters. It is often informative to convert from the currency of *parameters*, to that of *data*, in order to evaluate the tangible implications of the chosen priors[13].

Also, after we have fed our priors and likelihoods into the Bayesian formula, we are outputted with the posterior distribution for our parameters of interest. Fortunately, both of these are simple, due to the manipulable nature of probabilities.

4.5.1 Example: number of Republican voters within a sample

You find yourself working for a polling organisation, ahead of the next US Presidential election. Your job is to try to predict, out of a sample of 100 people, what will be the number voting for the Republican party. Based on previous work, you expect that the proportion of Republican voters in a sample, θ, can be represented by the prior distribution, $p(\theta)$, shown in the top-left of figure 4.9. To evaluate the implications of this prior, we would like to know what this means in terms of number, x, of people out of our sample

[13]See chapter 6 for a much more in-detail discussion.

of 100, who will vote for the Republicans; in other words, the *prior predictive distribution*. Fortunately, we can obtain this by manipulating probabilities, although we need to specify a likelihood function, $p(x|theta)$; in this case we pick a binomial distribution[14]. The prior predictive distribution here is essentially the marginal distribution of x, which we know from section

[14]The relevance and nice fit of this distribution is explained in full in chapter 5, so do not worry if you don't follow its use here.

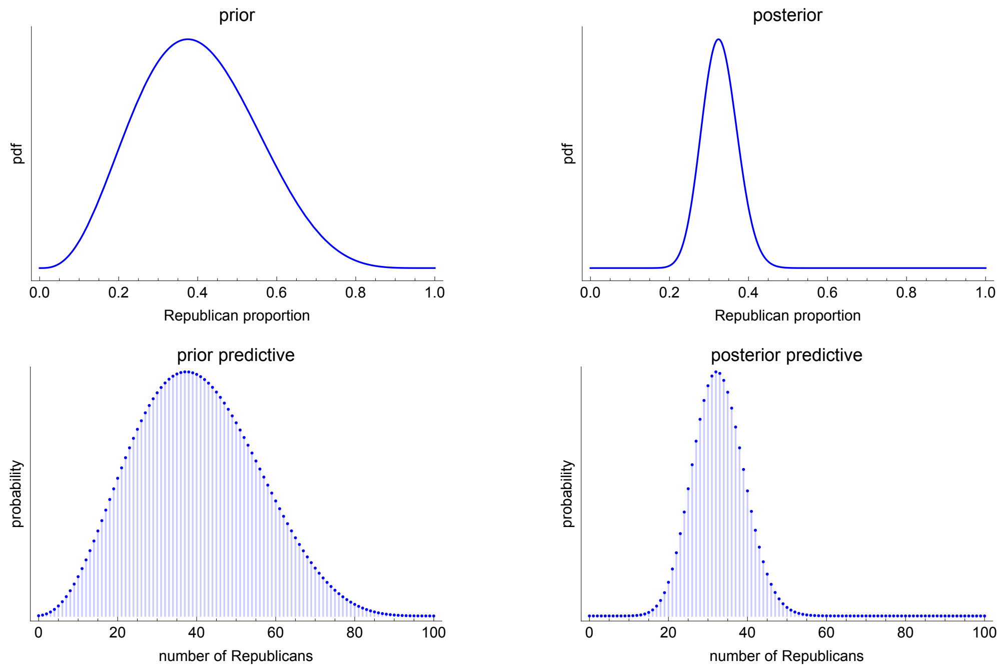


Figure 4.9: Top-left: the prior proportion of people voting for the Republican party in a sample of 100, resulting in the prior predictive distribution shown in the bottom-left. Top-right: the posterior proportion of people voting Republican, resulting the bottom-right posterior predictive distribution.

3.3.5 can be obtained by integrating[15] out the dependence of the parameter θ from the joint distribution $p(x,\theta)$:

$$\begin{aligned} p(x) &= \int_0^1 p(x,\theta)\mathrm{d}\theta \\ &= \int_0^1 p(x|\theta)p(\theta)\mathrm{d}\theta \end{aligned} \tag{4.7}$$

where to get to the second line from the first, we have used the conditional law of probability (3.24), to decompose the joint distribution into a

[15]Since the parameter is continuous here.

conditional and prior. Notice that the prior predictive distribution is essentially a sum of the probability of x conditional on θ, weighted by our prior probability to that proportion, θ, voting Republican. this case, this results in the intuitive result, where the prior density and the prior predictive distribution[16] exactly line up (albeit on different scales).

After our first sample of 100 people from that particular location, we find that 32 of them would vote Republican. We use this data to calculate our likelihood, then combine this with our prior via Bayes' rule; obtaining the posterior density shown in the top-right of figure 4.9. Our job is now to estimate the number of people who will vote republican in a new sample, of the same size. Ideally, we would like to have a probability distribution to describe all the possible outcomes. This distribution is known as the *posterior predictive distribution*, because it is that which we would predict *after* obtaining our data sample. It is written as $p(x'|x)$ where x' represents the number of people voting Republican in the new sample, and x is the number voting Republicans in the previous sample ($x = 32$). We can use a method exactly analogous to that which we used previously, by integrating out θ dependence of the joint distribution of x' and θ, although now we have to condition on x:

$$\begin{aligned} p(x'|x) &= \int_0^1 p(x', \theta|x) \mathrm{d}\theta \\ &= \int_0^1 p(x'|\theta, x) p(\theta|x) \mathrm{d}\theta \\ &= \int_0^1 p(x'|\theta) p(\theta|x) \mathrm{d}\theta \end{aligned} \tag{4.8}$$

Note that to get from the second line from the first, we have used the same conditional probability law (3.24), as we used for the prior predictive case. The only difference here is that we are additionally conditioning on x. To get from the second to the third line we have used a typical assumption, which is that once θ is known, the likelihood for our new sample does not depend on the previous data x. Another way to think about this is that all

[16]This is predictably called a Beta-binomial distribution, since it is made from combining

the information in x has been used to estimate θ; meaning that it does not confer any further information which is helpful for predicting x'. Again, we note in figure 4.9 how the posterior predictive distribution lines up exactly with the posterior distribution of θ. This makes intuitive sense, since if we predict the most likely *proportion* to vote Republicans is 0.32, we should expect that this will translate into the most likely *number* of $0.32 \times 100 = 32$, in a sample of 100.

4.5.2 Example: interest rate hedging

Suppose that you work as an analyst in an investment bank, focussing on predicting the actions of the central bank, so that your investments can be sufficiently hedged. The return of a certain investment, x, is probabilistically dictated by the rate of interest, and is also discrete; following a poisson distribution with rate parameter given by the chosen rate. Imagine that the current rate of interest is 1%, and the bank sets rates at 0.5% intervals; yielding a discrete distribution of possible values: $i \in$

a Beta prior, with a Binomial likelihood. Don't worry though, we will discuss this in detail in chapter 9.

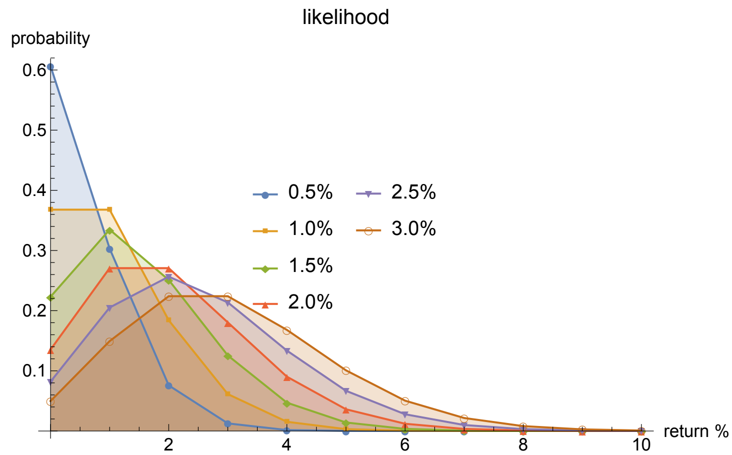


Figure 4.10: The likelihood of different rates of return across different central bank interest rates.

[0.5%, 1.0%, 1.5%, 2.0%, 2.5%, 3.0%]; resulting in the likelihood[17] shown in figure 4.10.

The pre-data-analysis probabilities that your in-house economist gives to each different central bank rate are shown in figure 4.11. After combining these expert views with a probabilistic model, which looks at historical rate decisions, this results in the posterior distribution of probabilities shown in figure 4.11; illustrating considerable discordance between the data's view,

[17]Do not worry about the use of likelihoods here, as we shall explain this concept fully in chapter 5.

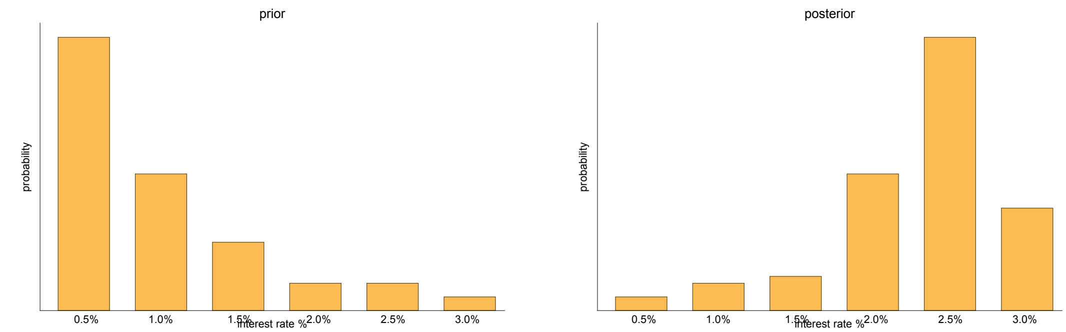


Figure 4.11: The prior and posterior probabilities of different central ban interest rates.

and those of your economist[18].

From the prior distribution, $p(i)$, we would like to estimate the probability of particular rates of investment return. Mathematically, what we want is the marginal distribution of investment return, $p(x)$. Fortunately, we know from section 3.3.5, that we can get a marginal distribution from a joint distribution by summing (for a discrete parameter) across values of the parameter:

$$\begin{aligned} p(x) &= \sum_{i=0.5\%}^{3.0\%} p(x, i) \\ &= \sum_{i=0.5\%}^{3.0\%} p(x|i)p(i) \end{aligned} \tag{4.9}$$

This distribution essentially weights each of the conditional probabilities, $p(x|i)$, by its corresponding prior probability $p(i)$ (see the left-hand side of figure 4.12), then sums them to yield the marginal probability of the data (see the right hand side of figure 4.12).

To calculate the posterior predictive distribution, we proceed in an analogous way to the prior case. What we would like to obtain is $p(x'|x)$, where x' indicates the predicted return on the investment, and x represents the historical data that has been used to produce the posterior. We can again

[18]This might indicate that your economist doesn't know his head from his hands, *or* he

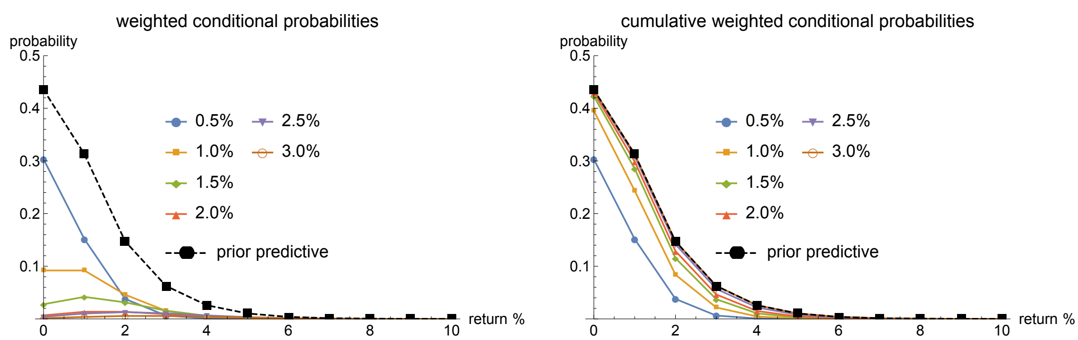


Figure 4.12: Building the prior predictive distribution, as a sum of weighted conditional probabilities. Left: the weighted conditional probabilities (compare with figure 4.10). Right: the cumulative weighted conditional probabilities, which converge on the prior predictive distribution.

achieve this by summing out all i dependence from the joint-conditional distribution $p(x', i|x)$:

$$\begin{aligned} p(x'|x) &= \sum_{i=0.5\%}^{3.0\%} p(x', i|x) \\ &= \sum_{i=0.5\%}^{3.0\%} p(x'|i, x)p(i|x) \\ &= \sum_{i=0.5\%}^{3.0\%} p(x'|i)p(i|x) \end{aligned} \tag{4.10}$$

Note that we have gone from the second line to the third line, because we suppose that, once i is known, the likelihood is independent of past values of x, and thus $p(x'|i, x) = p(x'|i)$. Also, note the similarity between this derivation and that for the prior predictive distribution in 4.9; the only difference in both second lines, is that in the posterior case, everything is conditioned on the historical values of x. We suppose that the likelihood function remains constant, so that like the prior predictive distribution, the posterior predictive distribution is obtained by weighting each of the conditional probability lines $p(x'|i)$ by a probability - in this case, the posterior probability for that value of i.

The resultant prior and posterior distributions are shown in figure 4.13. We notice that the posterior predictive distribution is shifted rightwards, towards higher investment returns, because of the fact that the posterior expected interest rate is higher than for the prior case.

4.6 Model comparison using the posterior

Suppose we have two competing models ($M1$ and $M2$) which we could use to explain a given dataset, and would like a way of evaluating their respective worth. The Bayesian framework can be used to incorporate the choice between two (or more) models, in a straightforward way. Suppose we denote a discrete parameter, $M \in [M1, M2]$, which indexes the choice of model. A way of gauging a model's performance is to calculate the

---

believes there is reason to suggest that this rate decision will be different to those historically.

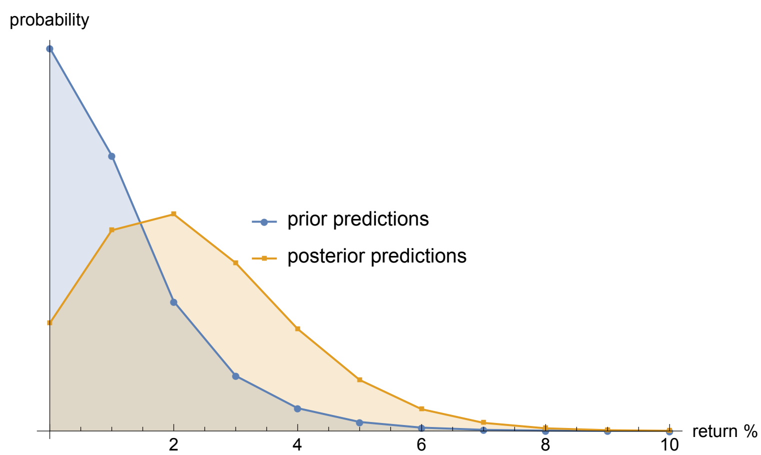


Figure 4.13: The prior and posterior predictive distributions for investment returns.

probability of the model *given* the data obtained, $p(M|data)$. Since we have introduced this new parameter M, we can use Bayes' rule to calculate this quantity:

$$p(M|data) = \frac{p(data|M)p(M)}{p(data)} \tag{4.11}$$

Looking at (4.11) in detail, we examine the following quantities: $p(data|M)$, which represents what we, in the single model set-up, simply call $p(data)$; the $p(M)$ which represents our *a priori* probability of each model being correct; and $p(data)$, which in this case represents the marginal probability of the model when we have summed out our dependence on the model choice:

$$\begin{aligned} p(data) &= \sum_{M=M1}^{M2} p(data, M) \\ &= \sum_{M=M1}^{M2} p(data|M)p(M) \\ &= p(data|M1)p(M1) + p(data|M2)p(M2) \end{aligned} \tag{4.12}$$

We can then calculate the ratio of the probabilities of each model, given the data obtained:

$$\begin{aligned}\frac{p(M1|data)}{p(M2|data)} &= \frac{p(data|M1)p(M1)}{p(data|M2)p(M2)}\\ &= \frac{p(data|M1)}{p(data|M1)} \times \frac{p(M1)}{p(M2)}\\ &= BF \times \frac{p(M1)}{p(M2)}\end{aligned} \tag{4.13}$$

If this ratio is greater than 1, then this test suggests that we should favour model 1, and vice versa if the ratio is less than 1. We have also defined a quantity, *BF*, called the *Bayes Factor*:

$$BF = \frac{p(data|M1)}{p(data|M1)} \tag{4.14}$$

This factor represents in a narrow sense, the strength of support of model 1, over model 2, provided by the data. Note that if our prior probabilities of each model are identical, $p(M1) = p(M2) = \frac{1}{2}$, then:

$$\frac{p(M1|data)}{p(M2|data)} = BF \tag{4.15}$$

And the choice of model is determined solely through the Bayes Factor.

However, there is uncertainty over when is the BF enough to prefer one model over another. Whereas a BF of 100 may be enough to prefer a given model, how about 1.01? Jeffrey's scale (introduced in chapter 10), is an arbitrary, albeit industry-standard, meter to determine when a given model should be preferred over another.

I do importantly want to state here, that this is *not* the correct way to choose between competing models. We will learn a much more nuanced, and reasonable way of choosing between models when we introduce the concept of *Posterior Predictive Checks* in chapter 10. That is not to say that these methods introduced above can't be used as additional evidence for or against a given model, it is just that they should not be used as *sole*, or even *primary*, method.

4.6.1 Example: epidemiologist comparison

Suppose you find yourself in the (bizarre) situation, where you would like to compare two epidemiologists in terms of their ability to predict the underlying proportion of individuals having colds. The first epidemiologist, named *optimist*, gives his estimation of the proportion of individuals with colds, θ, via the (prior) distribution shown in figure 4.14. The second, named *pessimist*, gives a slightly more conservative estimate of the underlying proportion of individuals with colds (also shown in figure 4.14)[19].

The data we have is 10 samples of 100 people, where survey respondents have indicated whether or not they are currently suffering from a cold, $\boldsymbol{x} = \{22, 18, 18, 12, 16, 15, 21, 19, 14, 15\}$.

We can then go through and calculate the $p(data|M)$ for each of the two different priors (see figure 4.14 for a graphical depiction of this), by simply integrating the joint density with respect to θ:

$$\begin{aligned} p(data) &= \int_0^1 p(data, \theta)\mathrm{d}\theta \\ &= \int_0^1 p(data|\theta)p(\theta)\mathrm{d}\theta \end{aligned} \tag{4.16}$$

[19]Here I have actually generated these priors by assuming, for the optimist, $\theta \sim Beta[3, 40]$ and $\theta \sim Beta[4, 15]$ for the optimist and pessimist respectively.

where $p(data|\theta)$ is the likelihood function[20]. Carrying out this integrand for each of the different priors results in the probabilities of data given in the bottom panel of figure 4.14. From this, we note two things: firstly both of the probabilities are very small. This is typical, and is unsurprising when you consider the vast array of possible outcomes that could have occurred. Secondly, we see that the probability of the data for the pessimist is higher than for the optimist. Indeed, we can calculate the Bayes factor via (4.14), and find here $BF \approx 6$ for the pessimist vs the optimist model; which, if we assign equal priors to each model, results in us favouring the pessimistic perspective.

I should add that this use of Bayes factors is atypical; it is not usually used as a means of testing between differing priors. However, since this forms part of a model, we can test its support with the data like any other part.

4.6.2 Example: customer footfall

Suppose that your job at a consultancy is to try to develop a statistical model which describes the footfall of customers into a particular store location at 1pm-2pm, over a span of 2 months. We have collected the data shown in the histogram in the left hand panel of figure 4.15.

Firstly, we fit a poisson likelihood to the data, since we know that the data were collected over a fixed period of time, and we might initially posit that

[20] Which here is taken to be a binomial density.

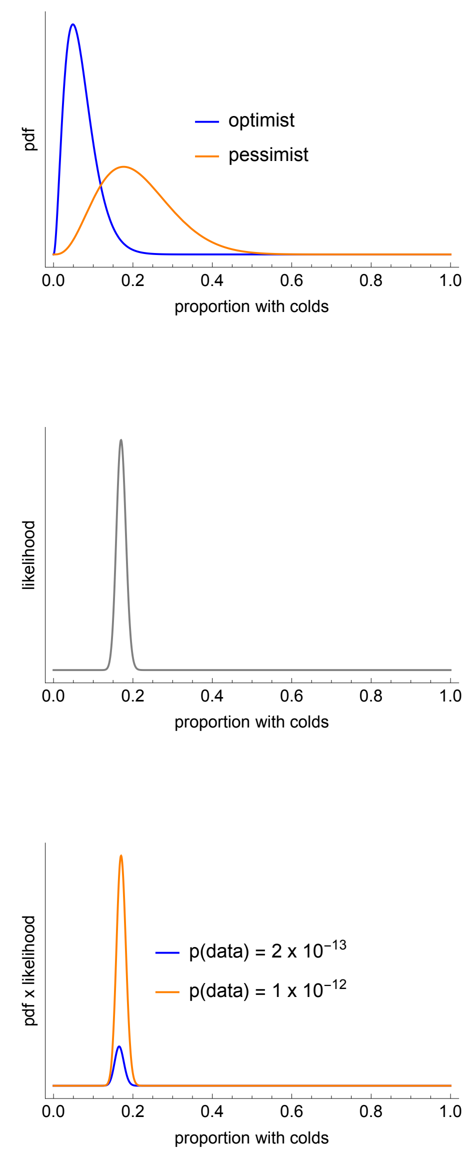
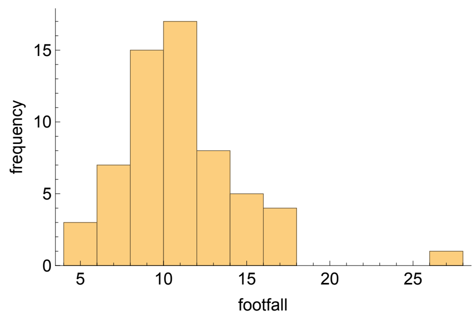


Figure 4.14: Calculating the probability of the data for the two epidemiologists' opinions on colds.

the entry of an individual into the store is independent of others' entry[21]. However, we notice the point to the far right of the histogram. A poisson distribution is not reasonably able to cope with this degree of extremity, since it has a variance which is given by its mean - in this case this is estimated to be 10.5. For a small dataset, it seems hard to believe that a data

[21] See chapter 8 for a more complete examination of this distribution.



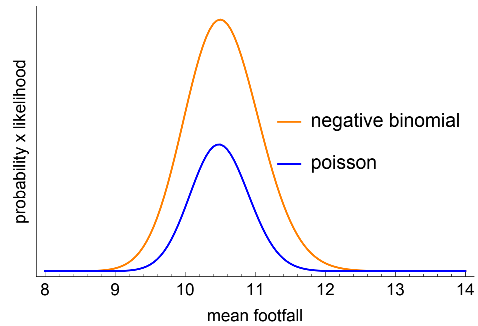


Figure 4.15: Left: the footfall data. Right: the area under the curves represents the probability of the data from each of the two models.

point at 26, could have been generated from such a model[22].

An alternative model, which allows for a variance which exceeds its mean is the negative binomial. This choice of distribution would make most sense if we believed that one consumers' entry into a store was not independent of another's. This would make sense if people tend to shop in groups, and if the shop is outside, with people tending to enter *en masse* when it rains. The expense of this extra freedom, compared to the poisson, is that it is a two-parameter distribution, opposed to a single one.

Again, we can go through and calculate the probability of the data in each case (see figure 4.15), and then take the ratio to find $BF \approx 3$ for the negative binomial vs the poisson, which is unsurprising given the extreme observation. Here though, we might be tempted to *a priori* favour the poisson, due to its relative parsimony; setting a prior for this model that more than accounted for its loss in explanatory power. This might mean that overall we end up using the poisson model, acknowledging its shortcomings in predicting extreme footfall. However this depends on the nature of these extreme events. If they account for a disproportionate part of sales, then it may be worth the extra flexibility of the negative binomial. If by contrast, the extreme footfall is often due to rain, where consumers simply enter to avoid getting wet, and fail to buy, we would perhaps not be so worried about the modelling of these events.

[22]This would be more easily seen by posterior predictive checks.

4.7 Model comparison through posterior predictive checks

The posterior is also the predominant means for testing the worth of a model, and comparing different models. The idea behind this method is that a reasonable model should be able to *simulate* data which is in some correspondence with the *real* data.

The idea behind this methodology is to simulate data, of the same dimension, and in the case of regression, with the same covariates, then compare this to the actual data. To go about simulating the data, we typically go through the following two steps:

1. Sample parameter values from the posterior: $\theta \sim p(\theta|x)$
2. Use these parameters in the likelihood function, then use this distribution to sample new data: $x' \sim p(x'|\theta)$

We shall see a full description of this methodology in chapter 10, so I only mention it briefly as a sign of things to come, and illustrate its basic mechanism through the following example.

4.7.1 Example: stock returns

Suppose we are tasked with modelling the daily-stock returns for a particular company over a one year period (see the leftmost panel of figure 4.16).

We firstly notice the symmetry of the returns, and reason that a normal likelihood may be a reasonable fit. We then use Bayes' rule, with appropriate priors (see chapter 9 for a good guide as to appropriate priors for the parameters of a normal distribution), to calculate posteriors for the mean and variance of this distribution. We then use these posterior distributions, which are fairly narrow, to firstly sample the mean, and variance, then use a normal likelihood to simulate a number of samples of the same size as the original data. For each of the simulated series, we compare the histogram of returns to that of the actual dataset. A typical plot of this form is shown in the middle panel of figure 4.16. We notice that the simulated data is a poor fit to the actual in a number of ways: it under-predicts the number of days with little stock movement; there are an over-abundance of days with

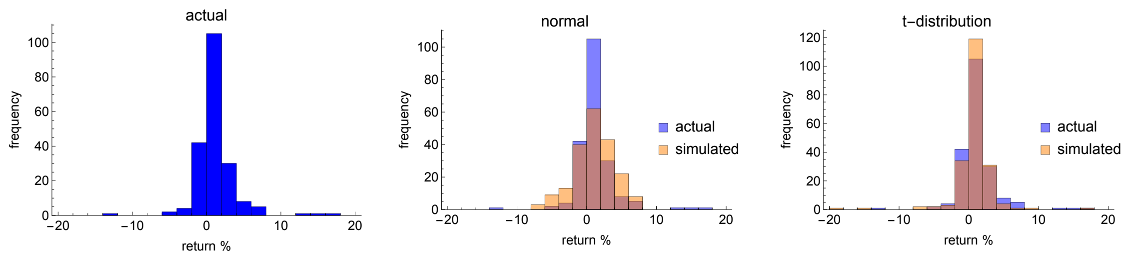


Figure 4.16: Left: the actual data. Middle: actual returns vs normal-simulated returns. Right: actual returns vs t-distribution-simulated returns.

moderate returns; and an under-weighting given to those days with more extreme stock movements. Overall, the dataset is not well represented by our model.

Instead of throwing in the towel, we decide to go for a distribution with fatter tails, which allows a greater degree of flexibility than that of a normal - a student's t distribution. We then go through the rigmarole of setting priors on its three parameters, and finally use Bayes' rule to find the posteriors of these. We then use the posterior distributions to sample these parameters, and then a t-likelihood to simulate the data a number of times. What we now typically see is a better fit (see the rightmost panel of figure 4.16), with a similar abundance of observations near zero, and a similar distribution of extreme daily movements.

Whilst these visual checks may become cumbersome with more complex data series, it is always a good idea to at least start with them, since the eyes can often be less misleading that, for example, p-values. However, we postpone a more complete discussion of posterior predictive checks until chapter 10.

4.8 Chapter summary

In this chapter we have seen how the posterior distribution can be used to produce summaries of parameters. In particular, we have used these distributions to produce Bayesian credible intervals. We then compared and contrasted these with the classically-equivalent confidence interval, and have reasoned that in many cases the Bayesian formulation is more straightforward and intuitive than the former. We then move from the realm of the parameter, to that of the data, in order to produce prior and posterior data probabilities; which can be used to sense-check the implications of statistical models. We finally discussed methods of model comparison, introducing firstly the Bayes Factor method, followed by a short introduction to posterior predictive checks. Although the former method is common in the literature, we advocated the more nuanced approach of posterior predictive checks, albeit postponing more in-depth discussion until chapter 10. Now that we have seen the utility of posterior probability distributions, we need to know how we can obtain them via Bayes' rule. In order to use the latter, we need to understand its constituent parts: the likelihood, prior and denominator. It is these three parts to which I devote the rest of the first part of this book.

4.9 Chapter outcomes

The reader should now be familiar with the following concepts:

1. Expressing uncertainty in a parameter's value through probability distributions.
2. The difference between the classical confidence interval and the Bayesian credible interval.
3. Summary measures of central tendency: posterior mean and median, and MAP. The issues with the latter.
4. How to obtain the prior and posterior predictive distributions from the prior and posterior distributions respectively.
5. Model evaluation through the posterior predictive distribution.

4.10 Problem set

4.10.1 The lesser evil

Suppose that you are a neurosurgeon and have been given the unenviable task of finding the position of a tumour within a patient's brain, and cutting it out. Along two dimensions - height and left-right axis - the tumour's position is known to a high degree of confidence. However, along the remaining axis - front-back - the position is uncertain, and cannot be ascertained without surgery. However, a team of brilliant statisticians has already done most of the job for you, and has generated samples from the posterior for the tumour's location along this axis, and is given by the data contained within the data file "data_Posterior_PS_tumour.csv" .

Suppose that the more brain that is cut, the more the patient is at risk of losing cognitive functions. Additionally, suppose that the damage inflicted varies:

1. Linearly with the distance the surgery starts away from the tumour.

2. Quadratically with the distance the surgery starts away from the tumour.

3. There is no damage if tissue cut is within 1mm of the tumour.

Under each of the three regimes above, find the best position along this axis from which to belong the surgery?

Which of the above loss functions do you think is most appropriate, and why?

Which loss function might you choose to be most robust to *any* situation?

Following from the previous point, which measure of posterior centrality might you choose?

4.10.2 Google word search prediction

Suppose you are chosen, for your knowledge of Bayesian statistics, to work at Google as a search traffic analyst. Based on historical data you have the data shown in table 4.4 for the actual word searched, and the starting string (the first three letters typed in a search). It is your job to help make the search engines faster, by reducing the search-space for the machines to lookup each time a person types.

	Barack Obama	**Baby clothes**	**Bayes**
Bar	50%	30%	30%
Bab	30%	60%	30%
Bay	20%	10%	40%

Table 4.4: The columns give the historic breakdown of the search traffic for three topics: Barack Obama, Baby clothes, and Bayes; by the first three letters of the user's search.

Find the minimum-coverage confidence intervals of topics that exceed 70%.

Find most narrow credible intervals for topics that exceed 70%.

Topic search volumes

Now we suppose that your boss gives you the historic search information shown in table 4.5. Further, you are told that what is most important is correctly suggesting the actual topic as one of the first auto-complete options irrespective of the topic searched.

Do you prefer confidence intervals or credible intervals in this circumstance?

	Barack Obama	**Baby clothes**	**Bayes**
Search volume	60%	30%	10%

Table 4.5: The historic search traffic broken down by topic.

Three-letter search volumes

Alternatively, suppose that you have the historic search volume by the starting characters of the search, shown in table 4.6. You are now told, that the most important thing is directing the user to the correct topic given the keyword they have started to type.

Do you prefer confidence intervals or credible intervals in this circumstance?

	Bar	**Bab**	**Bay**
Search volume	50%	40%	10%

Table 4.6: The historic search traffic broken down by keyword.

4.10.3 Prior and posterior predictive example (with PPCs maybe)

Ben to add later.

4.11 Appendix

4.11.1 The interval ENIGMA - explained in full

Chapter 5

Likelihoods

The world is everything that is the case. **Wittgenstein**

5.1 Chapter Mission statement

At the end of this chapter a reader will know how to go about selecting a likelihood which is appropriate to a given situation. Further the reader will understand the basis behind maximum likelihood estimation.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{5.1}$$

5.2 Chapter goals

The starting point of the right hand side of the Bayesian formula is the likelihood function. This chapter will explain what is meant by a likelihood function, and why it is incorrect to view it as a probability in Bayesian analyses. The choice over which likelihood to use for a given situation is often difficult; especially to those unfamiliar with statistics. This chapter will provide practical guidance on likelihood choice, describing a framework that can be used to select a model in a systematic way. As an important stepping stone to Bayesian estimation, this chapter will also explain how classical maximum likelihood estimation works.

5.3 What is a likelihood?

In all statistical inference, we use an idealised, simplified, model to try to mimic relationships between real variables of interest. This model is then used to test hypotheses about the nature of the relationships between these variables. In Bayesian statistics the evidence for a particular hypothesis is summarised in posterior probability distributions. Bayes' magic rule tells us how we can compute this posterior probability distribution for a given parameter within a model, θ:

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{5.2}$$

The first step to understanding this formula (so that we can ultimately use it!) is to understand what is meant by the numerator term, $p(data|\theta)$, which Bayesians call a *Likelihood*. Firstly, it's important to say that what we really mean by the numerator is:

$$p(data|\theta) = Probability(data|\theta, Model\ Choice) \tag{5.3}$$

(5.3) represents the probability that we would have obtained the 'data', given (this is represented by the | symbol) a particular value of θ and a particular choice of model. In other words, if our statistical model were true, and the value of the model's parameter were θ, (5.3) tells us the probability that we would have obtained our data.

But what does this mean in simple, everyday language? Imagine that we flip a *fair* coin. The most simple statistical model for coin flipping we can pick is to disregard the angle it was thrown at, as well as its height above the surface, along with any other details, and just pick the probability of the coin coming heads to be $\theta = \frac{1}{2}$. Furthermore, if a coin is thrown twice, we might choose to model the situation by assuming that the throwing technique is sufficiently similar between the two throws such that we can model each throw as independently having a probability of $\frac{1}{2}$. It's important to note that it is an assumption to forget about the throwing angle, as well as height of throw for each throw, and this forms part of our model of the situation.

We can use our simple model[1] to calculate the probability that we obtain two heads in a row:

$$\begin{aligned} Pr(HH|\theta, Model) &= Pr(H|\theta, Model) \times Pr(H|\theta, Model) \\ &= \theta \times \theta = \theta^2 \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \end{aligned} \tag{5.4}$$

The last row of (5.3) is obtained by assuming the probability of a head, $\theta = \frac{1}{2}$. If we continue to use this *same* value of θ, we can calculate the corresponding probabilities for all outcomes of throwing the coin twice. The most heads that can show up is 2, and the least being zero (if both flips come up tails). Figure 5.1 displays the probabilities for this model of the situation. The most likely number of heads to occur is 1, since this can occur in two different ways - either the first coin comes up heads, and the second is tails, or vice versa - whereas the other possibilities (all heads, or no heads) can each only occur in one way. However, the important thing to note about figure 5.1 isn't the individual probabilities, it is that it is a *valid*

[1] Albeit in practicality, this is a pretty reasonable representation of the situation for most

purposes.

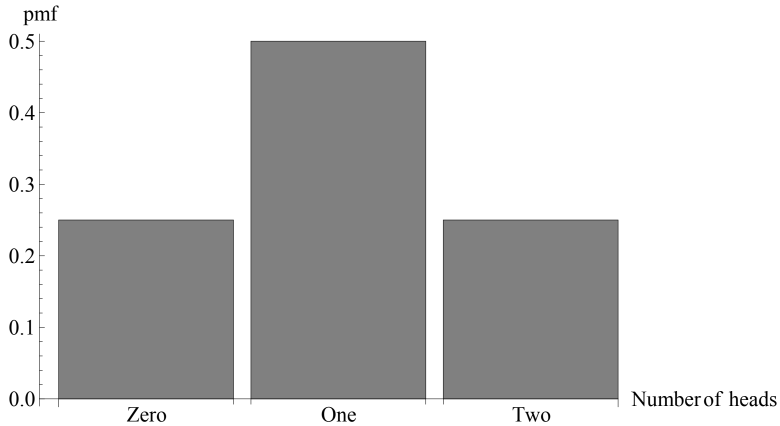


Figure 5.1: The probabilities of all possible numbers of heads for a fair coin.

probability distribution[2], because:

- The individual event probabilities are all non-negative.
- The sum of the individual probabilities is 1.

So it appears when we assume a particular value of θ, and vary the data (in this case the number of heads obtained), the collection of resultant probabilities form a probability distribution. So, why do Bayesians call $p(data|\theta)$ a 'likelihood', and eschew the name 'probability'?

5.4 Why use 'likelihood' rather than 'probability'?

When we hold the parameters of our model fixed, as when we held the probability of an individual throw turning up heads, $\theta = \frac{1}{2}$, we've reasoned that the first term of the numerator of Bayes' rule in (5.3) is a probability. So why don't we just keep calling it that, instead of renaming it a *likelihood*?

The reason is that in Bayesian inference, we *don't* keep the parameters of our model fixed! In Bayesian analysis, it is the *data* that is fixed, and the parameters that vary. This is because a posterior distribution shows the probability a parameter in a model lies in a particular range, assuming that

[2]See section 3.3.1 for a refresher if you are unsure what is meant by a valid probability

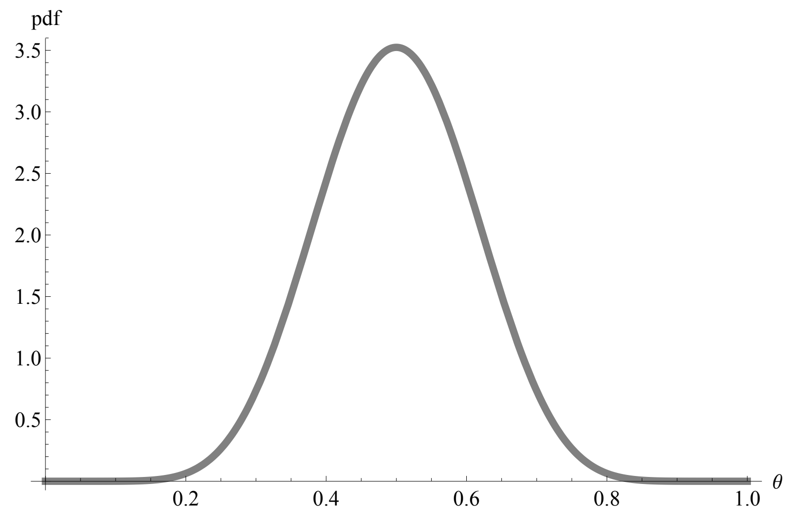


Figure 5.2: An example posterior distribution for the probability of obtaining a heads in a coin toss.

we have obtained our particular data sample. For the case of a coin, where we don't know the probability of a head beforehand, what we hope to get out is a probability distribution of the kind shown in figure 5.2, where the x-axis is the value of θ. In order to get $p(\theta|data)$ however, we must calculate $p(data|\theta)$ from the numerator of Bayes' rule in (5.3) for each *possible* value of θ. If we assume we obtained one head and one tail, then we can calculate the probability of this occurring for a fixed θ:

$$Pr(HT|\theta) + Pr(TH|\theta) = \theta(1-\theta) + \theta(1-\theta) \quad = 2\theta(1-\theta) \tag{5.5}$$

Since we are unsure as to the 'correct' value of θ, we can graph this expression as a function of this parameter, to try to understand which values of the parameter are more or less likely, given our data (see figure 5.3).

On first glances it appears that 5.3 could be a probability distribution, but first looks can be deceiving.

Checking off our necessary components of a probability distribution, we first note that all the values of the distribution in figure 5.3 are non-negative; which is what we require. However, if we calculate the area underneath

distribution.

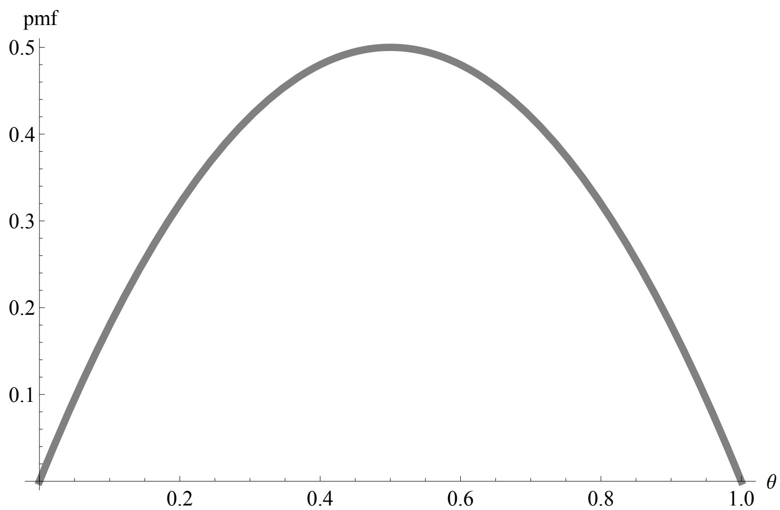


Figure 5.3: The likelihood function for obtaining a single head from two throws. The area under the curve is $\frac{1}{3}$.

the curve in figure 5.3:

$$\mathrm{I} = \int_0^1 2\theta(1-\theta)\mathrm{d}\theta = \frac{1}{3} \neq 1 \tag{5.6}$$

we find that it does not integrate to 1. Thus we have a violation of the second condition for a valid probability distribution. Hence, when we vary θ we find that, $p(data|\theta)$ is not a valid probability distribution! We thus introduce the term 'likelihood' to represent $p(data|\theta)$ when we vary the parameter, θ. Often the following notation is used to emphasise that likelihood is a function of the parameter θ with the data held fixed:

$$\mathcal{L}(\theta|data) = p(data|\theta) \tag{5.7}$$

However, in this book, we will persist with the original notation as this is most typical in the literature, under the implicit assumption that when we vary the parameters in question, the term is not strictly a probability.

To provide further justification for this argument, consider the following (albeit contrived) example. Suppose that, we throw a coin twice, and we are told beforehand that the probability of obtaining a head on a particular throw is one of six discrete values: $\theta \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. We can then use our model to calculate the probability of obtaining a number of heads, X:

$$Pr(X = 0|\theta) = Pr(TT|\theta) = Pr(T|\theta) \times Pr(T|\theta) = (1 - \theta)^2 \tag{5.8}$$

$$Pr(X = 1|\theta) = Pr(HT|\theta) + Pr(TH|\theta) = 2 \times Pr(T|\theta) \times Pr(H|\theta) = 2\theta(1 - \theta) \tag{5.9}$$

$$Pr(X = 2|\theta) = Pr(HH|\theta) = Pr(H|\theta) \times Pr(H|\theta) = \theta^2 \tag{5.10}$$

In (5.8), the probability is simply given by the product of the probabilities of not obtaining a head on the first throw, $(1 - \theta)$, by the probability of not obtaining a head in the second[3], which is also $(1 - \theta)$. The factor of two arises in (5.10) since there are two ways of getting one head: {HT,TH}.

We can represent the corresponding values of likelihood/probability as is shown in table 5.1. In this form we can see the impact of varying the data (moving along each row), and contrast it with the effect of varying θ (moving down each column). Note that if we hold the parameter fixed - regardless of this initial choice of θ - and move along each row summing the entries, we find that the values sum to 1; meaning that this is a valid probability distribution. By contrast, when we hold the number of heads fixed, and vary the parameter θ, moving down each column, summing the entries, we find that the values do not sum to 1. Hence, when we vary θ, we are not dealing with a proper probability distribution, meriting the use of the term 'likelihood'.

In Bayesian inference, we always vary the parameter, and implicitly hold the data fixed. Thus, from a Bayesian perspective it is important to use the term *likelihood* to indicate that we recognise we are not dealing with a probability distribution.

5.5 What are models and why do we need them?

All models are wrong. They are idealised representations of reality resultant from making assumptions, which if reasonable, may emulate some of the

captured in the parameter θ.

	Number of heads			
θ	**0**	**1**	**2**	**Total**
0.0	1.00	0.00	0.00	**1.00**
0.2	0.64	0.32	0.04	**1.00**
0.4	0.36	0.48	0.16	**1.00**
0.6	0.16	0.48	0.36	**1.00**
0.8	0.04	0.32	0.64	**1.00**
1.0	0.00	0.00	1.00	**1.00**
Total	**1.20**	**1.60**	**2.20**	

Table 5.1: The values of likelihood for the case of tossing a coin twice, where the probability of heads is constrained to take on a discrete value: {0.0,0.2,0.4,0.6,0.8,1.0}.

behaviour of a system of interest. Joshua Epstein in an article titled, 'Why model?' emphasises that we perennially build *implicit* mental models for various phenomena [4]. Before we go to bed at night we set our alarms for the next morning on the basis of a model. We imagine an idealised - model - morning when it takes us 15 minutes to wake up as a result of an alarm. We use this model to predict how long it will take us to rise from bed, shower, and get changed into clothes in sufficient time to get to work. Whenever we go to the Doctor, they use an internalised biological model of the human body to advise on the best course of treatment for a particular ailment. Whenever we hear expert opinions on TV about the outcome of an upcoming election, the pundits are using mental models of society to explain the results of current polls, as well as make forecasts. As is the case with all models, some are better than others. Hopefully, the models a Doctor uses to prescribe medicine are subject to less error than the opinions

[3] Since we have assumed a model whereby the results of the first and second throws are

independent, conditional on θ. In other words, all the similarity between the two throws is of pundits seen on TV[4]!

Epstein goes on to emphasise that the question, 'Why model?' really means why should we build an *explicit* - written down - model of phenomena? The point being that *implicit* models are by their very nature, opaque, and not subject to the sort of interrogation that can be obtained by writing the model on paper.

We can also ask more narrowly, what are we hoping to gain by building an *explicit* model of a situation? Epstein suggests the following motivations:

- Prediction
- Explanation
- Guide data collection
- Discover new questions
- Bound outcomes to plausible ranges
- Illuminate uncertainties
- Challenge the robustness of prevailing theory through perturbations
- Reveal the apparently simple (complex) to be complex (simple)

There are of course other reasons to build models, but we believe that this list is a reasonable starting point. However, we should not think of this list as static. Whenever we build a model, whether it is statistical, biological or sociological, we should ask, 'What are we hoping to gain by building this model, and how can I judge its success?'. Only when we have a grasp on the answers to these basic questions should we proceed to model building.

5.6 How to choose an appropriate likelihood?

Bayesians are acutely aware that their models are wrong. At best the abstraction from reality allows us to explain some aspect of real behaviour;

[4]For a great discussion of the performance of TV pundits, read Thinking Fast. Insert reference

at worst they can be very misleading. Before we use a model for prediction, we require that it can explain some reasonable proportion of the system's behaviour for the past and present. With this in mind we introduce the following model selection framework:

1. Write down the real life behaviour/data patterns that the model should be capable of explaining.

2. Write down the assumptions that it is believed are reasonable in order to achieve the above point.

3. Search the literature for models which utilise these assumptions; extracting only the relevant components.

4. Test your model's ability to explain said behaviour/data patterns. If unsuccessful go back to the second step and re-evaluate the appropriateness of your assumptions.

Whilst this methodology is useful for building a statistical model in general, it is more applicable for use with a full Bayesian model, resulting in a posterior distribution. In which case how do we go about specifying a likelihood for a given situation? To answer this we will start with going through a simple example.

5.6.1 A likelihood model for an individual's disease status

Suppose we work for the State as a healthcare analyst, and we want to build a statistical model to explain the prevalence of a certain disease within a sample, which can then be used to make inferences about the population incidence. Also, (unrealistically) let's imagine that we start off with a sample of only one person, for whom we have no prior information. Let the disease status of that individual be denoted by the variable X which takes on the following binary outcome values dependent on the disease status the individual:

$$X = \begin{cases} O & , No\ disease \\ 1 & , Positive\ diagnosis \end{cases} \tag{5.11}$$

The goal of our model is to output a probability that this individual has the disease. We might assume that a fraction θ of the population has the disease, and that this individual has come from that population. For each possible outcome, we can use this simple model to calculate the probability of each outcome:

$$Pr(X = 0|\theta) = (1 - \theta) \tag{5.12}$$

$$Pr(X = 1|\theta) = \theta \tag{5.13}$$

Note the similarity between these probabilities and those of the coin flipping example in the previous section. Often, a given model can be reused in a multitude of different settings.

However, we would like to write down a single rule which yields (5.12) or (5.13) respectively, dependent on whether $X = 0$ or $X = 1$. This can be achieved with the following:

$$Pr(X = \alpha|\theta) = \theta^{\alpha}(1 - \theta)^{1-\alpha} \tag{5.14}$$

Note that in (5.14) that $\alpha \in \{0, 1\}$ refers to the numeric value taken by the variable X. The function (5.14) is known as a *Bernoulli* probability density.

Although this rule for calculating a probability of a particular disease status, α, looks complex, we see that it reduces to (5.12) and (5.13) if the individual is disease -negative/-positive respectively:

$$Pr(X = 0|\theta) = \theta^{0}(1 - \theta)^{1} = (1 - \theta) \tag{5.15}$$

$$Pr(X = 1|\theta) = \theta^{1}(1 - \theta)^{0} = \theta \tag{5.16}$$

When we hold the datum X fixed and vary θ, (5.14) represents a likelihood. However, figure 5.4 shows that for a fixed value of *theta* the sum (here we mean the vertical sum) of the two probability densities is always equal to 1; demonstrating that in this case (5.14) is a valid probability density. Notice also in figure 5.4 that the sum of probability density is defined continuously on $\{0, 1\}$, whereas the sum of likelihoods is discrete.

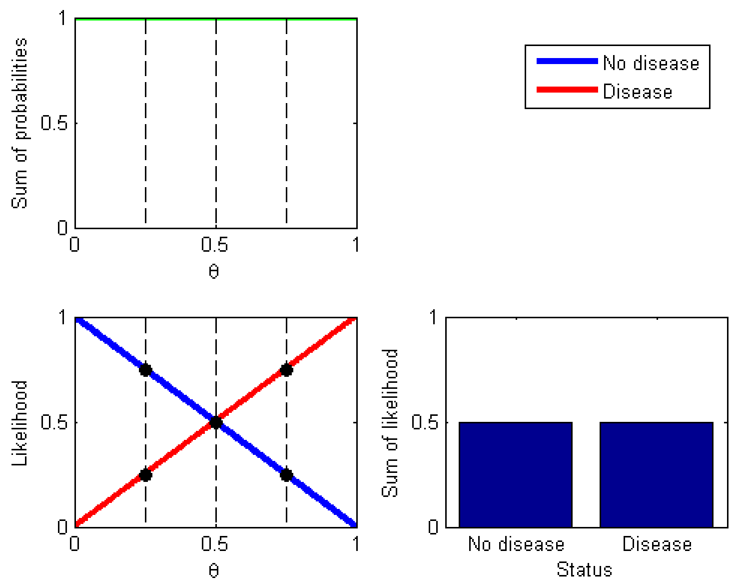


Figure 5.4: The likelihood function as theta varies for the case of the two possible data. The sum of likelihoods is found by the area under each line, whereas the sum of probabilities is a discrete sum.

5.6.2 A likelihood model for disease prevalence of a group

Now we imagine that instead of this solitary individual, we have a group of N individuals. What we would like to do is to calculate the develop a model which will tell us the probability of obtaining Z disease cases within our sample. We would also like to be able to use our model to predict the most likely number of individuals who have the disease in a sample, for a given value of the parameters[5].

In order to write down a model we first need to make some simplifying assumptions. We might assume that one individual's disease status tells us nothing about the probability of another individual in the sample having

[5]We are starting off by assuming that we know the parameters. Later in this chapter we will obtain a point estimate of the parameters using *Maximum likelihood* estimation.

the disease[6]. This would not be a reasonable assumption if the disease were contagious, and if the individuals in the sample came from the same neighbourhood or household. It also would not be a good assumption if (as is often the case with volunteer-dependent studies) the individuals who volunteered for the experiment, self-selected on the basis of some common pre-existing ailment/underlying-factor. If an advert for participants reads, 'Psychological experiment on sleep disorders: participants wanted', we might suspect that there would be an over-presence of insomniacs than is found in the population as a whole. This first assumption is that which in statistical language we call 'independence'. We also suppose that all individuals in our sample come from the same population - the one we are trying to draw conclusions about. If we knew beforehand that some individuals came from different populations, with significantly different prevalence rates, then we might abandon this assumption. Combining these two assumptions we say in statistical language that our data sample is *independent* and *identically-distributed*.

With our two assumptions in hand, we can begin to formulate a model for the probability of obtaining Z disease-positive individuals out of a total of N individuals. We start by considering each person's disease status individually, meaning we can reuse (5.14):

$$Pr(X = \alpha|\theta) = \theta^{\alpha}(1-\theta)^{1-\alpha} \tag{5.17}$$

Note that in (5.17) the $\alpha \in \{0, 1\}$ refers to a particular numeric value taken by the variable X. The assumption of *independence* means that we can get the

[6]Other than, if the disease prevalence were unknown, through our ability to estimate

overall disease prevalence from their individual statuses.

overall probability by multiplying together the individual probabilities[7]. In words, we obtain the probability that the first person has disease status X_1 *and* the second person has status X_2:

$$\begin{aligned} Pr(X_1 = \alpha_1, X_2 = \alpha_2|\theta_1, \theta_2) &= Pr(X_1 = \alpha_1|\theta_1) \times Pr(X_2 = \alpha_2|\theta_2) \\ &= \theta_1^{\alpha_1}(1-\theta_1)^{1-\alpha_1} \times \theta_2^{\alpha_2}(1-\theta_2)^{1-\alpha_2} \end{aligned} \tag{5.18}$$

In (5.18) we have assumed that each individual has a different predisposition to having the disease, denoted by θ_1 and θ_2 respectively.

The second assumption of *identically-distributed* individuals means that we can set $\theta_1 = \theta_2$:

$$\begin{aligned} Pr(X_1 = \alpha_1, X_2 = \alpha_2|\theta) &= \theta^{\alpha_1}(1-\theta)^{1-\alpha_1} \times \theta^{\alpha_2}(1-\theta)^{1-\alpha_2} \\ &= \theta^{\alpha_1+\alpha_2}(1-\theta)^{2-\alpha_1-\alpha_2} \end{aligned} \tag{5.19}$$

In (5.19) we have obtained the second line by using the simple exponent rule: $a^b \times a^c = a^{b+c}$, for the components θ and $(1-\theta)$ respectively.

For our sample of 2 we are now in a position to calculate the probability that we obtain Z cases of the disease. We first realise that we can get from X_1 and X_2 to Z by:

$$Z = X_1 + X_2 \tag{5.20}$$

We can then use (5.19) to generate the respective probabilities.

$$\begin{aligned} &Pr(Z = 0|\theta) = Pr(X_1 = 0, X_2 = 0|\theta) = \theta^{0+0}(1-\theta)^{2-0-0} = (1-\theta)^2 \\ &Pr(Z = 1|\theta) = Pr(X_1 = 1, X_2 = 0|\theta) + Pr(X_1 = 0, X_2 = 1|\theta) = 2\theta(1-\theta) \\ &Pr(Z = 2|\theta) = Pr(X_1 = 1, X_2 = 1|\theta) = \theta^{1+1}(1-\theta)^{2-1-1} = \theta^2 \end{aligned} \tag{5.21}$$

To complete our probability model we want to write out a single rule for calculating the probability of any value taken on by Z. To do this we note that we could rewrite (5.21) as:

[7] See section 3.5 for an explanation of this.

$$\begin{aligned} Pr(Z = 0|\theta) &= \ \theta^0(1-\theta)^2 \\ Pr(Z = 1|\theta) &= 2\theta^1(1-\theta)^1 \\ Pr(Z = 2|\theta) &= \ \theta^2(1-\theta)^0 \end{aligned} \tag{5.22}$$

In (5.22) we notice the common term $\theta^\beta(1-\theta)^{2-\beta}$ in each of the expressions, where $\beta \in \{0, 1, 2\}$ represents the number of disease cases found. Therefore this suggests that we may be able to write down a single rule as something similar to:

$$Pr(Z = \beta|\theta) \sim \theta^\beta(1-\theta)^{2-\beta} \tag{5.23}$$

The only problem with matching (5.23) with the previously obtained result is the factor of 2 on the middle line of (5.22). However, as a complete aside we note that when we expand a quadratic factor we get the following:

$$(x+1)^2 = x^2 + 2x + 1 \tag{5.24}$$

The numbers $\{1, 2, 1\}$ correspond here to the non-b-dependent coefficients of $\{x^2, x^1, x^0\}$ respectively. This sequence of numbers normally appears in early secondary school maths classes, and is either known as the binomial expansion coefficients or simply ${}^nC_r$. The expansion coefficients are normally written in compact form:

$$\binom{2}{\beta} = \frac{2!}{(2-\beta)!\beta!} \tag{5.25}$$

In (5.25) the ! has its usual meaning of factorial, and $\beta \in \{0, 1, 2\}$. We can therefore use this notation to help us to write down a single model for the probability of obtaining Z disease cases out of a total of 2 individuals using our model:

$$Pr(Z = \beta|\theta) = \binom{2}{\beta}\theta^\beta(1-\theta)^{2-\beta} \tag{5.26}$$

This likelihood function is illustrated for the three possible numbers of disease cases in figure 5.5.

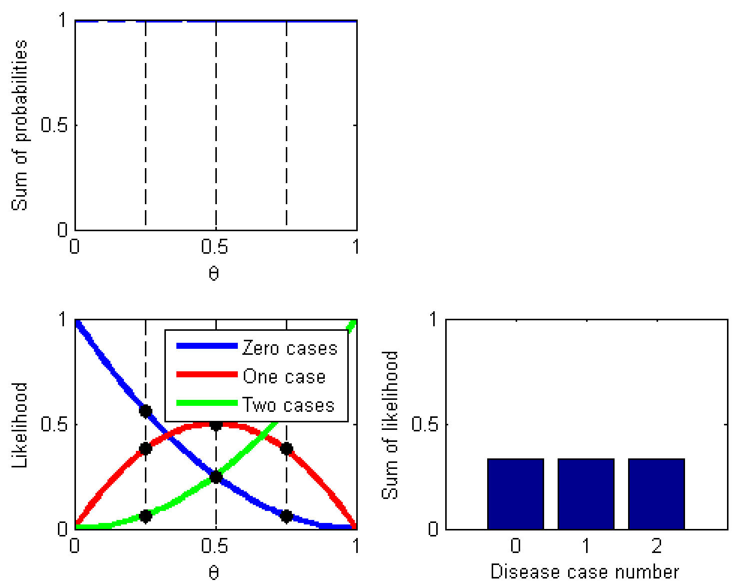


Figure 5.5: The likelihood function as theta varies for a sample of 2 individuals.

We will now extend the analysis to cover the case when we have groups of N individuals. Firstly, consider the case when we have a group size of 3. If we assume that the individuals are identically distributed, then the 4 probabilities are of the form:

$$\begin{aligned}
Pr(Z=0|\theta) &= Pr(X_1=0|\theta)Pr(X_2=0|\theta)Pr(X_3=0|\theta)\\
Pr(Z=1|\theta) &= 3Pr(X_1=1|\theta)Pr(X_2=0|\theta)Pr(X_3=0|\theta)\\
Pr(Z=2|\theta) &= 3Pr(X_1=1|\theta)Pr(X_2=1|\theta)Pr(X_3=0|\theta)\\
Pr(Z=3|\theta) &= Pr(X_1=1|\theta)Pr(X_2=1|\theta)Pr(X_3=1|\theta)
\end{aligned} \tag{5.27}$$

Again, we notice a numeric pattern in terms of the first part of each expression $\{1,3,3,1\}$, which happens to correspond exactly to the coefficients on terms for the expansion of $(x+1)^3$. Hence, we can again rewrite the likelihood using the binomial expansion notation:

$$Pr(Z=\beta|\theta) = \binom{3}{\beta}\theta^{\beta}(1-\theta)^{3-\beta} \tag{5.28}$$

We recognise a pattern in the likelihoods of (5.26) and (5.28) which allows us to deduce that, for a sample size of N, the likelihood is given by:

$$Pr(Z=\beta|\theta) = \binom{N}{\beta}\theta^{\beta}(1-\theta)^{N-\beta} \tag{5.29}$$

(5.29) is known as the *binomial* probability distribution.

If we had data, then we could test whether the assumptions made were appropriate by calculating the model-implied-probability of this outcome. For example, if we had a sample of 100 people of which 10 were disease-positive, and we assumed beforehand that the proportion of the population who have the disease is $\theta = 1\%$, then we could calculate the probability that we would have achieved a number of cases as bad, or worse than this using (5.29):

$$Pr(Z \geq 10|\theta = 0.01) = \sum_{Z=10}^{100}\binom{100}{Z}0.01^{Z}(1-0.01)^{100-Z} = 7.63\times10^{-8} \tag{5.30}$$

We have summed over all the disease cases from 10 to 100 here, because we wanted the probability that we would have obtained a result as bad, or worse, than the one which we actually achieved. This is a particular way of carrying out classical hypothesis tests, which we will dispense with later on, but for now it seems a reasonable way of testing our model.

The probability found in this case is extremely small. What does this tell us? Well, it basically says that there is something wrong with our model which we have chosen here. It could be that the actual disease incidence in the population is much higher than the 1% which we have assumed beforehand. It could also be that our assumption of *independence* is violated in this case, for example if we sampled whole households rather than individuals. This could mean that in a particular household, the chance of having the disease, if another member of your family has the disease, is substantially higher than for the population as a whole.

It is difficult to gauge what in particular is wrong with our model without knowing further details of data collection, as well as how the estimate of 1% incidence was estimated for the population. However, it does suggest that we need to do adjust one or more of our assumptions, and reformulate the model to take these into account. We should never simply accept that our model is *correct*. A model is only as good as its capability to reproduce the data which we see in real life. In this case we find it is not a good representation, and we should readjust appropriately.

5.6.3 The intelligence of a group of people

We are now tasked with formulating a model of intelligence test scores for a group of individuals for whom we have data. We are told that the test score is on a continuous scale from 0-200. We do not have any information on individual characteristics which might help us to predict scores, although we are going to, for this simplified example, assume that we do know the mean test score $\mu = 70$, and its variance $\sigma^2 = 81$ in the population (although we will relax this assumption in section 5.9). We might assume that there are a range of factors which overall result in an individual's performance on this test. For example, these might include their schooling, parental education, 'innate' ability, as well as how tired they were feeling on the day of the test. If we assume that there are a large range of such factors and the score which results is an average of all these, then we might assume that the Central Limit Theorem might be appropriate for determining the distribution of

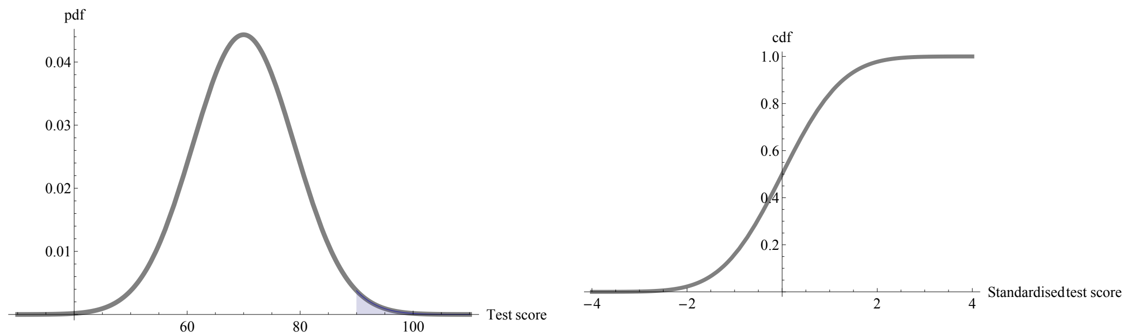


Figure 5.6: Left panel shows a normal with $\mu = 70$ and $\sigma^2 = 81$, with the area corresponding to a result as extreme as 90 indicated. This translates into a standard normal cdf shown in the right panel, which can be used to calculate this area from the first figure. This translation to the standard normal is done by taking away μ, and dividing through by σ. This is done since usually only standard normal cdf tables are available.

test scores[8]. In which case, we assume that a normal distribution for our likelihood function for an individual's test score, X:

$$p(X = \alpha|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha-\mu)^2}{2\sigma^2}} \tag{5.31}$$

Note that since this distribution is continuous, we have written p rather than Pr. The first p represents a density, whereas Pr represents a probability, which is only found in the continuous case by integrating over some bounds.

If we obtain an individual within our sample who achieved a test score of 90, we ask what's the probability of achieving a result as extreme as this? Using our idealised model, we just integrate the probability density (this is the continuous analogue to the discrete summing that we did in (5.6.2)):

[8] See section 3.6 for an introduction to the Central Limit Theorem.

$$\begin{aligned} Pr(X \geq 90|\mu = 70, \sigma^2 = 81) &= \int_{90}^{\infty} \frac{1}{\sqrt{2\pi \times 10}} e^{-\frac{(\alpha-70)^2}{2\times 10}} \mathrm{d}\alpha \\ &= 1 - \Phi\left(\frac{90-70}{9}\right) \approx 0.0131 \end{aligned} \tag{5.32}$$

In (5.32), Φ stands for the value of the *standard* normal cumulative distribution function[9] at the value of 90 (see figure 5.6 for an explanation). Since we find that the probability of obtaining this data point under our current model is extremely small, we conclude that it is likely that there is something wrong with our model, and go back to examine the various assumptions that were made in deriving it.

If we also assume that information regarding one individual's test score

[9] A standard normal has mean 0, and a variance of 1. By taking away the mean of 70,

and dividing through by the standard deviation, we transform from an arbitrary mean- and variance-normal, to a *standard* one.

tells us nothing about another's[10], then we might assume *independence* for our data. We might also assume that all individuals come from the same

[10] Apart from their joint reliance on μ and σ^2.

population; resulting in a *random sample*[11]. We calculate the joint probability density for a sample of N individuals by multiplying together the individual densities:

$$P(X_1 = \alpha_1, X_2 = \alpha_2, ..., X_N = \alpha_N|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha_i-\mu)^2}{2\sigma^2}} \tag{5.33}$$

We could then use (5.6.3) to calculate the probability of obtaining a given sample of observations as extreme as the values obtained, again by integrating. However, here it would be slightly more complicated than that of (5.32) since we would have to integrate across all individuals' variables.

5.7 Exchangeability vs random sampling

We have already been introduced to the concept of a *random sample*, in developing a probability model for the disease status of patients (section 5.6.2), and the intelligence of a group of people (see section 5.6.3). The use of this term is really just a shorthand for an *independent*, and *identically-distributed* sample of data. Often however, Bayesians eschew this term in want of a (slightly) weaker condition that still allows us to write down an overall likelihood as a product of individual likelihoods in many situations.

Suppose we have a sequence of random variables representing the height of individuals in a sample of size 3: $\{H_1, H_2, H_3\}$. If this sequence is as likely as the reordered sequence: $\{H_2, H_1, H_3\}$, or any other re-ordering, then the

[11]See section 5.7 for further discussion of random samples.

sequence is said to be *exchangeable*[12].

Since the assumption of random sampling is stronger than that of exchangeability, it turns out that any random sample is automatically exchangeable. However, the converse is not necessarily true. A particular example of this is for the case of an urn containing 3 red and 3 blue balls, which are drawn at random without replacement. The probability of obtaining the sequence RBR is given by:

$$Pr(RBR) = \frac{3}{6} \times \frac{3}{5} \times \frac{2}{4} = \frac{3}{20} \tag{5.34}$$

The sequence of random variables representing the outcome of this sampling *is* exchangeable, since we have that any permutation of this sequence is equally likely:

$$\begin{aligned} Pr(BRR) &= \frac{3}{6} \times \frac{3}{5} \times \frac{2}{4} &= \frac{3}{20} \\ Pr(RRB) &= \frac{3}{6} \times \frac{2}{5} \times \frac{3}{4} &= \frac{3}{20} \end{aligned} \tag{5.35}$$

However, this sequence of random variables is *not* a random sample. The probability distribution for the first ball drawn is different to that when the second is drawn. In the first case there are 6 balls in total, with equal numbers of each. However, for the second case there are only 5 balls, and *dependent* on the first draw, there may be either more red balls or blue balls remaining.

[12]Formally, a sequence which is exchangeable requires that the joint probability distribu-

tion is invariant under any permutation of the order.

In general we may not be able to assume we have a conditional[13] random sample of observations for reasons similar to that of the urn example. However, a brilliant theory originally by Bruno de Finetti allows us to assume that a sequence behaves as if it is a random sample, *so long as it is exchangeable.* Technically this requires that we need an infinite sample of observations, but for a reasonably large sample making this approximation is reasonable.

Much of the time we will have a random sample, and so do not need to worry about any of this. However, due to this theorem, we are often free to write down an overall likelihood as the product of individual likelihoods, so long as the observations are exchangeable.

5.8 The subjectivity of model choice

It is hoped that the analysis in the preceding sections has given us a taste of how we can go about specifying a likelihood for a hitherto unknown circumstance. We start by writing down the behaviours that we want to emulate, then make simplifying assumptions, which we then use to look for an appropriate model in the literature. This model is then used to test the validity of the assumptions with the sample data. If the model struggles to explain the data, then we should go back and iteratively modify, then test our model, until it adequately explains the range of behaviours.

However, it should be re-emphasised that by its nature, a model is always a simplification of reality. As such, no one model is *correct*. There are often many models that could be used to explain the data which we have to hand. We should always take care to test each of these against its ability to explain the aspect of the data in which we are interested, and only proceed with it if it is adequate in this regard. Real life is complicated, and thus with each of the assumptions that were used to justify a particular model, there will inevitably be a degree of *subjectivity*. As such, no analysis - whether Frequentist or Bayesian - can be thought to be purely *objective*. Hence, the human analyst cannot, and should not, be replaced by automata for statistical analysis. A degree of subjective judgement is always necessary in statistics, as in all other walks of life.

---

[13]Conditional on a distribution of a vector of parameters θ which sits above all the

5.9 Maximum likelihood - a short introduction

The analysis in section 5.6 assumes that we know beforehand the fraction, θ, of the populous that are predisposed to having the disease. In reality we rarely know such a thing. Often the main focus of building a statistical model is to try to estimate such parameters from our sample of data to which we have access. A popular Frequentist method for achieving this goal is the estimation strategy known as *Maximum Likelihood*. In this section we will examine how this estimation strategy yields estimates of parameters.

The principle of Maximum Likelihood estimation is simple. Firstly, we assume a model which we use to approximate the data generating process which resulted in our sample, based on the various assumptions about the real life process which we make. We then calculate what is known as the joint probability of obtaining the sample of observations, assuming that we do not know the parameters which specify completely those distributions. We then choose the parameters which *maximise* the likelihood of obtaining that particular sample of observations. We will go through some simple examples to illustrate this process.

5.9.1 Estimating disease prevalence

In section 5.6.2 we assumed that we knew beforehand the fraction of individuals who are disease-positive within the population. As mentioned previously, it is uncommon that such a thing be known before carrying out

---
observations.

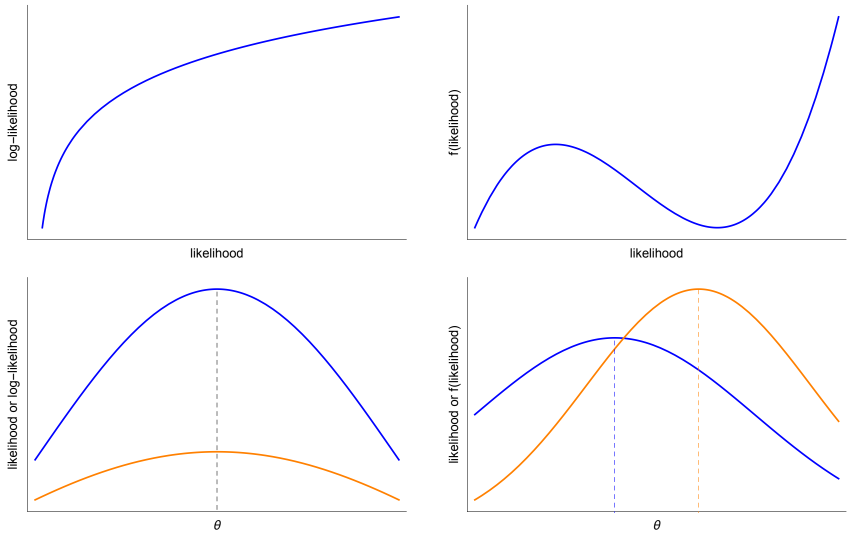


Figure 5.7: The monotonicity of log-likelihood (top-left), means that the peaks of likelihood and log-likelihood coincide (bottom-left). However, this is not the case for an arbitrary function (top-right and bottom-right).**Add legends to the bottom two graphs.**

an analysis. If in a sample of 100 individuals, 10 test positively[14], and we make the same assumptions as in section 5.6.2 - that of a random sample - then we can write down the overall likelihood function using (5.29) as:

$$L(\theta|data) = \binom{100}{10}\theta^{10}(1-\theta)^{90} \tag{5.36}$$

Remember, that since we are varying θ and holding the data constant here, that (5.36) is a *likelihood*, not a probability. We then need to simply choose θ so that we can maximise the likelihood. We could simply differentiate (5.36) as it stands, and set the derivative equal to 0; rearranging the resultant equation for θ. However, to make life a little easier for us, we are first going to take the *log* of this expression, then differentiate it, setting the derivative to 0; resulting in the same value of θ. We are able to do this because of the simple properties of the log transformation (see figure 5.7):

[14]Assuming for simplicity that there are no false-positives.

$$l(\theta|data) = Log\left(L(\theta|data)\right) = log\binom{100}{10} + 10log(\theta) + 90log(1-\theta) \tag{5.37}$$

Where to get the result (5.37), we have used the log rules:

$$\begin{aligned} log(ab) &= log(a) + log(b) \\ log(a^b) &= blog(a) \end{aligned} \tag{5.38}$$

We can now simply differentiate the log-likelihood $l(\theta|data)$:

$$\frac{\partial l}{\partial \theta} = \frac{10}{\hat{\theta}} - \frac{90}{1-\hat{\theta}} = 0 \tag{5.39}$$

If we set the derivative to 0 we then obtain the maximum likelihood *estimate*, $\hat{\theta} = \frac{1}{10}$ (see figure 5.8).

This estimator makes sense intuitively. The value of the parameter which results in the highest likelihood of obtaining the data occurs when the population prevalence exactly matches that obtained in our sample. In general if we found a number β of individuals out of a sample of size N, who were disease-positive, then we would again find that the preceding analysis results in an estimator[15] of the disease prevalence exactly equal to that in our sample:

$$\hat{\theta} = \frac{\beta}{N} \tag{5.40}$$

5.9.2 Estimating the mean and variance in intelligence scores

We are given a sample of individuals with test scores $\{75, 71\}$, and we model the test scores using a normal likelihood as described in section 5.6.3:

$$L(\mu, \sigma^2 | X_1 = 75, X_2 = 71) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(75-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(71-\mu)^2}{2\sigma^2}} \tag{5.41}$$

[15]An estimator is a mathematical function which outputs an estimate of a parameter in

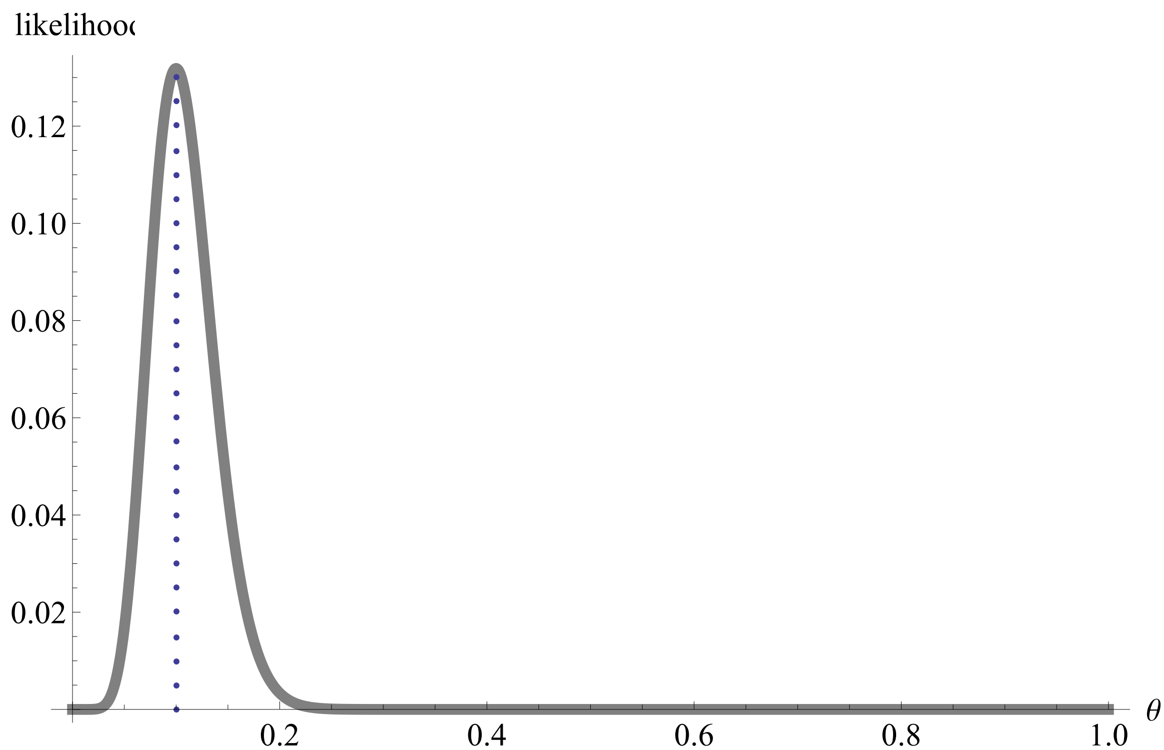


Figure 5.8: Log-likelihood of disease prevalence from section 5.9.1 as a function of the proportion of individuals which have the disease in a population, θ. The dotted line shows the maximum likelihood estimate $\hat{\theta} = 1/10$.

We can then proceed as we did in section 5.9.1 by taking the log of this expression before we differentiate it:

$$l(\mu, \sigma^2|X_1 = 75, X_2 = 71) = 2log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(75-\mu)^2}{2\sigma^2} - \frac{(71-\mu)^2}{2\sigma^2} \quad (5.42)$$

Where we have again used the log rules in (5.38) to achieve (5.42). We can now proceed to differentiate (5.42) with respect to both variables, holding the other constant, setting each to 0:

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{(75-\hat{\mu})}{\hat{\sigma^2}} + \frac{(71-\hat{\mu})}{\hat{\sigma^2}} = 0 \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{1}{\hat{\sigma^2}} + \frac{(75-\hat{\mu})^2 + (71-\hat{\mu})^2}{2\hat{\sigma^4}} = 0 \end{aligned} \quad (5.43)$$

The first of these expressions yields $\hat{\mu} = \frac{71+75}{2} = 73$, which when put into the second gives:

$$\hat{\sigma^2} = \frac{1}{2}\left[(75-73)^2 + (71-73)^2)\right] = 4 \quad (5.44)$$

Notice that the maximum likelihood estimators for the population mean

our model.

and variance are for this case the *sample mean* and *sample variance*[16]. In fact, this holds for the case of N individuals' data, then the maximum likelihood estimators for this case would be:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X} \tag{5.45}$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 = s^2 \tag{5.46}$$

5.10 Frequentist inference in Maximum Likelihood

We have now detailed how to derive point estimates of parameters using the method of maximum likelihood. However, at the moment we are unable to make any conclusions about the population. This is because we do not have any idea as to whether we obtained a particular estimate of a parameter due to picking a weird sample, or because it *actually* has a value in the population which is at this value. Frequentists get round this by examining a graph of log-likelihood near the maximum likelihood point estimate (see figure 5.9). If the log-likelihood is strongly peaked near the maximum likelihood estimate, then this suggests that only a small range of parameters would yield a similar valued likelihood. By contrast, if the log-likelihood is gently peaked near the ML estimate, then it is feasible that a large range of parameters would yield estimates close to this value. In the latter case, it seems logical that we should be less confident in the particular value of the parameter which is given by maximum likelihood. We can measure the 'peakedness' in the log-likelihood by looking at the magnitude

[16]Albeit a biased estimator of the population variance. The unbiased estimator would

divide by 1, rather than 2.

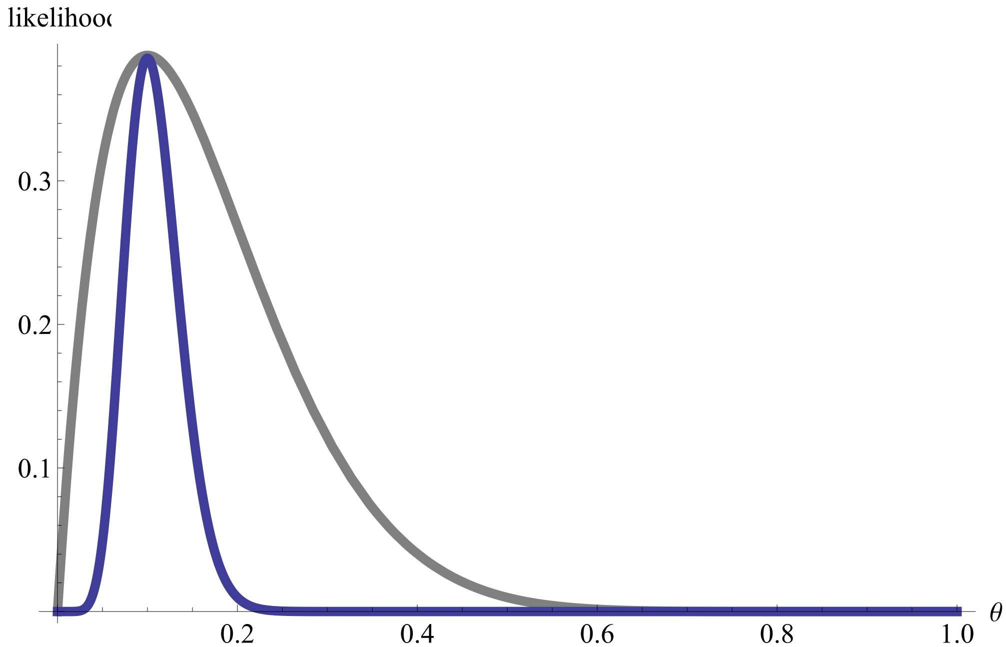


Figure 5.9: Two likelihoods which result in the same maximum likelihood estimates of parameters, at 0.1. The gray likelihood is less strongly-peaked, meaning we can be less confident about the estimates.

of the second derivative[17] of the function at the ML point estimate value. The more curved the log-likelihood, the more confident we can be of our estimated parameter value, and any conclusions drawn from this. Note however, that the Frequentist inference is not based on proper probability distributions (since we infer based on a likelihood). This contrasts with the Bayesian method which, by its nature, allows for a more adequate description of parameters, using probability distributions.

5.11 Chapter summary

We should now understand what is meant by a likelihood, and how to build probabilistic models of real life processes. However, the difficulty of modelling a process is governed by its degree of complexity and sensitivity

[17]The first derivative gives the gradient, the second derivative gives the rate of change of the gradient - a measure of curvature.

to violations of assumptions. Further we should also understand how the Frequentist method of Maximum Likelihood can be used to yield point estimates of parameters. We are however, currently restricted in our ability to make inferences based on full probability distributions over parameters. Bayes' rule tells us how we can convert a likelihood - itself not a proper probability distribution - to a posterior (*correct*) probability distribution for parameters. In order to use to do this though, we need to understand what is meant by a *prior* distribution and how we can specify this distribution to suit the particular situation. This is what is covered in the next chapter.

5.12 Chapter outcomes

The reader should now be familiar with the following concepts:

1. The difference between likelihoods and probabilities.
2. The need for probabilistic models in some circumstances.
3. How to go about choosing an appropriate likelihood for a given situation.
4. The exchangeability assumption.
5. Maximum likelihood, and how to carry out inference in this framework.

5.13 Problem set

5.13.1 Blog blues.

We suppose that visits to your newly-launched blog occur sporadically. Suppose you were interested in the length of time between consecutive first-time visits to your homepage. You collect the time data for a given 100 visits to your blog for a particular time period and day, and you set about thinking about building a statistical model.

What assumptions might you make about the first-time visits?

What model might be appropriate to model the time between visits?

Algebraically derive an estimate of the mean number of visits per hour

Data analysis:

you collect time data from Google Analytics[18] for 1000 visits. The data set is called XXXX. **Derive an estimate of the mean number of visits per hour.**

Graph the log likelihood near your estimated value. What does this show? Why don't we plot the likelihood?

Estimate confidence intervals around your parameter.

What is the probability that you will wait:

1. 1 minutes or more.
2. 5 minutes or more.
3. Half an hour or more.

before your next visit?

Evaluate your model.

What alternative models might be useful here?

What are the assumptions behind these models?

Estimate the parameters of your new model.

Use your new model to estimate the probability that you will wait:

1. 5 minutes.

[18] A popular provider of website analytics data.

2. 20 minutes.

3. 1 hour.

before your next visit?

Hints: the exponential is to the poisson model, what the ? is to the negative binomial.

5.13.2 Violent crime counts in New York counties

In data file XXX we have compiled a data set of the population, violent crime count and unemployment across New York counties in 2014.

Graph the violent crime account against population. What type of relationship does this suggest?

A simple model

A simple model here might be to assume that the crime count in a particular county is related to the population size by a poisson model:

$$crime_i \sim poisson(n_i\theta) \tag{5.47}$$

What are the assumptions of this model?

Estimate the parameter θ from the data. What does this parameter represent?

Do these assumptions seem realistic?

Estimate a measure of uncertainty in your estimates.

Evaluate the performance of your model.

A better model

An alternative model allows each county to be heterogeneous with respect to its susceptibility to crime, and has a specification of the form:

$$crime_i \sim poisson(n_i \theta_i) \tag{5.48}$$

Why is this model better?

What factors might affect θ_i?

Write down a new model specification taking into account the previous point.

Estimate the parameters of this new specification.

How does this new model compare to the previous iteration?

What alternative specifications might be worth attempting?

5.13.3 Monte Carlo evaluation of the performance of MLE in R

Ben to add later.

5.13.4 The sample mean as MLE

Ben to add later.

Chapter 6

Priors

6.1 Chapter Mission statement

At the end of this chapter a reader will know what is meant by a prior, and the different philosophies that are used to understand and construct them.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{6.1}$$

6.2 Chapter goals

Bayes' rule tells us how to convert a likelihood - itself not a proper probability distribution - into a posterior probability distribution for parameters, which can then be used for inference. We are required in the numerator to multiply the likelihood by a pre-experimental weighting of each set of parameter values described by a probability distribution, which is known as a *prior*. Priors are without doubt the most controversial aspect of Bayesian statistics, with opponents criticising its inherent *subjectivity*. It is hoped that by the end of the chapter we will have convinced the reader that, not only is subjectivity inherent in *all* statistical models - both Frequentist and Bayesian

- but the explicit subjectivity of priors is more transparent, and hence open to interrogation, than the implicit subjectivity abound elsewhere.

This chapter will also explain the differing interpretations which are ascribed to priors. The reader will come to understand the types of method that can be used to construct prior distributions, and how they can be chosen to be minimally subjective, or otherwise to contain informative pre-experimental insights from data or opinion. Finally, the reader will understand that if significant data are available then the conclusions drawn should be insensitive to the initial choice of prior.

Inevitably, this chapter will be slightly more philosophical and abstract than other parts of this book, but it is hoped that the examples given will be sufficient to ensure its practical use.

6.3 What are priors, and what do they represent?

Chapter 5 introduced us to the concept of formulating a likelihood, and how this can be used to derive Frequentist estimates of parameters, using the method of maximum likelihood. This pre-supposes that the parameters in question are immutable, fixed quantities that actually exist, and can be estimated by methods that can be repeated, or imagined to be repeated many times [7]. As Gill (2007) indicates, this is unrealistic for the vast majority of social science research.

> It is simply not possible to rerun elections, repeat surveys under exactly the same conditions, replay the stock market with exactly matching market forces, or re-expose clinical subjects to identical stimuli.

Furthermore, parameters only exist because we have *invented* a model, hence we should innately be suspicious of any analysis which assumes an existence of a single certain value for any aspect of these abstractions.

For Bayesians, it is the data that are treated as fixed, and the parameters that vary. We know that the likelhood - however useful - is not a proper probability distribution. Bayes' rule tells us how to combine a likelihood with something called a *prior* to obtain a proper posterior distribution for

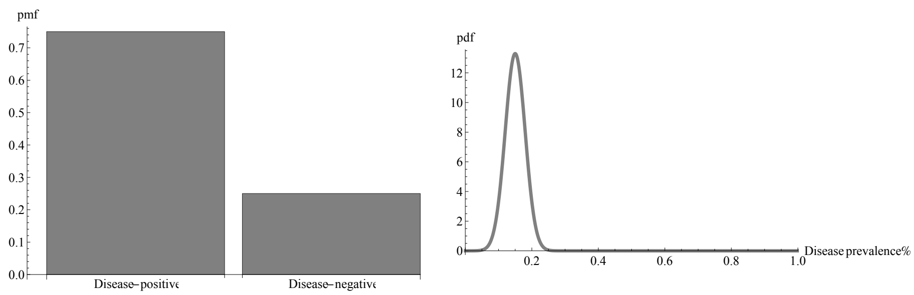


Figure 6.1: Left - a prior for a doctor's pre-testing diagnostic probability of an individual having a disease. Right - a prior which represents pre-sample uncertainty in disease prevalence.

the parameter in question, which can then be used for inference. But what does it actually mean for a parameter to have a prior distribution?

Gelman et al. (2013) suggests that there are two different interpretations of priors: the *state of knowledge* interpretation, where we specify our knowledge and uncertainty in a parameter as if regarding it as a draw from a probability distribution; alternatively in the more objective *population* interpretation where the current value of a parameter is the result of a draw from a true population distribution [5]. In both viewpoints the model parameters are not viewed as static, unwavering constants as they are taken to be in Frequentist theory.

If we adopt the *subjective* state of knowledge viewpoint above, then we can think of the prior as representing our pre-experimental/data certainty in the parameter in question. For example, imagine that a Doctor is asked to evaluate the probability that a given individual has a particular disease, before the results of a blood test become available. Using their knowledge of the patient's history and their expertise on the particular condition, they assign a prior disease probability of 75% (see figure 6.1).

Alternatively, imagine we are tasked with estimating the proportion of the UK population that has this disease. We may have some idea of its prevalence, as well as the variance in the mean prevalence of a disease across a range of previous samples of individuals which have been tested. In this case, the prior is continuous and represents our uncertainty in our estimate of the prevalence (see figure 6.1).

In all cases a prior is a proper probability distribution, and hence can be used to elicit our prior expectations as to the value of a parameter. For example, we could use the prior probability distribution for the proportion of individuals having a particular disorder in figure 6.1 to estimate a pre-experimental mean of approximately 15% prevalence.

Adopting the *population* perspective, we imagine the value of a parameter of current interest to be drawn from a population distribution. If we imagine the process of flipping a coin, we could if we knew the angle at which it is tossed, as well as the height from which it is thrown above the surface[1] predict deterministically the side on which the coin would fall face up. We could then hypothetically enumerate the (infinitely) many angles and heights of the coin, and for each set determine whether the coin would fall face up or down. Each time we throw the coin we are implicitly choosing an angle and height from the set of all possible combinations, which determines whether a heads or tails falls face up. Some ranges of the angle and the height will be more frequently chosen than others, albeit relatively agnostic with regards to final state of the coin. Hence we could think of this choice as the realisation from a distribution of all possible sets. Thus we could think about the choice of angle and height as being a realisation from this *population* distribution, and hence determines the fate of the coin toss.

Interactive : see the interactive tool XXX which allows you to investigate the population interpretation of priors for local literacy rates in the US.

Alternatively, going back to the disease prevalence example, we could imagine that each time we pick a sample, the data we obtain is partly determined by the exact characteristics of the sub-populations from which these individuals were drawn. The other part of variability is sampling variation within those sub-populations. Here we can view the particular sub-population characteristics as draws from an overall population distribution of parameters, representing the entirety of the UK.

6.4 Why do we need priors at all?

A question we might ask is, why do we need priors at all? Can't we simply let the data speak for itself, without the need of these subjective beasts?

[1] Also assuming that we knew the physical properties of the coin and surface.

Frequentists without knowing it actually do use something equivalent to priors, by setting the size of statistical tests[2]. However, can't we as Bayesians side-step this subjective jump completely?

The answer to this question is provided by Bayes' rule. Its inclusion in Bayes' rule, which is the only correct way to update beliefs, means that if we are to be consistent with the laws of probability, we are required to provide this part for any Bayesian inferential procedure.

If you find this description somewhat unsatisfying, then another way of re-phrasing this argument, is that Bayes' rule is really only a way to *update* our initial beliefs, to yield new beliefs which reflect the weight of the data obtained:

$$initial\ belief \xrightarrow{Bayes'rule} new\ beliefs \tag{6.2}$$

Viewed in this light, it is clear that we need to specify an initial belief, otherwise we have nothing to update! Bayes unfortunately doesn't tell us how to formulate this initial belief, but fear not, in this chapter we will delve into how we can go about setting priors in practice.

6.5 Why don't we just normalise likelihood by choosing a unity prior?

Another question that can be asked is, 'Why can't we simply let the prior be the same for all values of θ?'; in other words set $p(\theta) = 1$ in the numerator of Bayes' rule; resulting in a posterior that takes the form of a normalised likelihood:

$$p(\theta|data) = \frac{p(data|\theta)}{p(data)} \tag{6.3}$$

This would surely negate the need for specification of a prior, and thwart all attempts to denounce Bayesian statistics as *subjective*. So why don't we do just that?

[2]See section 2.9 for a further discussion.

There is a pedantic, mathematical argument against this, which is that $p(\theta)$ must be a proper probability distribution to ensure the same properness of the posterior. If we choose $p(\theta) = 1$ (or in fact any positive constant), then the integral $\int_{-\infty}^{+\infty} p(\theta)\mathrm{d}\theta \rightarrow \infty$, and we can no longer think of the distribution, $p(\theta)$ as representing a probability distribution. It may still be possible that even if the prior is improper, that the resultant posterior also satisfies the required properties of a proper probability distribution, but care must be taken when using these distributions for inference, as technically they are *not* probability distributions, due to the abuse of Bayes' rule. In this case the posteriors can only be viewed, at best, as approximations to the result we would have obtained under some limiting prior distribution.

Another, perhaps more persuasive argument, is that by assuming all parameter sets have an equal likelihood of being chosen beforehand, then this can result in nonsensical resultant conclusions being drawn. Consider the following example:

We are given some data on a coin which has been flipped twice, with the result $\{H, H\}$. We are given the choice of deciding whether the coin is fair, with an equal chance of both heads and tails occurring, or biased with a very strong weighting towards heads. We denote fairness by a parameter $\theta = 1$, if the coin is fair, and $\theta = 0$ otherwise.

Figure 6.2 illustrates how assuming a uniform prior in this case results in a very strong posterior weighting towards the coin being biased. This is because from a likelihood perspective - $p(data|\theta)$ - if we assume that the coin is biased, then the probability of obtaining two heads is high. Whereas if we assume that the coin is fair, then the probability of obtaining this data is only $\frac{1}{4}$. Thus, by ignoring common sense - that it is likely the majority of coins are relatively unbiased - we end up with a result that is nonsensical.

$\boxed{\textit{Interactive}:}$ see the interactive tool XXX which allows you to play around with your prior probability of a coin being biased, and note its effect on the posterior.

Of course, in this example we would hope that by collecting more data, in this case, throws of the coin, we could be confident in the conclusions drawn from the likelihood. However, Bayesian analysis allows us to achieve such a goal with a smaller sample size, should we be relatively confident about our pre-data knowledge.

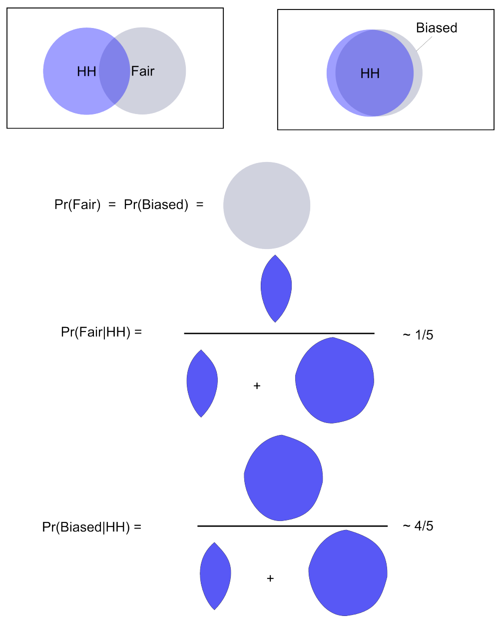


Figure 6.2: Ignoring common sense, by setting a uniform prior between a coin being fair and biased, results in an unrealistic posterior; supposing that the coin is *very* likely to be biased.

6.6 The explicit subjectivity of priors

Opponents of Bayesian approaches to inference criticise the subjectivity inherent with choice of prior. However, all analysis involves a degree of subjectivity, particularly in regard to choice of statistical model. This choice is often formulated implicitly as being *objectively* correct, with little justification or discourse given to the underlying assumptions necessary to arrive there. The statement of a prior, necessary for any full description of a Bayesian analysis, is at least *explicit*; leaving this aspect of the modelling subject to the same interrogation and academic examination to which any analysis should be subjected. A word that is often used by protagonists of Bayesian methods, is that it is *honest* due to the *explicit* statement of assumptions. The statement of pre-experimental biases actually forces the analyst to self-examine, and perhaps also leads to a decline in the temptation to manipulate the analysis to one's own ends.

6.7 Combining a prior and likelihood to form a posterior

This chapter thus far has given more attention to the philosophical and theoretical underpinnings of Bayesian analysis. Now we change tack to illustrate to the reader the mechanics behind Bayes' formula; specifically how the prior is combined with the likelihood to yield a posterior probability distribution. The following examples introduce an illustrative method, known as *Bayes' box* described in detail in [18] and [3], which illustrates the functioning of Bayes' rule, in which the parameter, prior, likelihood, and posterior are all displayed in a logical manner.

6.7.1 The Goldfish game

Imagine a covered bowl of water, containing 5 fish, each of which is red or white, and suppose we are tasked with inferring the total number of red fish which are present in the bowl, on the basis of a single fish which we pick out, and find to be red. Before we pull the fish out from the bowl, we have no prejudice for a particular number of red fish, and so suppose that all possibilities - 0 to 5 - are equally likely, and hence have the probability of

$\frac{1}{6}$ in our discrete prior. Our model for the likelihood is that a number Y of the fish are red, and that the result of an individual picking a fish from the bowl tells us nothing about future picks, apart from their joint dependence on Y. In this oversimplified example, this independence assumption seems reasonable, particularly if the fish are picked out in a randomised manner and have no distinguishing features. Further suppose that the random variable $X \in \{0, 1\}$ indicates whether a fish is white or red respectively. The analogy with the disease status of an individual described in section 5.6.1 is evident, and hence we choose a likelihood of picking a red fish of the form:

$$P(X = 1|Y = \alpha) = \frac{\alpha}{5} \tag{6.4}$$

In (6.4), $\alpha \in \{0, 1, 2, 3, 4, 5\}$ represents the number of red fish in the bowl.

We can then illustrate the functioning of Bayes' rule in the *Bayes' box* shown in table 6.1. We start by listing all the possible numbers of red fish that can exist in the bowl in the leftmost column. We then introduce our prior probabilities that we associate with each of the six potential numbers of red fish that can be in the bowl. In the third column we then calculate the likelihoods for each of the outcomes using the simple rule given in (6.4). We then multiply the prior by the likelihood in the fourth column, which on summation gives us $p(data) = \frac{1}{2}$, which we use to create a proper probability distribution for the posterior in the last column. For a mathematical description of this process see section 6.13.1.

The Bayes' box illustrates the straightforward and mechanical working of Bayes' rule for the case of discrete data. We also note that when we sum the likelihood over all possible numbers of red fish in the bowl - in this case the parameter which we are trying to infer - we find that this to be equal to 3; illustrating again that a likelihood is not a valid probability distribution. We also see that at a particular parameter value, if either the prior or the likelihood are found to be zero as is the case of 0 red fish being in the urn (impossible since we have at least one), then this ensures that the posterior distribution is zero at this point. This makes it important that we use a prior that gives a positive weight to *all* possible ranges of parameter values. The results are also displayed graphically in figure 6.3.

Now suppose that we had reason to believe that the game-maker had a prejudice towards more equal numbers of both fish, and as a result we alter our prior to have a greater weight towards these numbers of red fish(see

Table 6.1: A Bayes' box showing how to calculate the posterior for the case of drawing fish from a bowl containing a 5 fish mixture of red and white fish, one of which has been drawn and shown to be red. Here we assume that pre-experiment all possible numbers of red fish are equally likely, by adopting a uniform prior.

Number of red fish	**Prior**	**Likelihood**	**Prior x likelihood**	**Posterior=** $\frac{Prior \times Likelihood}{p(data)}$
0	1/6	0	0	0
1	1/6	1/5	1/30	1/15
2	1/6	2/5	1/15	2/15
3	1/6	3/5	1/10	3/15
4	1/6	4/5	2/15	4/15
5	1/6	1	1/6	5/15
Total	**1**	**3**	$p(data) = 1/2$	**1**

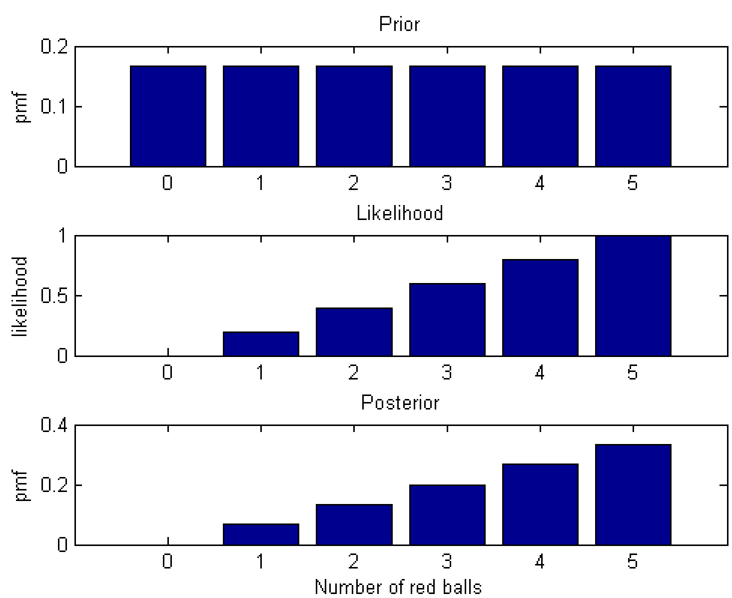


Figure 6.3: The prior, likelihood and posterior for the fish example described in 6.7.1. The prior in the upper panel gives uniform weighting to all possible numbers of red fish. This is then multiplied by the likelihood (in the middle panel) at each number of fish, and normalised to make the posterior density shown in the bottom panel.

table 6.2 and figure 6.4).

Table 6.2: A Bayes' box showing how to calculate the posterior for the case of drawing fish from a bowl containing a 5 fish mixture of red and white fish, one of which has been drawn and shown to be red. Here we assume that pre-experiment all possible numbers of red fish are equally likely, by adopting a uniform prior.

Number of red fish	**Prior**	**Likelihood**	**Prior x likelihood**	**Posterior**= $\frac{Prior \times Likelihood}{p(data)}$
0	1/12	0	0	0
1	1/6	1/5	1/30	1/15
2	1/4	2/5	1/10	1/5
3	1/4	3/5	3/20	3/10
4	1/6	4/5	2/15	4/15
5	1/12	1	1/12	1/6
Total	**1**	**3**	**1/2**	**1**

$\boxed{\textit{Interactive}:}$ see the interactive tool XXX to play the goldfish game!

6.7.2 Disease proportions revisited

Suppose that we substitute our goldfish bowl from section 6.7.1 for a sample of 100 individuals taken from the UK population. Suppose also that we continue to assert the independence of individuals within our sample, and make explicit the assumption that individuals are from the same population, and are hence identically-distributed. We are now interested in making conclusions about the overall proportion of individuals within the population who have the disease, θ. Since the parameter of interest is now continuous, we cannot use Bayes' box as there would be infinitely many rows (corresponding to the continuum of possible θ) over which to sum. Let's suppose that within our sample of 100 we find 3 of them who are disease-positive[3]. We could then use the assumptions of independence and identical-distribution to write down a likelihood of the form introduced in section 5.6.2:

[3]We also suppose that there are no false-positives here.

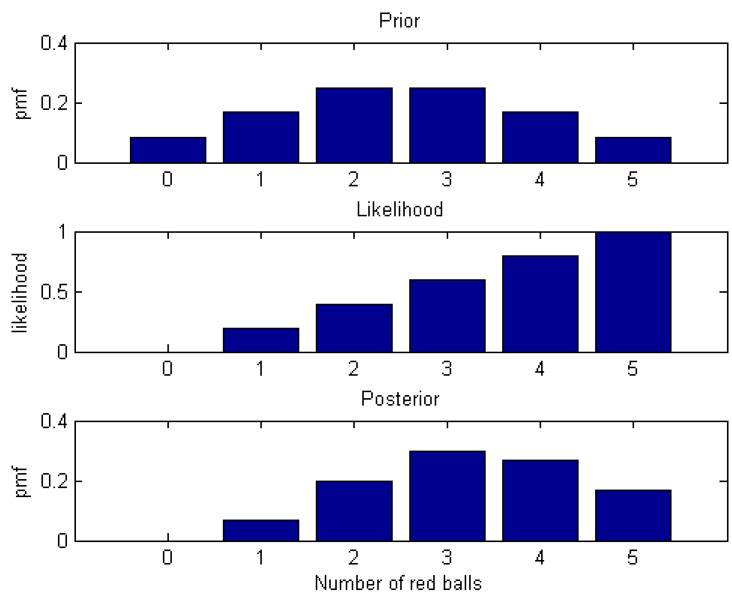


Figure 6.4: The prior, likelihood and posterior for the goldfish game example described in 6.7.1. The prior in the upper panel gives more weighting to more equal numbers of red and white fish. This is then multiplied by the likelihood (in the middle panel) at each number of fish, and normalised to make the posterior density shown in the bottom panel.

$$P(Z = 3|\theta) = \binom{100}{3}\theta^{\beta}(1-\theta)^{100-3} \tag{6.5}$$

The reason for the $\binom{100}{3} = 161,700$ term at the beginning of (6.5) is that we have to count the number of different permutations of getting 3 individuals who are disease-positive within a sample size of 100.

We suppose that at the beginning of the experiment all values of θ are equally likely. However, we would expect researchers to have a pre-experimental idea as to the most probable frequencies of the disease within the population, meaning that a flat prior which is given is likely understating a prejudice towards a certain range of θ values. Whilst, this is the case, it is often assumed in research papers - for the sake of objectivity - that priors are flat, in order to try to minimise the effect which assumptions here make on the outcome of an analysis.

The impact of using a flat prior here is that the posterior is peaked at the same value of θ as the likelihood (see figure 6.5).

Interactive effect of the prior on the posterior

See the following link to explore the effect of the prior, as well as the data sample, on the posterior for the disease proportions example.

6.8 Constructing priors

There are a number of different methodologies and philosophies when it comes to the construction of a prior density. In this section we consider briefly how priors can be engineered so as to be relatively uninformative - better-termed vague - or alternatively can be used to assemble pre-experimental knowledge in a logical manner.

6.8.1 Vague priors

When there is a premium placed on the objectivity of analysis, as is often the case in regulatory work - drug trials, public policy and the like - then use of

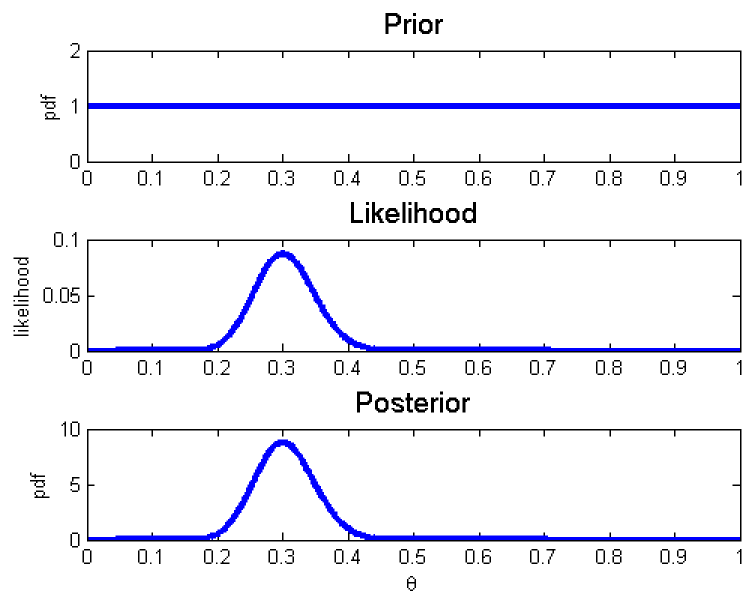


Figure 6.5: The prior, likelihood and posterior for the disease proportion example described in section 6.7.2. Each point in θ along the continuous prior curve (top panel) is multiplied by the corresponding value of likelihood (middle panel), to form the numerator of Bayes' rule. The numerator is then normalised to make the posterior probability density shown in the bottom panel.

a relatively 'uninformative' prior is often desired. If we were uncertain as to the proportion of individuals within a population who have a particular disease, then a uniform prior (see figure 6.6) is often employed to this end.

The use of a prior that has a constant value, $p(\theta) = constant$, is attractive because in this case:

$$\begin{aligned} p(\theta|data) &= \frac{p(\theta) \times p(data|\theta)}{p(data)} \\ &\propto p(\theta) \times p(data|\theta) \\ &\propto p(data|\theta) \end{aligned} \tag{6.6}$$

In (6.8.1) we thus see that the shape of the posterior distribution is solely determined by the likelihood function. This is seen as a merit of uniform priors since they 'let the data speak for itself' through the likelihood. This is used as the justification for using a flat prior in many analyses.

The flatness of the uniform prior distribution is often termed 'uninformative', but this is misleading. If we assume the same model as described in section 6.7.2, then the probability that one individual has the disease is θ, and the probability that two randomly sampled individuals both have the disease is θ^2. If we assume a flat prior for θ, then this implies a decreasing prior shown in figure 6.6 for θ^2. Furthermore, when we consider the probability that within a sample of ten individuals, all of whom are diseased, we see that a flat prior for θ implies an even more accentuated prior for this event; meaning that we beforehand give little weight to this event. For the mathematical details of these graphs see section 6.13.2.

We can hence see that even though a uniform prior for an event looks, on first glances, to convey no information, we are actually making quite informative statements about other events. This aspect of choosing flat priors is swept under the carpet for most analyses, partly because often we care most about the particular parameter for which we create a prior. All priors contain some information, so we prefer the use of the terms "vague" or "diffuse" to represent situations where a premium is placed on drawing conclusions from only the data at hand.

$\boxed{\textit{Interactive}:}$ see the interactive tool XXX which allows you to select between different priors, and investigate their effect on the probability that a number of individuals have a disease.

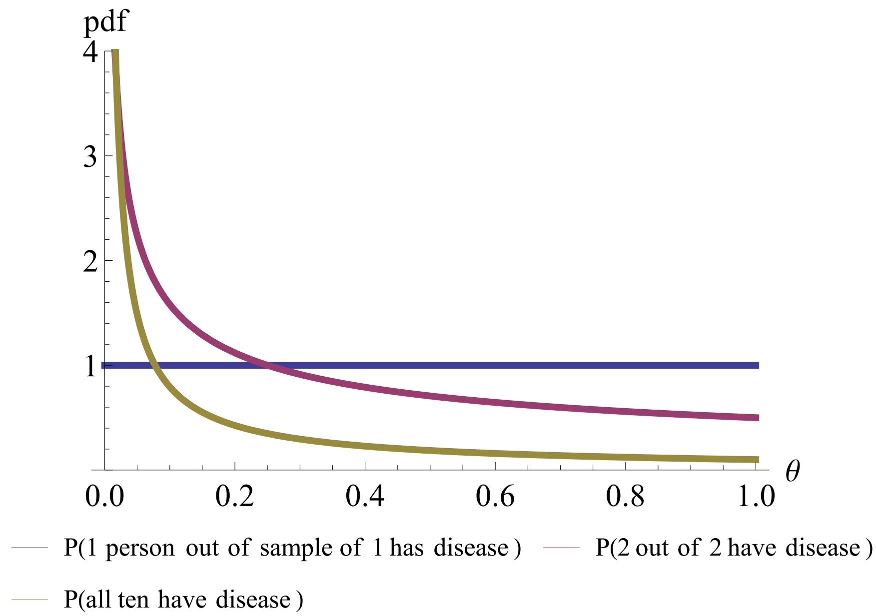


Figure 6.6: The probability density for obtaining all diseased individuals within sample sizes of 1, 2 and 10 respectively. Starting out with a flat prior for the probability that one individual has a disease has resulted in non-flat priors for the other 2 probabilities.

There are methods for constructing priors that seek to limit the information contained within priors, so as to not colour the analysis with pre-experimental prejudices. However, we will leave a discussion of these methods until chapter 11 on *Objective Bayes*.

Whilst uniform priors are relatively straightforward to specify when we aim to infer about a parameter which is bounded - such as in the previous example where $\theta \in \{0, 1\}$, or in the case of discrete parameters - we run into issues for parameters which have no predefined range. An example of this would be if we were aiming to determine the mean, μ, time of onset of lung cancer for individuals who develop the disease, after they begin to smoke. If we remove all background cases (assumed not to be caused by smoking), then μ has a lower bound of 0. However, there is no obvious point at which to draw an upper bound. A naive solution to this would be to use a prior for $\mu \sim Unif(0, \infty)$. This solution, although at first appears to be reasonable, is not viable for two reasons; one statistical, another which is practical. The statistical reason is that $\mu \sim Unif(0, \infty)$ is not a valid probability density, because any non-zero constant value for the pdf will mean that the area under the curve is ∞ because the μ axis stretches out forever. The common sense argument is that we would never ascribe the same likelihood to an individual having onset of lung cancer after 10 years as for it occurring after 250 years! The finiteness of human lifespan dictates that we select a more appropriate prior. If we were to ignore these two concerns although it is possible that the posterior could behave as a valid probability distribution[4], it would not actually be one (see section 6.5 for an explanation). A better choice of prior to use in this example would be one which ascribes zero probability to negative values of μ, and ever decreasing values of the pdf for high values of μ such as the one shown in figure 6.7. Alternatively, we could choose a uniform prior on a reasonable range of μ, and allow the pdf to be zero elsewhere (see figure 6.7).

6.8.2 Informative priors

We have seen in section 6.8.1 that priors are frequently chosen to give a strong voice to the recent data; minimising the impact of existing prejudices. There are however occasions when the choice of prior acknowledges that the analysis is based on more than the latest data. This choice of prior can be used to incorporate previous data, conclusions from older studies, or to

[4]Although not assured.

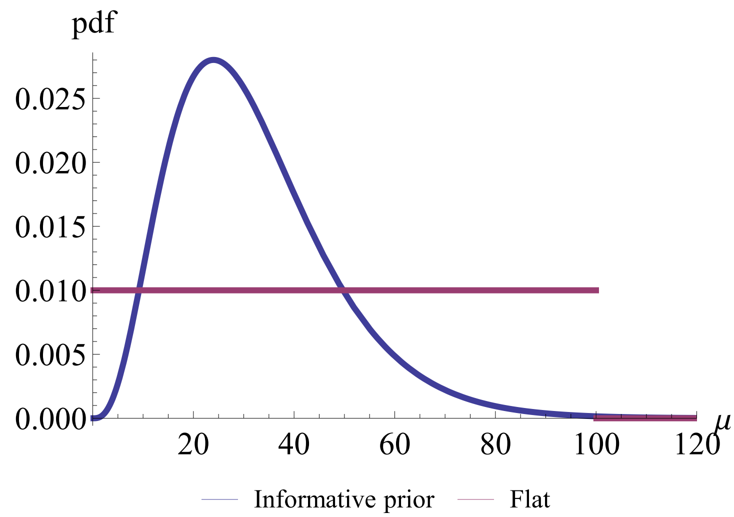


Figure 6.7: Two viable prior distributions for the average time taken before the onset of lung cancer after patients begin smoking.

include expert opinion.

In cases where data is available from previous studies, the construction of a prior can proceed methodically via a method that is known as *moment-matching*. Suppose that we have the data shown in figure 6.8 for SAT scores of past participants of a particular class. We might think that to a reasonable approximation the data could be modelled as having come from a normal distribution[5]. We typically characterise normal distributions via two parameters: its mean, μ, and variance, σ^2. In moment-matching a normal prior to this previous data, we choose the mean and variance to be equal to their sample equivalents, in this case $\mu = 1404$, and $\sigma^2 = 79,716$, respectively.

Whilst this simple methodology can result in priors that closely approximate pre-experimental datasets, note that it was a arbitrary choice to fit the first two moments of the sample. We could have used the skewness and kurtosis (measures related to the third and fourth centred moments respectively). Also, moment-matching is not Bayesian in nature, and can often be difficult to apply in practice. When we discuss hierarchical models in chapter 18, we will learn a more pure Bayesian method which can be used to create

[5] A weakness of this model is that it allows for scores outside of the 600-2400 range of

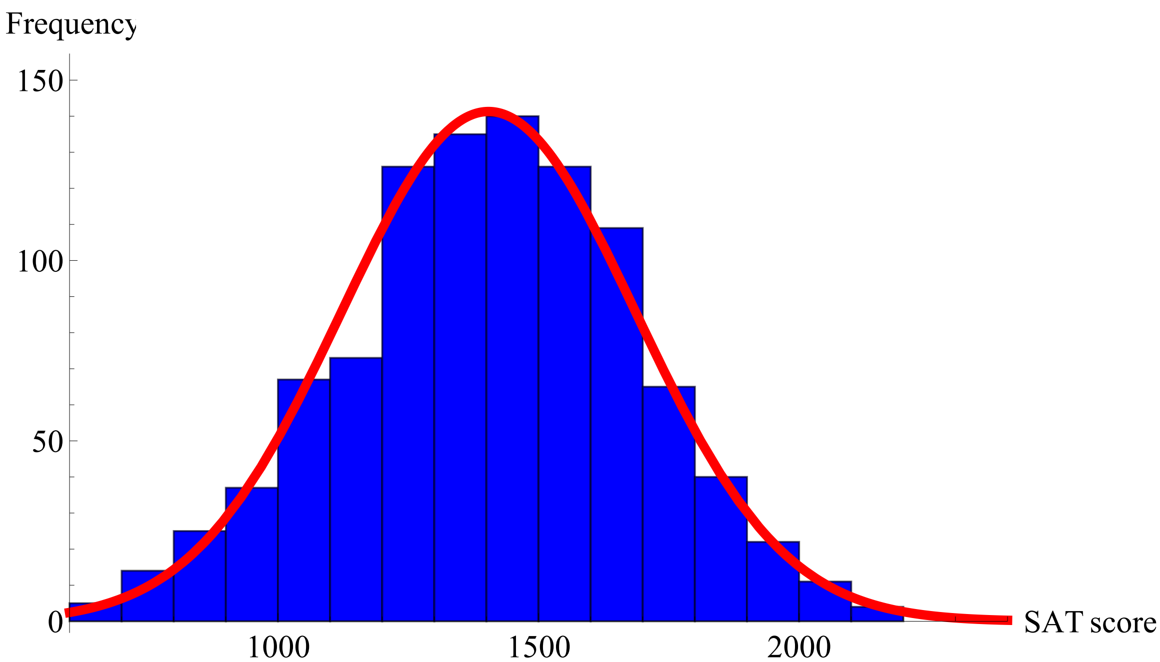


Figure 6.8: The SAT scores for past students of a class. The mean and variance of this hypothetical sample are 1404, and 79,716 respectively, which are used to fit a normal distribution to the data, and is shown in red.

prior densities.

6.8.3 The numerator of Bayes' rule determines the shape

We notice for both the examples described in sections 6.7.1 and 6.7.2 that the overall shape of the posterior distribution is determined by the prior, $p(\theta)$, multiplied by the likelihood, $p(data|\theta)$. This is the numerator of Bayes' rule:

$$p(\theta|data) = \frac{p(\theta) \times p(data|\theta)}{p(data)} \propto p(\theta) \times p(data|\theta) \tag{6.7}$$

The shape of the posterior is determined by how it varies with θ. Since the denominator is independent of θ, the numerator completely describes how the gradient and curvature of the posterior density varies with θ, which allows us to write the above $\propto p(\theta) \times p(data|\theta)$ statement. Viewed another way, the denominator is a nuisance normalisation factor which allows us to ensure that the posterior density when summed (discrete) or integrated (continuous) is equal to 1. We will return to a discussion of these concepts in depth in the chapter 7, but it doesn't hurt to see where we may be headed at present.

6.8.4 Eliciting priors

A different sort of informative prior is often required, which is not derived from prior data, but from expert opinions. In particular these priors are often required for evaluating clinical trials, and clinicians are interviewed before the trial is conducted. However, there is a raft of research in the social sciences which also make use of these methods for prior construction. Whilst there are a plethora of methods for creating priors from subjective views (see [7] for a detailed discussion), we go through a simplified example in order to explain a potential way in which these methods are used.

Suppose that we asked a range of economists to give their estimates of the 25th and 75th percentiles, $wage_{25}$ and $wage_{75}$, of the wage premium which one extra year of education spent at college commands on the job

---

permissible SAT scores.

market on average. If we were to assume a normal prior for the data, then we can relate these two quantiles back to the corresponding values of a standardised normal distribution for each expert:

$$\begin{aligned} z_{25} &= \frac{wage_{25} - \mu}{\sigma} \\ z_{75} &= \frac{wage_{75} - \mu}{\sigma} \end{aligned} \tag{6.8}$$

In (6.8), z_{25} and z_{50} are the 25th and 75th percentiles of the standard normal distribution respectively. These two simultaneous equations can be solved for each expert, giving an estimate of the mean and variance of a normal variable. These could then be averaged to get estimates of the mean and variance across all the experts. However, a better method relies on linear regression. The expressions in (6.8) can be rearranged to the following:

$$\begin{aligned} wage_{25} &= \mu + \sigma z_{25} \\ wage_{75} &= \mu + \sigma z_{75} \end{aligned} \tag{6.9}$$

We now recognise that each equation is of the form of a straight line $y = mx + c$, where in this case $c = \mu$ and $m = \sigma$. If we then fit a linear regression line to the data from all the panel, we can then use the values of the y-intercept and gradient for μ and σ to estimate the mean and square root of the variance respectively (see figure 6.9).

6.9 A strong model is not heavily influenced by priors

Returning to the example of section 6.7.2 for estimating the prevalence of a disease within a population, we now examine the effects of using an informative prior on the analysis. Suppose we choose a prior which represents our pre-data view that the prevalence of a disease within a particular population is high (see the topmost row of figure 6.10). If we only have a sample size of 10, and obtain 1 individual in our sample who tests positive for the disease we see that the posterior is located roughly equidistant between the peaks of the prior and likelihood functions respectively (see the left hand column of figure 6.10). Now if we increase the sample size to 100, keeping the same percentage of individuals who are disease-positive within our sample, we then find that the posterior is peaked much closer to the

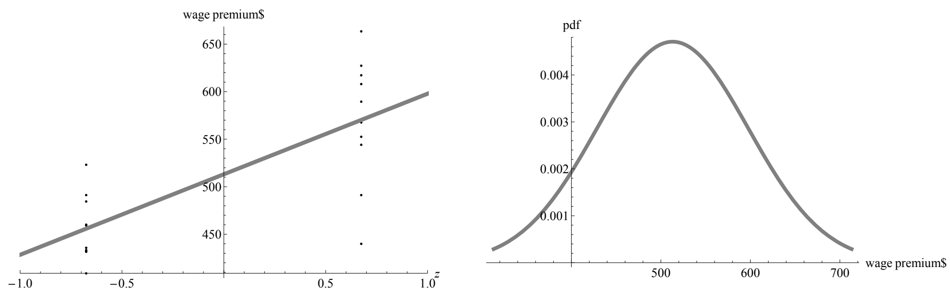


Figure 6.9: Hypothetical data for the 25th and 75th percentiles of the estimated wage premium from 10 experts. In the left hand panel we regress these percentiles on the corresponding percentiles from a standard normal distribution, yielding estimates of the mean and variance of a normal prior, which is shown on the right.

position of the likelihood peak (see the middle column of figure 6.10). If we increase sample size further, maintaining the percentage of individuals with a disease in the sample, we see that the posterior peak's position appears indistinguishable from that of the likelihood (see the rightmost column of figure 6.10).

We can see from figure 6.10 that the effect of the prior on the posterior density decreases as we collect more data. Alternatively, we see that the likelihood - the effect of current data - increases as we have access to further data points. This makes intuitive sense, since when we collect more evidence that comes solely from the data we should lend this source more weight, and pay less attention to our pre-experimental prejudices.

In general, in Bayesian analysis, when we collect more data our conclusions become less influenced by priors. The use of a prior allows us to make inferences in small sample sizes by using pre-experimental knowledge of a situation, but in larger samples, and for more appropriate models, we should see the effect of choice of priors decline. We have an obligation to report when choice of priors heavily influences the conclusions that we draw from an analysis, and *sensitivity analysis* is a field which actually allows a range of priors to be specified, and combined into a single analysis. However, if we have sufficient data and a strong model, then we should see that the conclusions we draw are not heavily affected by choice of priors within a sensible range.

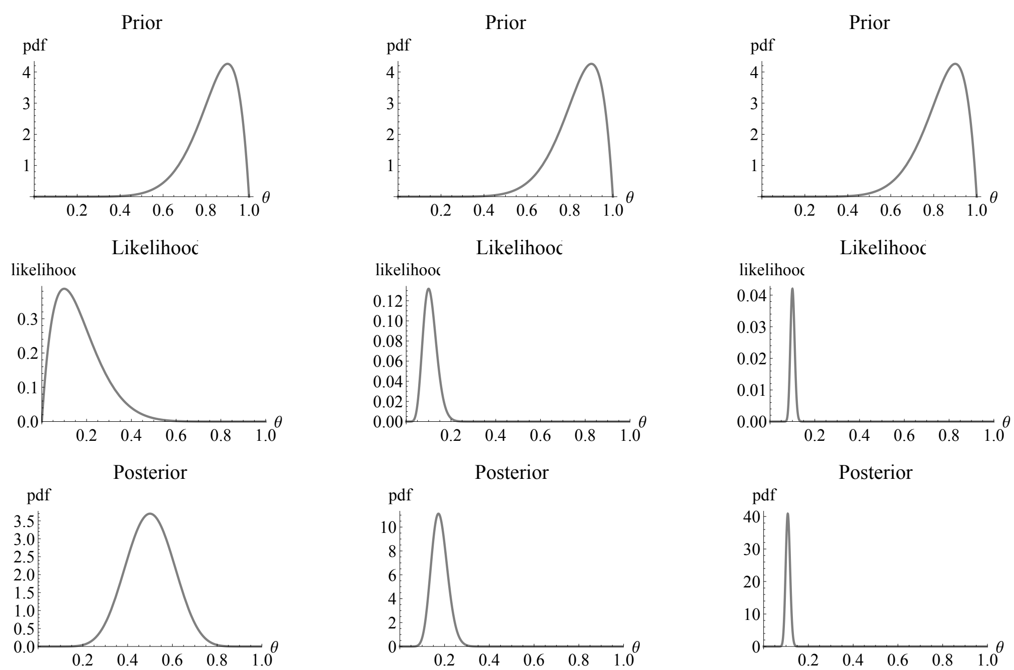


Figure 6.10: The effect of increasing sample size on the posterior density for the prevalence of a disease in a population. The leftmost column has N=10, the middle N=100, and the rightmost N=1,000. All three have the same proportion of disease cases in the sample.

6.10 Chapter summary

We now know that a *prior* is a probability distribution that represents our pre-experimental/-data knowledge about a particular situation. We also understand the importance of selecting a proper prior density, and the need to test and interpret a posterior carefully that results from using an improper prior. Further we understand that when an emphasis is placed on drawing conclusions solely from the data, that a vague prior may be most appropriate. This contrasts with situations in which we wish to use pre-experimental data or expert knowledge to help us to draw conclusions, in which case we may choose a more informative prior. In all cases however, we realise the need to be aware of the how sensitive our inferences are to choice of prior. We also realise that as the amount of data increases, or a better model is chosen, then the posterior density is less sensitive to choice of prior.

We are now nearly in a position to start doing Bayesian analysis, all that we have left to cover is the denominator of Bayes' rule. This aspect appears relatively benign on first glances, but is actually where the difficulty lies in Bayesian approaches to inference. Appropriately then we devote the next chapter to studying this final part of Bayes' rule.

6.11 Chapter outcomes

The reader should now be familiar with the following concepts:

1. Why do we need priors for Bayesian analysis?
2. The need to use proper priors in order to ensure posterior probability distributions result.
3. Bayes' box for discrete parameters.
4. How Bayes' rule can be used to combine a prior and a likelihood to yield a posterior.
5. The difference between vague and informative priors.
6. How expert knowledge can be encoded in priors.
7. A strong model is non overly-sensitive to priors.

6.12 Problem set

6.12.1 Counting sheep

Suppose in an attempt to stay awake you decide to construct a probabilistic model for counting sheep which jump over a one-way fence into a neighbouring field. You know the total number of sheep on your side of the fence to be 100, but do not know the probability of a single sheep jumping the fence for the time period that you are awake.

What likelihood model might you use here?

What are the assumptions underpinning this model?

Introducing a prior

Suppose that you choose to use a Beta prior to represent your pre-sheep-viewing knowledge about the probability of an individual sheep jumping the fence.

Graph the resultant prior for the following specific priors:

1. $Beta(1,1)$

2. $Beta(0.5,0.5)$

3. $Beta(0,0)$ - hint: consider this $\lim_{\epsilon \to 0} Beta(\epsilon,\epsilon)$.

Which of the previous priors is most uninformative?

Suppose that you observe 10 sheep jumping over the fence, calculate the posterior distribution for each of the different priors using your chosen likelihood.

For what numbers of jumping sheep would one of your posteriors run into problems?

6.12.2 Investigating priors through US elections

Suppose that the probability that the Republicans win in any particular US state is given by a probability θ. If we assumed a prior distribution on θ that is given by:

1. $p(\theta) = 1$
2. $\theta \sim Beta(2, 2)$
3. $\theta \sim Beta(0.5, 0.5)$

In each case calculate what this implies about the probability that out of 10 US states, the Republicans win in:

1. none of them.
2. 5 of them.
3. 8 of them.

6.12.3 Choosing prior distributions.

If we choose a prior distribution that sets $p(log(\sigma)) = 1$, what does this imply about the distribution of $p(\sigma)$?

For a parameter θ, which choice of prior will leave it invariant under transformations?

Pre-experimental data prior setting

Moment matching as one way of doing this. Ben to add later.

6.12.4 Expert data prior example

Ben to add later.

6.12.5 Data analysis example showing the declining importance of prior as data set increases in size

Ben to add later.

6.13 Appendix

6.13.1 Bayes' rule for the urn

In this case the application of the discrete form Bayes' rule takes the following form:

$$
\begin{aligned}
P(Y=\alpha|X=1) &= \frac{P(X=1|Y=\alpha)\times P(Y=\alpha)}{P(X=1)} \\
&= \frac{P(X=1|Y=\alpha)\times P(Y=\alpha)}{\sum\limits_{\alpha=0}^{5} P(X=1|Y=\alpha)\times P(Y=\alpha)} \\
&= \frac{\frac{\alpha}{5}\times\frac{1}{6}}{\sum\limits_{\alpha=0}^{5}\frac{\alpha}{5}\times\frac{1}{6}}
\end{aligned}
\tag{6.10}
$$

6.13.2 The probabilities of having a disease

We assume that the probability of an individual having a disease is θ, and we assume a uniform prior on this probability, $p(\theta) = 1$. We can calculate the probability that out of a sample of two, $P(Y) = P(\theta^2)$ by applying the change of variables rule:

$$P(Y) = P(\theta(Y)) \times |\theta'(Y)| \tag{6.11}$$

In (6.11), $\theta(Y) = Y^{-\frac{1}{2}}$ is the inverse of $Y = \theta^2$, and θ' means derivative wrt Y. Now substituting in this, we derive the probability density for two individuals having the disease:

$$P(Y) = \frac{1}{2\sqrt{Y}} \tag{6.12}$$

Chapter 7

The devil's in the denominator

7.1 Chapter mission

At the end of this chapter, the reader will understand what is represented by the denominator term, $p(data)$, in Bayes' rule. Furthermore, they will also have an appreciation of the inherent complexity of this term, and similar expressions encountered in applied analysis, as well as an idea of how modern computational methods can be used to bi-pass this.

Insert a graphic with the likelihood part of Bayes' formula circled, as in the equation shown below for the part highlighted in blue.

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)} \tag{7.1}$$

7.2 Chapter goals

Bayesian inference uses probability distributions, called *posteriors*, to make inferences about the world at large. However, to be able to use these powerful tools, we must ensure they are probability distributions. The denominator of Bayes' rule, $p(data)$, ensures that the posterior distribution is a *valid* probability distribution, by constraining the sum of its values to be 1.

$p(data)$ is a marginal probability density obtained by a sum across all parameter values of the numerator. The seeming simplicity of the previous statement belies the fact that for many circumstances its calculation can be complicated, and often practically intractable. In this chapter we will learn the circumstances when this difficultly arises, as well as a basic appreciation as to how modern computational methods sidestep this issue. We will leave the details of how these methods work in practice to part IV, but this chapter will lay the foundations for this later study.

7.3 An introduction to the denominator

7.3.1 The denominator as a normalising factor

We know from chapter 5 that the likelihood is not a valid probability density, and hence we reason that the numerator of Bayes' rule - the likelihood multiplied by the prior - is similarly not constrained to be one either. The numerator will satisfy the first condition of a valid probability density: that its values are non-negative. However, the sum of the numerator across all parameter values will not generally be 1; meaning it fails the second test.

A natural way to normalise the numerator to ensure that the posterior is a valid probability density, is to divide by its sum; thus ensuring that its transformed variable's sum is always 1. The denominator of Bayes' rule, $p(data)$, is just this normalising factor. Notice that it does not contain the parameter, θ. This is because $p(data)$ is a *marginal* probability density (see section 3.3.5), obtained by summing/integrating out all dependence on θ. This parameter-independence of the denominator ensures that the dependence of the posterior distribution $p(\theta|data)$ on θ is solely through the numerator (see sections 6.8.3 and 7.5).

There are two varieties of Bayes' rule which we will employ in this chapter, which use slightly different[1] formulations of the denominator. When θ is a discrete parameter we are required to *sum* over all possible parameter values, in order to obtain a factor which normalises the numerator:

[1] Although conceptually identical.

$$\begin{aligned} p(data) &= \sum_{All\ \theta} p(data, \theta) \\ &= \sum_{All\ \theta} p(data|\theta) \times p(\theta) \end{aligned} \tag{7.2}$$

We will leave multiple-parameter inference largely to chapter 9, although will discuss how this leads to added complexity in section 7.4. However, the method proceeds in an analogous manner to (7.2), with the single sum replaced by a number of summations[2].

For continuous parameters we use the continuous analogue of the sum - an integral - resulting in a denominator of the form:

$$\begin{aligned} p(data) &= \int_{All\ \theta} p(data, \theta)\mathrm{d}\theta \\ &= \int_{All\ \theta} p(data|\theta) \times p(\theta)\mathrm{d}\theta \end{aligned} \tag{7.3}$$

Similarly, for multiple-parameter systems the single integral is replaced by a multiple-integral. We will now demonstrate how to use (7.2) and (7.3) through two examples in sections 7.3.2 and 7.3.3 respectively.

7.3.2 Example: disease

Imagine that we are a medical practitioner tasked with evaluating the probability that a given patient has a particular disease. We use θ to represent the two possible outcomes:

$$\theta = \begin{cases} O & , Disease\ negative \\ 1 & , Disease\ positive \end{cases} \tag{7.4}$$

Using our experience and the patient's medical history we estimate that there is a probability of $\frac{1}{4}$ that this patient has the disorder; representing our prior. We then obtain test information, and are asked to re-evaluate the

[2]The number of summations corresponds to the number of parameters in the model.

probability that the patient is disease-positive. In order to do this, we are required to state our likelihood. In this case we choose a likelihood of the form:

$$Pr(\textit{test positive}|\theta) = \begin{cases} \frac{1}{10} & ,\theta = 0 \\ \frac{4}{5} & ,\theta = 1 \end{cases} \tag{7.5}$$

In (7.5), we implicitly assume that the probability of a negative test result is given by 1 minus the positive test probabilities. Also, by stating that there is a non-zero probability for $Pr(\textit{positive}|\theta = 0)$, we are assuming that false-positives do occur.

Suppose that the individual tests positive for the disease. We can now use (7.2) to calculate the denominator of Bayes' rule in this case:

$$\begin{aligned} Pr(\textit{test positive}) &= \sum_{\theta=0}^{1} Pr(\textit{test positive}|\theta) \times Pr(\theta) \\ &= Pr(\textit{test positive}|\theta = 0) \times Pr(\theta = 0) + Pr(\textit{test positive}|\theta = 1) \times Pr(\theta = 1) \\ &= \frac{1}{10} \times \frac{3}{4} + \frac{4}{5} \times \frac{1}{4} = \frac{11}{40} \end{aligned} \tag{7.6}$$

Furthermore, it turns out the denominator is also a valid probability density[3], meaning that we can calculate the counter-factual $Pr(\textit{test negative}) = 1 - Pr(\textit{test positive}) = \frac{29}{40}$. We need to be careful with interpreting this last result, since it didn't actually occur. It's best to think of $Pr(\textit{test negative})$ as the probability that we would assign to an individual testing negative before we carry out the test.

We can then use Bayes' rule to obtain the posterior probability that the individual has the disease, given that they tested positively:

[3]Due to the fact that we have removed the θ dependence that confounds attempts to

$$
\begin{aligned}
Pr(\theta = 1|test\ positive) &= \frac{Pr(test\ positive|\theta = 1) \times Pr(\theta = 1)}{Pr(test\ positive)} \\
&= \frac{\frac{4}{5} \times \frac{1}{4}}{\frac{1}{10} \times \frac{3}{4} + \frac{4}{5} \times \frac{1}{4}} \\
&= \frac{8}{11}
\end{aligned} \tag{7.7}
$$

We see that in this case, even though we started off with a fairly optimistic prejudice - a probability that the individual has the disease of $\frac{1}{4}$ - the strength of the data has shone through, and we now are fairly confident of the alternative (see figure 7.1 for a graphical depiction of this change of heart). Bayesians are fickle by design!

7.3.3 Example: the proportion of people who vote for conservatively

We now are in the position of interpreting exit polls in a general election, and are tasked with inferring the proportion of voters, θ, that have voted for the conservative party. We suppose that conservatives are relatively unpopular at the time of the election, and hence we assume that, at most, 45% of the electorate will vote for them, meaning we choose a cut-off uniform prior of

view the numerator as one.

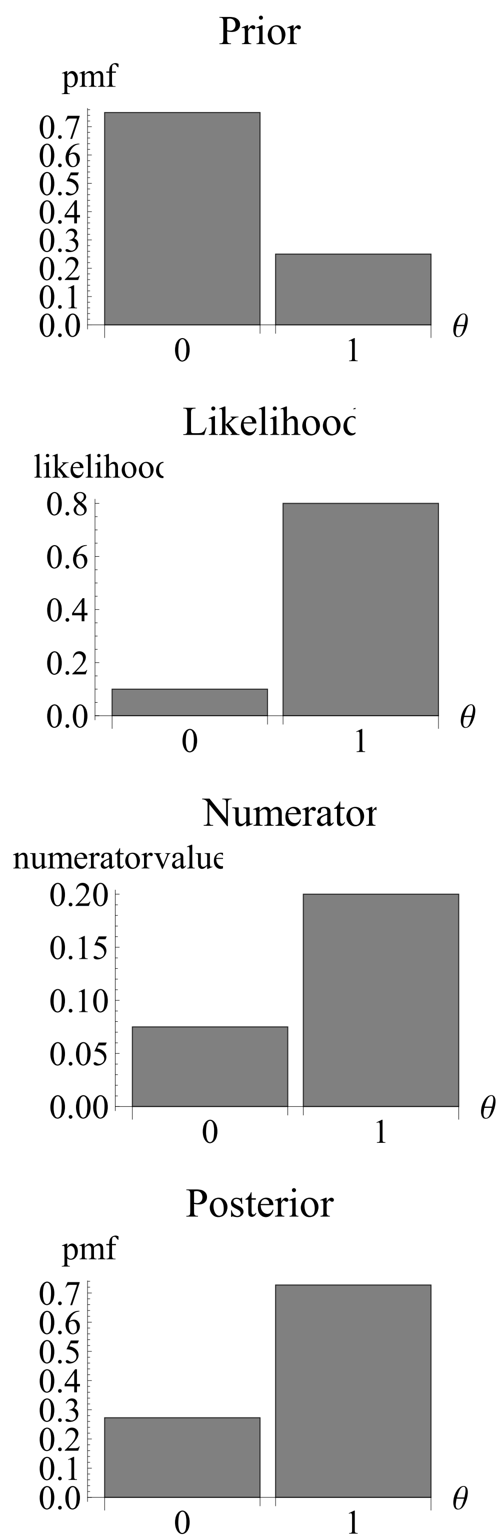


Figure 7.1: The prior is multiplied through by the likelihood, resulting in the numerator (the penultimate panel), which is then normalised by the sum over its values, to obtain the denominator.

the form shown in figure 7.2[4]. For data we obtain voter preference data from 100 individuals leaving a particular polling station. To simplify the analysis, we will assume that there are only two political parties, and all voters must choose between either of these two options. We will assume that the polling station chosen is thought to be representative of the electorate as a whole, and voters' choices are independent of one another. In this situation we can use the results of section 5.6.2, and use a binomial likelihood function:

$$Pr(Z = \beta|\theta) = \binom{100}{\beta}\theta^{\beta}(1-\theta)^{100-\beta} \tag{7.8}$$

In (7.8), Z is a variable that represents the number of individuals who vote conservatively in the sample. $\beta \in [0, 100]$ is the value which corresponds to the number of conservative voters. We assume in this case that 40 people out of the sample of 100 voted conservatively resulting in the likelihood shown in figure 7.2, which is peaked at the Maximum Likelihood estimate of $\theta = 40\%$.

We then find the denominator by using (7.3), where $\theta \in \{0, 1\}$:

$$\begin{aligned} Pr(Z = 40) &= \int_0^1 Pr(Z = 40|\theta) \times p(\theta)\mathrm{d}\theta \\ &= \int_0^{0.45} \binom{100}{40}\theta^{40}(1-\theta)^{60} \times \frac{20}{9}\mathrm{d}\theta \\ &\approx 0.018 \end{aligned} \tag{7.9}$$

In (7.9), we have used the fact that since $p(\theta) = 0$ for $\theta > 0.45$, we can restrict

[4]This isn't really a reasonable prior in this case, since it is unrealistic to allow the probability density to jump from 0 at 46% to above 2 at 45%! However, we will stick with it to

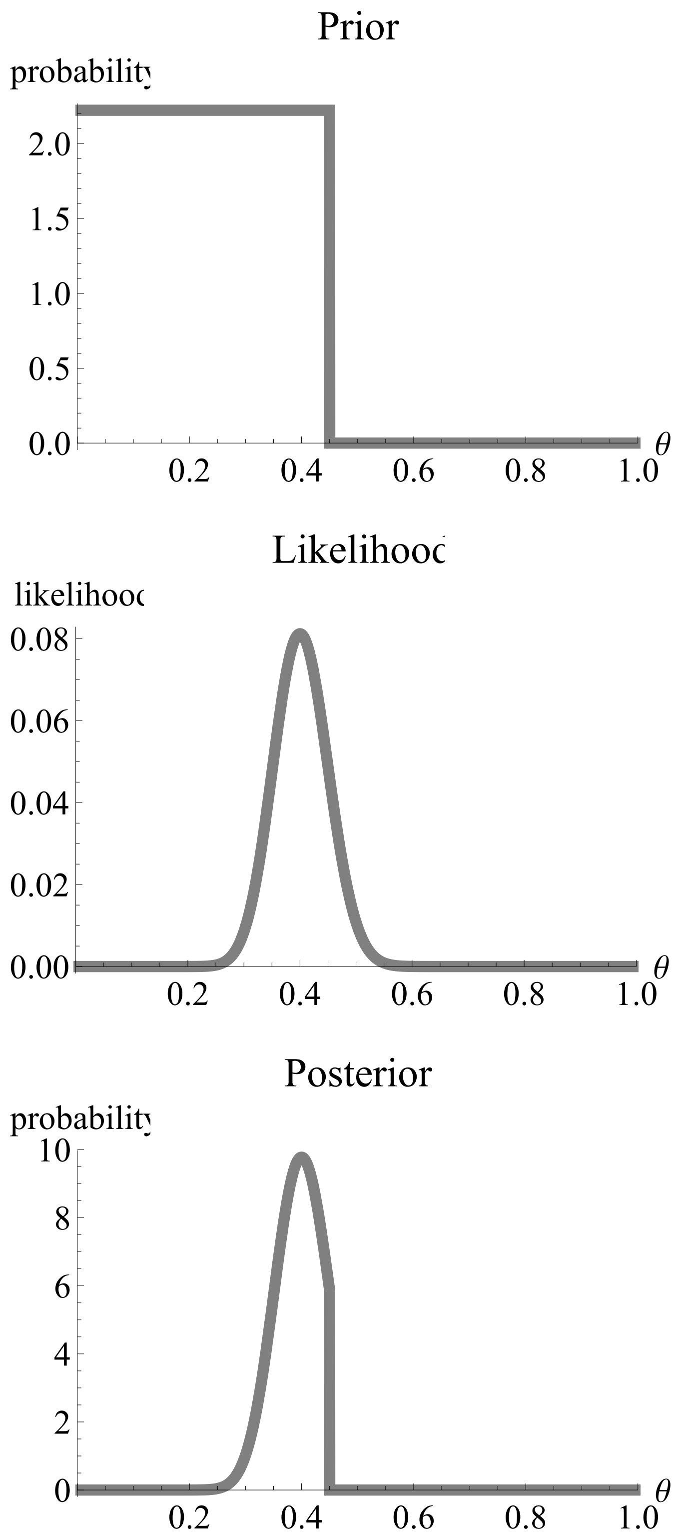


Figure 7.2: The prior, likelihood and posterior for the proportion of individuals voting for the conservative party in a general election, where we have found 40 people out of a sample of 100 voted conservative.

demonstrate its effect on inference.

the integral to only the region below that value. The value $\theta \approx 0.018$ has come by numerically integrating the second line.

Now that we have the value of the denominator, we can use it to normalise the product of the prior and the likelihood, resulting in the posterior distribution seen in figure 7.2. Notice the effect of truncating the uniform distribution at $\theta = 0.45$ is to truncate the posterior distribution at this value; resulting in a discontinuous jump in the posterior, which could be seen as an undesirable consequence of this prior.

7.3.4 The denominator as a probability

Another way to view the denominator is as the *probability of the data given choice of model*. Where *model* here encompasses both the likelihood and the prior. It is actually a *marginal* probability density that is obtained by summing/integrating out the dependence on the parameter(s) of the joint density $p(data, \theta)$:

$$\begin{aligned} p(data) &= \int_{All\ \theta} p(data|\theta) \times p(\theta) \mathrm{d}\theta \\ &= \int_{All\ \theta} p(data, \theta) \mathrm{d}\theta \end{aligned} \tag{7.10}$$

In (7.10) we have assumed that the parameter(s) is/are continuous. We have obtained the second line of (7.10) from the first by using the conditional probability formula introduced in section 3.3:

$$p(data|\theta) = \frac{p(data, \theta)}{p(\theta)} \tag{7.11}$$

We are thus able to characterise the joint density of the data and θ in Bayesian statistics. We can draw the joint density for each of the examples in sections 7.3.2 and 7.3.2 respectively, by taking the product of the likelihood and prior. In the disease example of section 7.3.2 this results in the discrete joint density shown in table 7.1, with a graphical depiction of the density shown

[4] The factor $\frac{20}{9}$ is from the uniform density for $\theta \leq 0.45$.

in figure 7.3. In the continuous case we obtain a joint probability density with a landscape of the form shown in figure 7.4.

		Disease status		
	Test Results	**Negative**	**Positive**	
Likelihood	0	0.90	0.20	
	1	0.10	0.80	
		$\times$	$\times$	
Prior		0.75	0.25	
		=	=	
	Test Results			$p(data)$
Joint density	0	0.675	0.05	**0.725**
	1	0.075	0.20	**0.275**

Table 7.1: Shows the derivation of the joint density for the disease example described in section 7.3.2. Each column of the likelihood - corresponding to a given disease status - is multiplied by the corresponding prior, resulting in the joint density. By summing the joint density across the different disease statuses of the patient, this results in $p(data)$. **Add pluses and equals to the calculation of** $p(data)$**. Also add in the posterior calculation.** See figure 7.3 for a graphical depiction of this joint density.

7.3.5 Using the denominator to choose between competing models

The denominator represents the accumulation of evidence for our particular model, with the result being a trade-off between our the data and our pre-experimental pre-conceptions. It represents the *average* fit of our model to the data across all parameter values. To see this note that the denominator is actually the expected value of the likelihood - the fit - given choice of prior[5]:

[5]This comes from the mathematical definition of the expected value of a quantity.

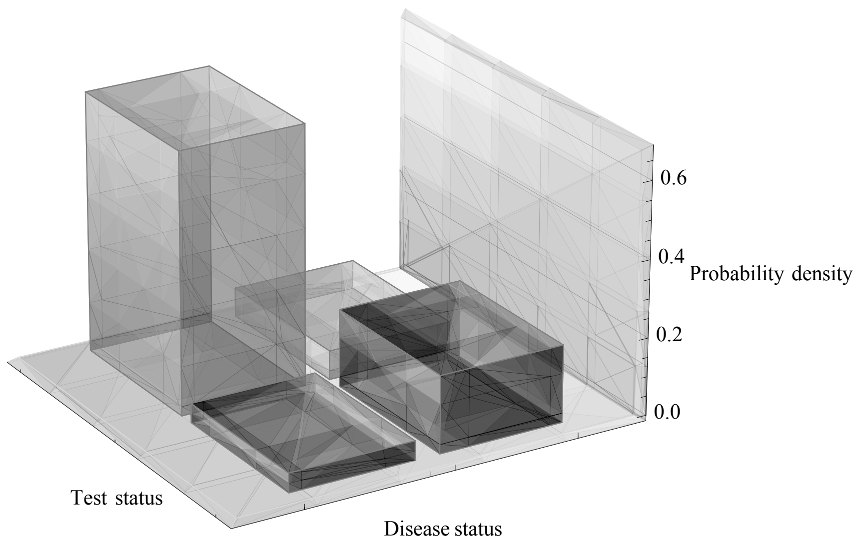


Figure 7.3: The joint density of the data and the parameter for the disease example described in section 7.3.2. When we uncover that the test result is positive, we are confined to look at the bars in dark grey; finding that the probability that an individual is diseased is significantly higher than the alternative (see the bottom panel of figure 7.1). **Perhaps redo this figure with a contour plot opposed to a 3D graph, and show how the posterior is obtained in another panel. Or just get rid of it, the table does pretty much cover it.**

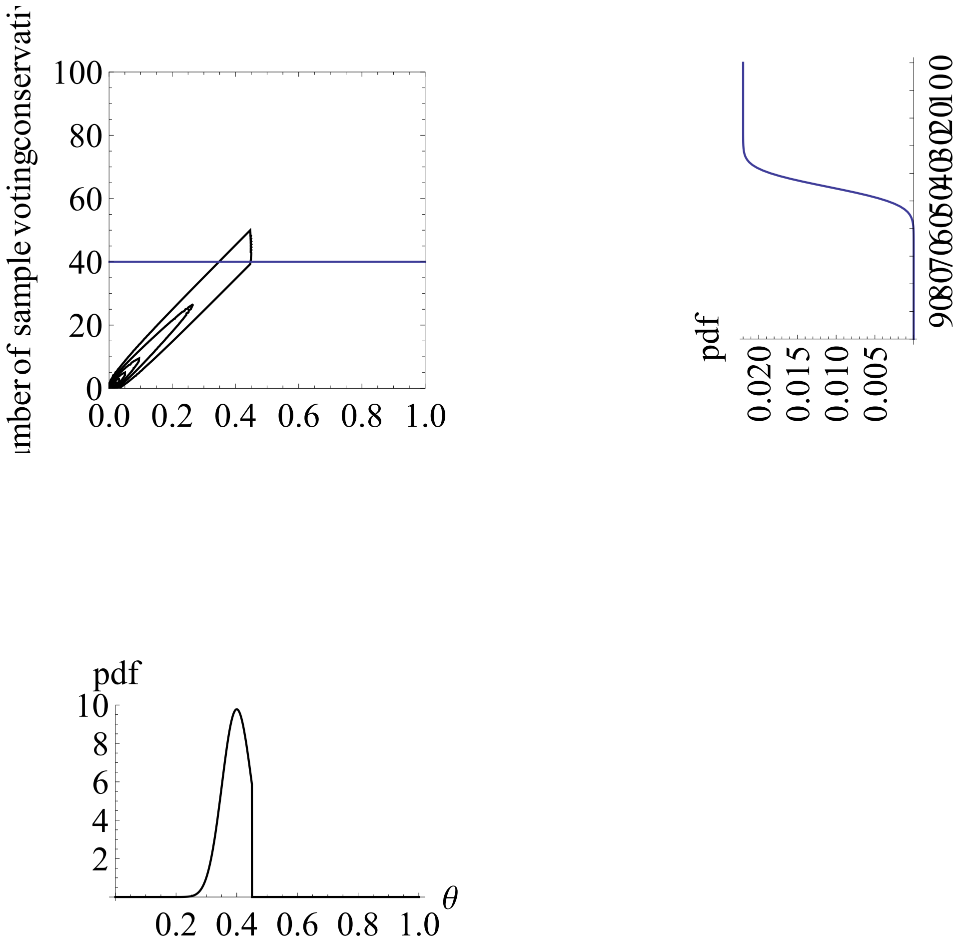


Figure 7.4: Top-left: a contour plot of the joint density of the voting example described in section 7.4. Top-right: the marginal density of $p(data)$ obtained by summing across all values of θ. Bottom-left: the posterior obtained by summing the joint density across the line shown at 40. Note that in reality the data variable is discrete, but I have drawn it here as continuous to make the plot simpler to interpret. **The line at 40 may be dashed in the final version. The axes all need to be aligned.**

$$p(data) = \mathbb{E}_{\theta}\left[p(data|\theta)\right] = \int_{All\ \theta} p(data|\theta) \times p(\theta)\mathrm{d}\theta \tag{7.12}$$

In (7.12) we have assumed that the parameter(s) are continuous, necessitating an integral rather than a discrete sum.

Since it really represents the evidence for our model, we can use it to compare two competing models. We could simply calculate the ratio of $\frac{p(data|model_1)}{p(data|model_2)}$, and use this as a guide to choose between models, but we would ideally like to do model selection in a more complete Bayesian manner. What we really care about for model choice is $p(model|data)$, rather than what we currently have $p(data|model)$. To obtain this we can use Bayes' rule, but now conditioning on choice of *model* rather than *parameter*:

$$p(model|data) = \frac{p(data|model) \times p(model)}{p(data)} \tag{7.13}$$

In (7.13), the denominator is *not* the same as that which we see in our previous applications of Bayes' rule, and represents the probability of obtaining the data across *all* models. Notice also that we also have introduced $p(model)$ which represents our prior faith in this particular model. We can now use (7.13) to choose between two models by calculating the ratio:

$$\frac{p(model_1|data)}{p(model_2|data)} = \frac{p(data|model_1)}{p(data|model_2)} \times \frac{p(model_1)}{p(model_2)} \tag{7.14}$$

If we have no prior leaning towards either of the two models then it seems reasonable to set $p(model_1) = p(model_2)$, and we are reduced to our previously proposed way of choosing between models. In fact the first ratio on the RHS of (7.14) is sufficiently used to merit its own name, the *Bayes' factor* is:

$$Bayes\ factor(model_1, model_2) = \frac{p(data|model_1)}{p(data|model_2)} \tag{7.15}$$

We shall come to discuss the usefulness of the Bayes factor in chapter 10 for choosing between models, as well as comparing hypotheses.

In both the examples discussed in sections 7.3.2 and 7.3.3, we found the denominator as a means to obtaining the posterior distribution through

Bayes' rule. However, as an ends in itself it is less useful, unless it is calculated across a number of models/hypotheses and then used to choose amongst them.

7.4 The difficulty with the denominator

We have come to realise that the denominator of Bayes' rule is obtained by summing/integrating the joint density $p(data, \theta)$, where the latter is obtained by the product of the prior and the likelihood. The examples in section 7.3.4 indicate how this procedure works when there is a single parameter in the model. However, in most real-life applications of statistics, the likelihood is a function of a number of parameters. For the case of a two parameter discrete model, the denominator is given by a double sum:

$$p(data) = \sum_{All\ \theta_1} \sum_{All\ \theta_2} p(data, \theta_1, \theta_2) \tag{7.16}$$

And for a two-dimensional continuous parameter vector, we are now required to do a double integral:

$$p(data) = \int_{All\ \theta_1} \int_{All\ \theta_2} p(data, \theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2 \tag{7.17}$$

Whilst the two-parameter forms (7.16) and (7.17) may not look more intrinsically difficult than their single parameter counterparts, (7.2) and (7.3) respectively, this aesthetic similarity is misleading, particularly in the continuous case. Whilst in the discrete case, it is possible to enumerate all parameter values, and hence - by brute force - calculate the exact value of $p(data)$, for continuous parameters, the integral may be difficult to undertake. This difficulty is amplified the more parameters we include within the model, rendering the analytic[6] calculation of the denominator practically impossible, for all but the simplest models.

[6]This just means to write down a relation for the denominator in closed form.

7.4.1 Multi-parameter discrete model example: the comorbidity between depression and anxiety

In medicine co-morbidity refers to the concurrence of two or more conditions. An example of this is the frequent coincidence of depression and anxiety in a patient. Let $D \in \{0, 1\}$ and $A \in \{0, 1\}$ be random variables representing the depression and anxiety statuses of a particular patient respectively. Now that we have two parameters, we must specify a joint prior distribution. An example prior is shown at of the top of table 7.2, in which we have also calculated the marginal prior distributions by summing over all values of the other variable. We suppose that *a priori* the clinician undertaking this case believes that the patient is unlikely to meet all the criteria necessary for them to be defined as having both disorders, which is reflected in a prior probability of $p(D = 1, A = 1) = 0.2$.

We can also use this joint distribution to calculate prior conditional probabilities. For example, we can calculate the probability that an individual has anxiety, *given* that they have depression:

$$\begin{aligned} p(A = 1|D = 1) &= \frac{p(A = 1, D = 1)}{p(D = 1)} \\ &= \frac{0.2}{0.35} \\ &= \frac{4}{7} \approx 0.57 \end{aligned} \tag{7.18}$$

This shows that it is considerably more likely that a patient has anxiety, if they are already depressed (compared with the unconditional $p(A = 1) = 0.25$), indicating our prior beliefs regarding the comorbidity of these two conditions.

We assume that the patient takes a personality diagnostic test which provides some extra information regarding whether the individual has either of these conditions. Let's assume for simplicity that the result of the test, $X \in \{0, 1\}$, has the likelihood shown in the second panel of table 7.2. The maximum likelihood estimator would be that the individual has $(D = 1, A = 1)$, with the lowest likelihood going to the disorder-free case.

We would now like to calculate the joint posterior probability of the two conditions, given that an individual tests positive ($X = 1$). We can write this using Bayes' rule, although now we must now make sure to condition

Prior			**A**		
			0	**1**	$p(D)$
	D	**0**	0.6	0.05	**0.65**
		1	0.15	0.2	**0.35**
		$p(A)$	**0.75**	**0.25**	
Likelihood (X=1)			**A**		
			0	**1**	
	D	**0**	0.05	0.4	
		1	0.4	0.8	
Numerator = Prior x Likelihood			**A**		
			0	**1**	
	D	**0**	0.03	0.02	
		1	0.06	0.16	
p(X=1)=0.03+0.03+0.06+0.16=0.27					
Posterior			**A**		
			0	**1**	$p(D\vert X=1)$
	D	**0**	0.11	0.07	**0.19**
		1	0.22	0.59	**0.81**
		$p(A\vert X=1)$	**0.33**	**0.67**	

Table 7.2: Calculation of posterior for the anxiety and depression example described.

the likelihood on both parameters. However, we can denote the parameter vector, $\boldsymbol{\theta} = (D, A)$, and apply Bayes' rule just as before:

$$\begin{aligned} p(\boldsymbol{\theta}|X=1) &= \frac{p(X=1|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(X=1)} \\ &= \frac{p(X=1|A,D) \times p(A,D)}{p(X=1)} \end{aligned} \tag{7.19}$$

In (7.19), we have simply substituted the definition of, $\boldsymbol{\theta} = (D, A)$, into the top line to get the final expression. Therefore, just like before we multiply the likelihood by the prior to obtain the numerator of Bayes' rule. We finally sum over all numerator values, and use this to obtain the posterior distribution (see table 7.2). In table 7.2 we have also calculated the marginal conditional posterior probabilities by using the law of conditional probability, and we find an 81 % probability that the individual has depression, and 67 % chance that they have anxiety. The probability that they have both disorders is 59%.

$\boxed{\textit{Interactive}:}$ see the interactive tool XXX which allows you to see how altering the priors and likelihoods affects the posterior.

7.4.2 Continuous multi-parameter example: mean and variance of IQ

We now consider a situation where the parameters of interest are continuous. It is hoped that this section will provide evidence for the complexity of analytic multi-parameter inference in Bayesian statistics, and hence by its very nature, the material covered here may be difficult to fully grasp. However, we will cover it in more detail in part III.

We suppose we are interested in estimating the mean IQ of some population of interest, of which we only possess a sample of three persons' IQ data of $\boldsymbol{IQ} = \{100, 50, 150\}$. We suppose that since intelligence - as measured by IQ - is dependent on many additive factors, and hence as an approximation we assume a normal likelihood[7]:

[7] We have used the central limit theorem here - see section 3.6 for a full explanation.

$$p(IQ_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(IQ_i-\mu)^2}{2\sigma^2}\right) \tag{7.20}$$

For sake of simplicity, we will assume that IQ is measured on a fixed scale, $IQ \in [0, 300]$. We also assume that prior independence between μ and σ^2, which means that we can calculate the joint prior by multiplying together the individual probabilities:

$$p(\mu,\sigma^2) = p(\mu) \times p(\sigma^2) \tag{7.21}$$

Since $\sigma^2 \geq 0$, we might be tempted to specify a prior distribution for $\sigma^2 \sim Unif(0,\infty)$. However, this does not appear sensible because this would assign the same probability to an infinite variance, which is not possible on finite-scaled data. A frequently-used alternative is to specify a prior as uniform in $log(\sigma^2)$ space. This serves two purposes, firstly, because the inverse of a log (the exponent) is always non-negative for real inputs, this ensures that this condition is satisfied by σ^2. Secondly, and most importantly, when we transform a uniform prior on $log(\sigma^2)$ back to σ^2 space, we find that the prior density is equivalent to[8]:

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{7.22}$$

This results in a joint prior for (μ,σ^2) shown in figure 7.5. We importantly note that this prior is improper, since $\int_0^\infty \frac{1}{\sigma^2} d\sigma^2 \to \infty$, and hence must take case when interpreting the resultant 'posterior' distribution (see section 6.8.1).

We imagine we only observe a sample of one individual, from which we would like to find the posterior distribution of the joint distribution of $(\mu,\sigma^2) = \boldsymbol{\theta}$. This is found by application of Bayes' rule:

[8] See the chapter appendix for a full mathematical treatment of this result.

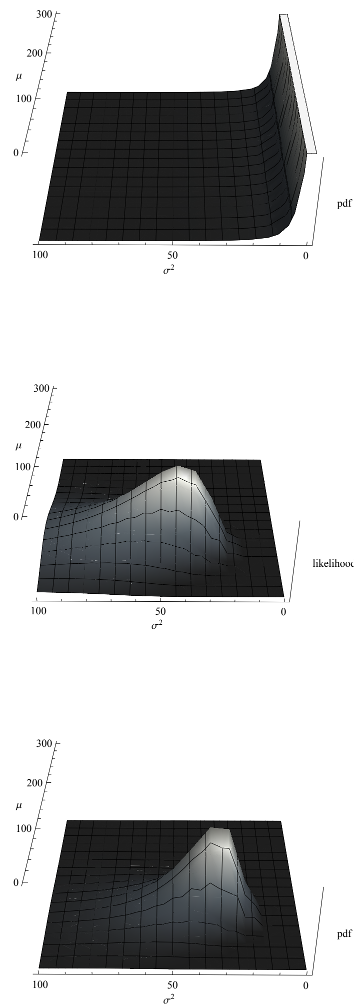


Figure 7.5: The prior, likelihood, and posterior distributions for the mean and variance of IQ example described in section 7.4.2.

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{IQ}) &= \frac{p(\boldsymbol{IQ}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\boldsymbol{IQ})} \\
&= \frac{p(\boldsymbol{IQ}|\mu,\sigma^2) \times p(\mu,\sigma^2)}{p(\boldsymbol{IQ})} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\sum_{i=1}^{3}(IQ_i-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma^2}}{\int_0^{300}\int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\sum_{i=1}^{3}(IQ_i-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma^2} \mathrm{d}\sigma^2 \mathrm{d}\mu} \\
&\propto \sigma^{-3} exp\left(-\frac{\sum_{i=1}^{3}(IQ_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}
\tag{7.23}
$$

In (7.23), the second line was obtained from the first by simply substituting in for $\boldsymbol{\theta} = (\mu, \sigma^2)$. We then substituted for the likelihood[9] and prior from (7.20) and (7.22) respectively.

We can choose, in this rather simplified example, to go through and actually evaluate the posterior exactly, by calculating the denominator by brute force:

$$
\begin{aligned}
p(IQ) &= \int_0^{300}\int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{\sum_{i=1}^{3}(IQ_i-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma^2} \mathrm{d}\sigma^2 \mathrm{d}\mu \\
&= \frac{e^{-\frac{2500}{\sigma^2}}\left(\mathrm{erf}\left(\frac{50\sqrt{6}}{\sigma}\right) + \mathrm{erf}\left(\frac{100\sqrt{6}}{\sigma}\right)\right)}{4\sqrt{3}\pi\sigma^4} \\
&\approx 3.2 \times 10^{-7}
\end{aligned}
\tag{7.24}
$$

Do not worry if you find the above calculation hard to follow! We have included it to emphasise the difficulty which calculating this term can pose even for relatively simple examples. We carried out this calculation using Mathematica's symbolic calculation engine, but would not want to attempt it by hand.

[9] We have assumed independence for the data, meaning that to get the overall likelihood,

Using the above denominator results in a posterior density shown in figure 7.5. We can then obtain the marginal densities by integrating out any dependence of the parameter not in interest:

$$p(\mu|\boldsymbol{IQ}) = \int_{0}^{\infty} p(\mu, \sigma^2|\boldsymbol{IQ})\mathrm{d}\sigma^2 \tag{7.25}$$

$$p(\sigma^2|\boldsymbol{IQ}) = \int_{0}^{300} p(\mu, \sigma^2|\boldsymbol{IQ})\mathrm{d}\mu \tag{7.26}$$

$$\tag{7.27}$$

Although, here we could go through and analytically derive these posteri-

---

we multiply together the three individual likelihoods.

ors[10], it is hoped that this example gives a little insight into the complexity of calculating the denominator in Bayesian models. The degree of difficulty in calculating the denominator increases rapidly in the number of unknown parameters within a model. In fact, at some point, the denominator becomes practically infeasible to calculate for models more complicated than only a few parameters.

However, all is not lost, as we discuss in section 7.5.

7.5 How to dispense with the difficulty: Bayesian computation

The Herculean task of calculating the denominator for continuous parameters would seem to put a real spanner in the works for Bayesian statistics, such its reliance on the denominator of Bayes' rule. However, all is not lost. There are two solutions to the difficulty:

- Use priors conjugate to the likelihood (See chapter 9).
- Abandon analyticity, and opt to sample from the posterior instead.

The first of these workarounds still allows for exact derivation of an expression for the posterior distribution, by choosing a mathematically *nice* form for the prior distribution. This simplifies the analysis, since one can simply look up formulae for the posterior which have already been tabulated for us, avoiding to have to do any maths at all. However, frequently in real life applications of Bayesian statistics, we need to stray outside this realm of mathematical convenience. The price for a more varied choice of priors and likelihoods is that we have to give up our aspirations for closed-form calculation of the posterior density. However, it turns out in these circumstances we can still *exactly* sample from the posterior, and then use sample summary statistics to describe the posterior distribution in a very adequate way. We will leave a full description of these computational methods to part IV, but to provide a clue as to where we may be heading, we note that the posterior density can be written:

[10] Although we have chosen to omit the exact closed-form results here for brevity. Post-

$$\begin{aligned} p(\theta|data) &= \frac{p(data|\theta) \times p(\theta)}{p(data)} \\ &\propto p(data|\theta) \times p(\theta) \end{aligned} \tag{7.28}$$

In (7.28) we have arrived at the second line due to $p(data)$ being independent of θ; it is essentially a constant that we use to normalise the posterior. The numerator of Bayes' rule tells us everything that we need to know about

poning such a full derivation until part III.

the *shape*[11] of the posterior distribution, whereas the denominator merely tells us about its *height*. Computational methods use the shape of the posterior distribution to generate samples from it based on local comparison of relative probabilities.

Furthermore, the use of sampling allows one to bypass the difficulty with calculating some of the most important summary measures of parameter estimates. Even if we were able to calculate the denominator, we usually want to use the joint posterior to calculate marginal quantities of interest. For example, we are often interested in the expected value of a parameter:

$$
\begin{aligned}
\mathrm{E}[\theta_1] &= \int\limits_{\Theta_1} \theta_1 \int\limits_{\Theta_2} p(\theta_1, \boldsymbol{\theta_2}|data)\mathrm{d}\boldsymbol{\theta_2}\mathrm{d}\theta_1 \\
&= \int\limits_{\Theta_1} \theta_1 p(\theta_1|data)\mathrm{d}\theta_1
\end{aligned}
\tag{7.29}
$$

In the above, we have got to the second line by marginalising out the parameter vector $\boldsymbol{\theta_2}$ (see section 3.3.5 for further explanation). As the difficulty with calculating the denominator proves, these highly-dimensional integrals can be practically intractable. In Part IV we will see how sampling again comes to our rescue.

7.6 Chapter summary

This chapter should have introduced you to the different interpretations of the denominator of Bayes' rule; firstly as a nuisance normalising factor; secondly as a probability of the data. We then discussed the difficulty that the denominator poses for Bayesians. Fortunately, this problem can be sidestepped via use of conjugate priors (see chapter 9), although this is often extremely limiting for all but the most simple of data processes. We then reasoned that since the curvature of the posterior is solely determined by the numerator of Bayes' rule - the prior multiplied by the likelihood - we can learn much of the posterior from this quantity. In particular modern computational methods use the numerator of Bayes' rule, to generate samples from the posterior distribution. These can then be used to summarise

[11] It's dependence on θ.

the posterior distribution, as well as for any of the uses of posterior distributions described in chapter 4. Before we introduce computational methods in full in part IV, we firstly must understand the multitude of distributions which we have in our arsenal. We then discuss how conjugate priors can be used to analytically derive posteriors, which can be useful before moving to approximate computational methods.

7.7 Chapter outcomes

The reader should now be familiar with the following concepts:

1. The denominator as a normalising factor.
2. The denominator as a probability.
3. The difficulty with computing the denominator.
4. How computational methods dispense with the difficulty.

7.8 Problem set

7.8.1 New disease cases

As an intern at the WHO you have been tasked with developing a statistical model of the number of outbreaks of new diseases throughout the world. You have data on the number of outbreaks that have occurred in the month of June for the past 10 years.

What likelihood model might be appropriate here?

What are the assumptions of this model? Are they appropriate here?

A gamma prior

A colleague of yours suggests that a gamma prior may be appropriate to model the mean number of cases occurring within each month, λ. The gamma distribution PDF they suggest has the following parameterisation:

$$p(\lambda|\alpha,\beta) = \frac{\lambda^{\alpha-1}\beta^{\alpha}e^{-\beta\lambda}}{\Gamma(\alpha)} \tag{7.30}$$

Assuming your likelihood and the above prior, find $p(data)$. In other words the prior predictive distribution.

Data

Suppose that the data you have for the past 10 years is: {2,1,0,5,2,1,1,7,3,2}.

Using the aforementioned model along with a $Gamma(2,1)$ prior to find $p(data)$.

Find the posterior distribution for the mean parameter λ

7.8.2 The comorbidity between depression, anxiety and psychosis

Suppose that you are a medical researcher interested in the coincidence of depression, anxiety and psychosis. You have collected historical data for patients who were admitted to your clinic, and record it in tabular form shown in table 7.3.

Calculate the probability that a patient is depressed.

Calculate the probability that a patient is psychotic.

What is the probability that a patient is psychotic given that they are depressed, and anxious?

What is the probability that a patient is not psychotic if they are not depressed?

7.8.3 Finding mosquito larvae after rain

It is important for those combating malaria to understand the factors that affect mosquito larvae development. Suppose that you work as part of a

	Not psychotic		Psychotic	
	Depression		**Depression**	
Anxiety	0	1	0	1
0	0.32	0.16	0.04	0.04
1	0.08	0.24	0.04	0.08

Table 7.3: The joint distribution representing historical patient diagnoses upon entering the clinic.

field team, and collect data on the position of larval sites within a lake, after a rainfall. Specifically, you estimate the distance of these sites away from the bank.

Suppose experience tells you that a lognormal likelihood function is sensible for the distance of the larval sites away from the bank. Further suppose that the scale parameter of this distribution is known to be 1, leaving the first parameter of this distribution μ to be estimated.

Find the mean of a $lognormal(d, 1)$ distribution.

Finding the mode

Suppose that further, you choose to use a $Gamma(2, 1)$ distribution for the parameter μ. You collect the following data for distances from the most recent rainfall {3.3, 1, 0.5, 0.4, 3.3}.

Using computational simulation, find the position of the mode.

Use a computer to graph the shape of the posterior distribution.

Calculate the posterior (difficult).

7.9 Appendix

Part III

Analytic Bayesian methods

7.10 Part mission statement

7.11 Part goals

Chapter 8

An introduction to distributions for the mathematically-un-inclined

8.1 Chapter mission statement

At the end of this chapter the reader should be familiar with the commonplace likelihood and prior distributions used in a significant proportion of applied analyses.

8.2 Chapter goals

Often texts on Bayesian analysis assume a familiarity with probability distributions that evade all but the statistics fanatics. This presumed knowledge often makes it difficult to penetrate the literature, and makes Bayesian inference appear more difficult than it actually is. The idea behind this chapter is to redress the balance; by going through the commonly-used distributions, highlighting their interrelationships, and their use for a wide range of example situations. It is hoped that by addressing their practical applicability, that we can turn an oft-dry subject into something more palatable for bedtime reading.

Generally, there are two classes of distributions: those that can be used for likelihoods, and those which are commonly used for priors. There is considerable overlap between these groupings (particularly for continuous distributions), meaning that we will often encounter the distributions twice. It is important to note that

Throughout the chapter there are considerable interactive elements. It is hoped that you will take the opportunity to use this material to gain hands-on experience. If the text becomes dense in places, there are videos which take you through some of the basic properties and uses of the distributions.

It should be stressed that this chapter, perhaps less than some of the others needn't necessarily be read in its entirety. Whilst it is recommended that you read all of section 8.3, it is not necessary to read the entirety of section 8.5. Whilst there is much to be gained from persevering, to avoid boredom, it may be better to refer to the relevant section in the event that you come across a distribution which is hitherto unknown to you.

8.3 Sampling distributions for likelihoods

The first class of distributions which we came across in Part II were those that were used to write down a probability model to describe the situation under inspection. As well as expressing our assumptions used to simplify the system, these *likelihood* models also define the parameters over which we are interested in conducting inference. Specifying a likelihood that is adequate to the specifics of the system under consideration is the most important part of conducting Bayesian inference. As such, it is essential to have a reasonable grasp of some of the most common likelihood distributions on which many analyses are built. Whilst there are an infinity of potential sampling distributions out there, we believe that the collection here are the most frequently used, and are often a starting point for the construction of more complex models.

8.3.1 Bernoulli

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Discrete data.
2. A single trial.
3. *Only* two trial outcomes: *success* and *failure*. These do not need to literally represent successes and failures, but this shorthand is typically used, and we adopt it here.

Example uses: outcome of flipping a coin, a single clinical trial, or a presidential election!

Imagine you are interested in the outcome of a single horse race. For added simplicity, we suppose that we only care whether the horse wins, or loses; forgetting about its position if it doesn't win. In this framework, we formalise the outcomes of the race by creating a mathematical object - a *random variable* - which associates a numerical value with each of the outcomes, X:

$$X = \begin{cases} O & ,\textit{horse wins} \\ 1 & ,\textit{horse loses} \end{cases} \tag{8.1}$$

In this set-up it makes sense to model the outcome of a single race as being influenced by a background probability of success, $0 \leq \theta \leq 1$. We don't actually witness this probability, and after the discussion in chapter 2, we aren't sure it *really* exists. In the Frequentist paradigm θ represents the proportion of races that the horse has won historically (and will win in total if we were to continue re-racing forever). Whereas to Bayesians, θ merely gauges our subjective confidence in the event of the horse winning (see section 2.5).

Now that we have a model, we can work out the probability of the two distinct possible outcomes. The probability that the horse wins is straightforwardly $p(win) = \theta$; meaning the probability of a loss *must*[1] be given by $p(loss) = 1 - \theta$.

We now are in a position to work out the likelihood of a given outcome. Remember that a likelihood is *not* a valid probability density, and is found by holding the *outcome*, or in other words *data*, constant, whilst varying

[1] It *must* be given by $1 - \theta$, so that the Bernoulli distribution sums to 1, and hence is a valid probability density.

the parameters. Suppose that the horse has had a good meal of carrots this morning, and goes on to win by a country mile. We know that the likelihood of this event is given by $l(\theta|X=1) = p(X=1|\theta) = \theta$ (see the red line in figure 8.1).

Alternatively, if our horse spent the night in the midst of night-mares, and went on to lose, then the likelihood is given by $l(\theta|X=1) = 1-\theta$ (see the blue line in figure 8.1).

As you can see, in this situation, the maximum likelihood estimates in each case are given by $\hat{\theta} = 1$, and $\hat{\theta} = 0$, if the horse wins or loses respectively.

We can actually write down a single expression that yields the likelihood/probabilities[2] for both potentialities:

$$p(X|\theta) = \theta^X (1-\theta)^{1-X} \tag{8.2}$$

This distribution is known as the *Bernoulli*, after the Swiss mathematician Jacob Bernoulli who first probed this much-discussed distribution.

$\boxed{\textit{Properties}:}$

$$\boxed{\begin{aligned} \textit{notation} &: X \sim Ber(\theta) && (8.3a) \\ \textit{pdf} &: p(X|\theta) = \theta^X(1-\theta)^{1-X} && (8.3b) \\ \textit{mean} &: \mathrm{E}[X] = \theta && (8.3c) \\ \textit{variance} &: var[X] = \theta(1-\theta) && (8.3d) \end{aligned}}$$

It is rare in life to come up against one-off events, and we typically have data for a number of races historically. However, this discussion has given us the necessary grounding to make the step to the next distribution up the rung, the *binomial*, which is perfect for this more realistic setting.

$\boxed{\textit{Video}:}$ Need to make an intuitive video here. One exists but is overly mathematical.

8.3.2 Binomial

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have**

[2] Dependent on whether we vary the parameter or data respectively.

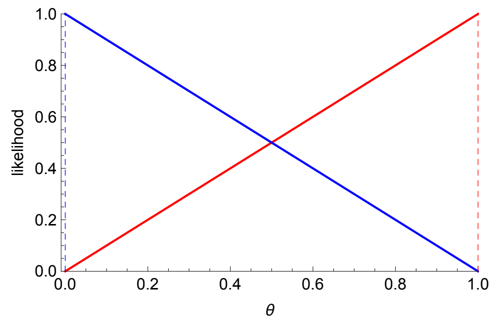


Figure 8.1: Bernoulli likelihoods for the event that the horse wins (red), and loses (blue). The maximum likelihood estimates are shown as dotted lines for each of the cases.

ticks next to these points.)

1. Discrete data.
2. Multiple trials.
3. Each trial has only two outcomes (see the Bernoulli): we for lack of better words, deem these *successes* and *failures*.
4. Individual outcome probabilities are not determined by any other factor that varies across individuals for which we have data.
5. Probability of success is same in each trial.
6. Overall data we measure is the aggregate number of successes.
7. Trials are independent[3].

Example uses: clinical drug trials, Democrat voters exiting a poll station, or the number of mosquito bites over the course of a day.

Now we jump immediately to a much more practically-relevant case where we have data on the outcome of a number of horse races, Republican votes, or in this case, we shall imagine we are examining the outcome of a clinical trial of a new flu drug. We have a sample of 10 willing students, who have chosen to take the pain of a sweaty week in pyjamas, for the betterment

[3]See section 3.5.1

of humanity. At the week's start, the students are infected with the virus via an injection. At the end of the week, the consulting physician records the number of volunteers that still show flu symptoms, and those that do not. In order to build a model, we need some assumptions. Firstly, we assume that the students' data are exchangeable[4]. This might be violated if we knew that some of the volunteers have asthma, and are likely not to be so rapidly cured by the wonder drug. We also create a random variable X which represents the outcome of the trial for a single volunteer, which takes the value 1 if the trial is successful, in other words the volunteer is no longer symptomatic, and 0 if not.

However, we have data on the success of the trials for all 10 of our volunteers. Here, it makes sense to create another helper random variable, $0 \leq Z \leq 10$, which measures the aggregate outcome of the trial:

$$Z = \sum_{i=1}^{10} X_i \tag{8.4}$$

Carrying out the trial we find that 5 volunteers successfully recovered from the virus of the week. We reason that the following outcomes *could* have lead to this aggregate result: $\boldsymbol{X} = \{1, 1, 1, 1, 1, 0, 0, 0, 0, 0\}$; meaning the arbitrarily-chosen first 5 volunteers happened to react well to the treatment. Feeling satisfied we present our results to the pharma company executives who developed the drug in the first place. They look slightly perturbed, and say that $\boldsymbol{X} = \{1, 0, 1, 0, 1, 0, 1, 0, 1, 0\}$ would also have been possible. Realising our mistake, we counter with the argument that it is also possible that the latter 5 volunteers were the ones who recovered. This back-and-forth continues, until you realise that there is a unifying mathematical formula

[4] In this example, this really means that the students constitute a *random sample*; meaning the data are independent and identically-distributed.

that can cover any situation - the binomial nCr formula[5]. You reason that there are possibly $\binom{10}{5} = 252$ possible overall combinations of results with the same aggregate outcome!

With this realisation you make the next step; writing down the likelihood. Since we have supposed that the observations are *independent*, we are able to calculate the overall probability by multiplying together the individual probabilities, taking into account the 252 possible combinations:

$$\begin{aligned} p(Z = 5|\theta) &= 252 \times p(X_1 = 1|\theta)...p(X_5 = 1|\theta)p(X_6 = 0|\theta)...p(X_10 = 0|\theta) \\ &= 252 \times \theta^5(1-\theta)^5 \end{aligned} \tag{8.5}$$

We can graph this likelihood as a function of θ (see the blue line of figure 8.2), and note that the maximum likelihood estimator of the parameter occurs at the intuitive result of $\theta = \frac{1}{2}$.

If we were less lucky and only found 2 patients recovered, then we see that the likelihood is shifted leftwards; peaking now at $\theta = \frac{1}{5}$ (see the red line of figure 8.2). By contrast if our patients respond well, and 9 out of 10 recover in the week period, then the likelihood shifts right (see the grey line of figure 8.2).

As for the Bernoulli case, we would like a compact way of writing down the likelihood to cover any eventuality. We now suppose that the number of volunteers is given by n, the probability of individual treatment success is θ, and k cases turn out to be successes:

$$p(Z = k|n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{8.6}$$

This distribution is known as the *Binomial* distribution.

If we hold n constant, and increase the probability of success θ, we see that the probability distribution of data shifts to the right (see the right-hand panel of figure 8.2).

[5]See section 5.6.2 for a more complete discussion.

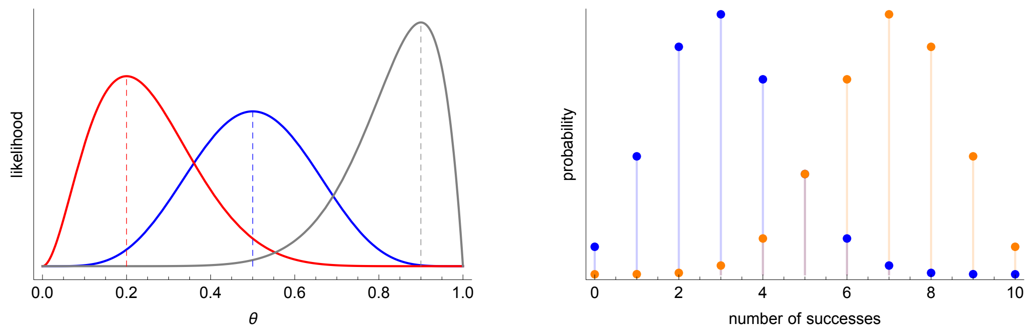


Figure 8.2: **Left**: Binomial likelihoods for the event that 5 (blue), 2 (red), and 9 (grey) volunteers recovered in the week period. The maximum likelihood estimates are shown as dotted lines for each of the cases. **Right:** the probability distribution of successful trials if $\theta = 0.3$ (blue) and $\theta = 0.7$ (orange).

$\boxed{Properties:}$

$$\boxed{\begin{aligned} notation &: X \sim B(n, \theta) && (8.7a)\\ pdf &: p(X|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} && (8.7b)\\ mean &: \mathrm{E}[X] = n\theta && (8.7c)\\ variance &: var[X] = n\theta(1-\theta) && (8.7d)\end{aligned}}$$

$\boxed{Video:}$ Need to make an intuitive video here. One exists but is overly mathematical. $\boxed{Interactive:}$ again make.

8.3.3 Poisson

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Count of discrete events.
2. Individual events occur at a given rate, and independently of other events.

3. Fixed amount of time or space, in which the events can occur.

Example uses: estimating the failure rate of artificial heart valves, the first appearance of crop disease in a large farming area, approximating the binomial for the percentage of the UK population who suffer from autism.

Relationships with other distributions: approximates *binomial*(n, θ) if n is large, and θ is small. **It would be good to add a graph showing interrelations of the various distributions.**

Suppose that you are interested in estimating the rate at which new outbreaks of Legionella disease[6] occur worldwide. Public health officials have collected data of the count of independent disease outbreaks that occur each year, X. In this case, a reasonable model would be to assume that the outbreaks appear spontaneously, and at a mean rate λ which is independent of any systematic macro factors for which we might have data. In this case, a sampling distribution of the form:

$$p(X = k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{8.8}$$

The distribution is called a *poisson* model, after a prominent French mathematician born in the 18th century.

The likelihood is found by varying the mean rate of occurrence, λ, holding the data sample constant. The likelihood is shown for three data samples with $\bar{X} = 0.5, 3, 8$ respectively on the left hand side of figure 8.3.

This distribution is only defined for non-negative integer X data, which makes it perfect for modelling the count of occurrences. However, the mean rate λ is not constrained to the integers; being able to take on any non-negative real value. It is important to stress that this distribution is only valid if the individual events being counted occur *independently* of each other. In our Legionella example, this would be violated if there was an issue with the data collection, meaning that some of the outbreaks were not in fact new, and were actually caused by contagion from existing outbreaks.

The distribution is shown for different values of λ on the right hand side of figure 8.3

[6] A nasty disease carried by bacteria that thrive in warm water, named for its first reported occurrence among conference attendees at a Philadelphia convention of the American Le-

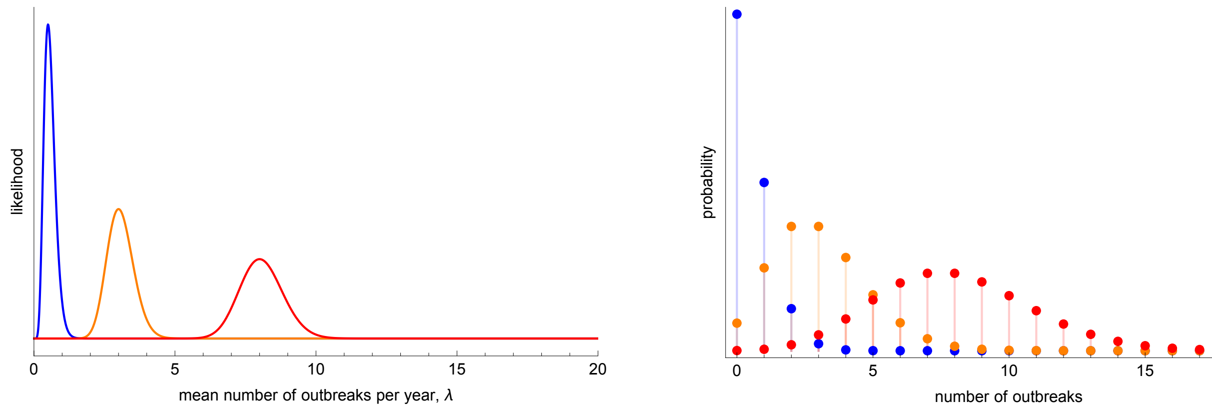


Figure 8.3: **Left**: The poisson likelihood for three different samples. **Right**: the poisson distribution for different values of the mean rate. **Add legends: left xbar=0.5,3,8. Right: lambda=0.5,3,8.**

Notice in the below listed properties, that the mean and variance of this distribution are the same, and equal to λ. This restriction hence limits the use of this distribution to cases where the data sample satisfies these properties, at least approximately. However, we will see next that the addition of an extra parameter, which governs the degree of over-dispersion in the data, can allow for sufficient flexibility to handle data with a variance greater than the mean.

$\boxed{Properties:}$

$$notation: X \sim poisson(\lambda) \tag{8.9a}$$

$$pdf: p(X = k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{8.9b}$$

$$mean: \mathrm{E}[X] = \lambda \tag{8.9c}$$

$$variance: var[X] = \lambda \tag{8.9d}$$

$\boxed{Video:}$ Need to make an intuitive video here. $\boxed{Interactive:}$ Need to make an interactive element here.

gion.

8.3.4 Negative binomial

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Count of discrete events.
2. Non-independent events. It is sometimes said that the events can exhibit contagion; meaning that if one event occurs, it is more likely that another will do.
3. A variance that exceeds the mean.
4. Fixed amount of time or space, in which the events can occur.

Example uses: everything the poisson can do and more, numbers of measles cases on an island, collapses of banks in a financial crisis.

In real life it is not uncommon for us to want to model the count of events which occur in clumps; either in time or space. This 'clumping' behaviour indicates that the occurrence of one event makes it more likely that another will occur afterwards.

Again we imagine we are in the role of epidemiologists, this time modelling the occurrence of influenza over a winter season in three small villages: Socialville, Commuterville, and Hermitville, all of the same size. To keep things simple, we imagine that in each village, the magnitude of the outbreak is determined by a parameter θ. If this number were known, we assume that the numbers of flu cases, X is given by a poisson distribution:

$$X|\theta \sim Pois(\lambda\theta) \tag{8.10}$$

θ can be thought of as measuring the type and degree of social interactions of the villagers. In Studentville, $\theta = 1.35$ meaning that each person has a lot of interaction with fellow villagers, and the outsider world. In Commuterville, $\theta = 1$ meaning that each person has no contact with their fellow villagers, but have a large number of contact with the outside world. In Academicville, $\theta = 0.65$ meaning only a few villagers have contact with the outside world, and again none within the village. Across all villages, it is

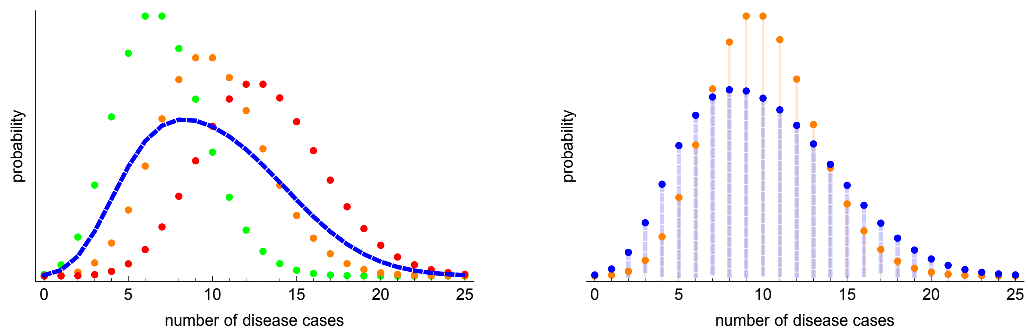


Figure 8.4: **Left**: the overall distribution (blue, dashed) is essentially an average across those for each of the towns: Studentville (red), Commuterville (orange), and Academicville (green). **Right**: a comparison of the overall distribution (blue, dashed), with a poisson distribution (orange) of the same mean.

straightforward (see below) to show that the mean number of cases is λ. Here we take $\lambda = 10$.

Clearly, in this case we will on average get the most cases of the flu in Studentville, where people are likely to contract the disease, then pass it on to others. In Commuterville, where all the people have no contact with one another, then the number of flu cases is governed by how many people *independently* get sick. Finally in Academicville, only some of the people will likely get sick since some people are complete hermits! And, those that do get sick, will again not pass on the disease to their fellow villagers. We can think about building a distribution which allows for disease case numbers across all three villages, by essentially averaging out across all the distributions (see the left hand side of figure 8.4)

If we want to build a model for the disease cases across all three villages, we should expect that even though the overall mean is 10, that the distribution will be wider than that of a simple poisson distribution with the same mean (see the right hand side of figure 8.4).

The above example illustrates how contagion can result in a higher number of cases than would be expected if the flu cases were independent. We also see that θ captures this degree of contagion. If we suppose that we are now interested in predicting the number of flu cases across a large number of equally sized villages. We assume that we have no data on the connectivity of the various villages, and that it is reasonable to assume that $\theta \sim \Gamma(\kappa, \frac{1}{\kappa})$[7], meaning that $E[\theta] = 1$ across all villages, meaning that unconditionally, the above distribution has the property that the mean number of flu cases is given by:

$$\begin{aligned} E[X] &= E\left[E[X|theta]\right] \\ &= E[\theta]E[\lambda] \\ &= \lambda \end{aligned} \tag{8.11}$$

[7] Wait until section 8.5.3 for a full explanation of this distribution.

where to obtain the right-hand-side of the first line, we have used the law of iterated expectations[8].

Here because of the inherent variance in the connectivity across villages, we will get a greater variability of results than would be predicted, if all villages were equally connected. In this latter case we can suppose that $\theta = 1$, which means that all villages would have a $Pois(\lambda)$ number of cases. In the former, we can calculate the variance in flu cases:

$$\begin{aligned} var\left[X|\,\lambda,\kappa\right] &= \lambda + \lambda^2\kappa \\ &\geq var(X_{Poisson}) \\ &= \lambda \end{aligned} \tag{8.12}$$

Do not worry about the provenance of this expression, that is not what is important. The only thing that matters here, is that because we have allowed variance in the parameter θ, which governs the connectivity in each village, we see that we have a wider variance than we would have got if all the villages were the same as Commuterville.

More generally, we see that this distribution allows for a variance that is greater than that of the Poisson distribution. However this extra flexibility comes at a cost of extra complexity, because it is characterised by two parameters rather than one. The parameter κ measures the degree of dispersion of the data; the larger its value, the greater the variability (see figure 8.5). This greater variance of this distribution has lead some to categorise it as an *over-dispersed* Poisson. However, its more common name is the *negative binomial*.

This distribution has a number of commonly used parameterisations, each of which we illustrate below.

Properties :

[8] Do not worry if you don't know what this is. It is not important, but I have a video here `https://www.youtube.com/watch?v=Ki2HpTCPwhM` which explains it in full.

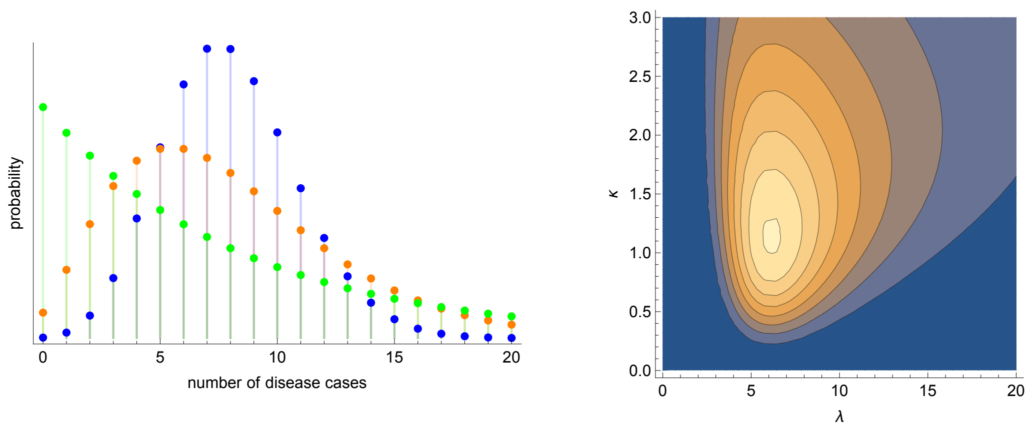


Figure 8.5: **Left**: The negative binomial distribution for low dispersion (blue), medium dispersion (orange) and high dispersion (green). All three cases have a mean of 8. **Right**: a contour plot of the likelihood surface for the data sample: 0, 5, 0, 5, 8, 10, 15. The parameter κ here represents the dispersion, with smaller values indicating less dispersion. **Add legends**

Here κ represents the dispersion.

$$\boxed{\begin{aligned} notation &: X \sim NB(\lambda, \kappa) && (8.13a)\\ pdf &: p(X=y|\lambda,\kappa) = \frac{\Gamma(y+\frac{1}{\kappa})}{y!\Gamma(\frac{1}{\kappa}+1)}\left(\frac{\lambda}{\lambda+\frac{1}{\kappa}}\right)^{y}\left(\frac{\frac{1}{\kappa}}{\lambda+\frac{1}{\kappa}}\right)^{\frac{1}{\kappa}} && (8.13b)\\ mean &: \mathrm{E}[X] = \lambda && (8.13c)\\ variance &: var[X] = \lambda + \lambda^2\kappa && (8.13d) \end{aligned}}$$

Here κ represents the inverse of the dispersion.

$$\boxed{\begin{aligned} notation &: X \sim NB(\lambda, \kappa) && (8.14a)\\ pdf &: p(X=y|\lambda,\kappa) = \frac{\Gamma(y+\kappa)}{y!\Gamma(\kappa+1)}\left(\frac{\lambda}{\lambda+\kappa}\right)^{y}\left(\frac{\kappa}{\lambda+\kappa}\right)^{\kappa} && (8.14b)\\ mean &: \mathrm{E}[X] = \lambda && (8.14c)\\ variance &: var[X] = \lambda + \frac{\lambda^2}{\kappa} && (8.14d) \end{aligned}}$$

This final parameterisation is probably the most common. Here we suppose we have a number of Bernoulli trials; each of which can be a failure or success. In this case, r represents the number of failures before we stop the experiment, and θ is the probability of success in each experiment.

$$notation : X \sim NB(r, \theta) \tag{8.15a}$$

$$pdf : p(X = y|r, \theta) = \frac{\Gamma(y + r)}{y!\Gamma(r)}(1 - \theta)^r \theta^y \tag{8.15b}$$

$$mean : \mathrm{E}[X] = \frac{r\theta}{1 - \theta} \tag{8.15c}$$

$$variance : var[X] = \frac{r\theta}{(1 - \theta)^2} \tag{8.15d}$$

Be careful, Γ in the above distributions is the gamma function, *not* the gamma distribution.

It would be good to combine each of these into a single table.

$\boxed{\textit{Video} :}$ Need to make an intuitive video here. $\boxed{\textit{Interactive} :}$ Need to make an interactive element here.

8.3.5 Beta-binomial distribution

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Discrete data.
2. Multiple trials.
3. Each trial has only two outcomes (see the Bernoulli): we for lack of better words, deem these *successes* and *failures*.
4. Individual outcome probabilities are not determined by any other factor that varies across individuals for which we have data.
5. Overall data we measure is the aggregate number of successes.
6. Probability of success varies across studies/groups/time.

Example uses: everything the binomial can do and more, breast cancer prevalence across heterogeneous patient groups, number of voters for the Republican party across counties.

This distribution is to the binomial what the negative binomial is the poisson. By assuming that there is heterogeneity in the probability of success, this leads to a variance that exceeds that of the basic binomial. This has lead some to say this is another example of a *over-dispersed* distribution. As the following example illustrates, this effect is common when there is heterogeneity across different groups.

Imagine that you are interested in quantifying the efficacy of a particular drug, which aims to cure depression We conduct a number of separate trials, in each case measuring the number of recoveries out of the total number of participants. Here we only consider formulating a model for the group who are actually treated. For one group we might imagine that the number of successes, X, can be modelled by a binomial distribution: $X \sim B(n, \theta)$. Here n represents the number of individuals within that sample, and θ is the success probability for an individual patient.

We now imagine extending the analysis to cover two patient groups, each of size 10 people indexed by their depression levels: *mild* and *severe*. We suppose that the drug is most efficacious for the mild group, meaning that the corresponding success probability $\theta_{mild} > \theta_{severe}$. Here we expect the number of recoveries in the mild group to exceed that of the severe group. If we were to build a model to allow for results that encompass both groups, we could think about the distribution as representing a sort of average of the other two (see the left hand side of figure 8.6). We also think that due to this extra variability, we expect that across both samples we would likely see a greater degree of variability than we would have if the groups had the same trial success probability (see the right hand side of figure 8.6)[9].

More generally, if we consider a large number of such groups, and suppose that the success probability in each group is drawn from an overarching distribution, $\theta \sim Beta(\alpha, \beta)$,.then the number of trial successes in a given study is given by a *beta-binomial* distribution, with properties shown below. The variance of this distribution can be shown to exceed that of the binomial distribution (see problem set at end of chapter), meaning that this is a useful distribution for when the variance in your data appears to be large between studies.

[9] See problem set at end of chapter

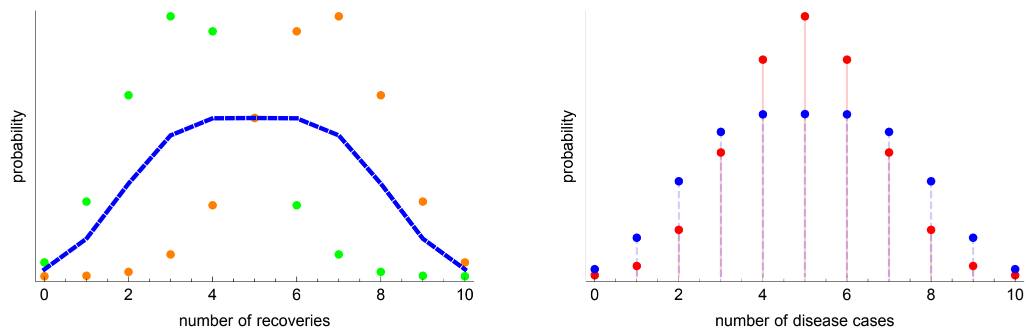


Figure 8.6: **Left:** a distribution that encompasses outcomes from both groups can be thought of as an average of the *mild* and *severe* groups. **Right:** a comparison of the resultant distribution with a binomial distribution with the same mean. **Add legends**

See figure 8.7 for a graphical depiction of this distribution in response to changes in its parameters, as well as its likelihood surface for a sample of six different studies, with numbers of trial successes given by: 1, 10, 8, 3, 8, 9.

$$pdf : p(X = y|n, \alpha, \beta) = \binom{n}{y} \frac{BetaFn(y + \alpha, n + y - \beta)}{BetaFn(\alpha, \beta)} \tag{8.16a}$$

$$mean : \mathrm{E}[X] = \frac{n\alpha}{\alpha + \beta} \tag{8.16b}$$

$$variance : var[X] = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{8.16c}$$

Here *BetaFn* means the beta function, rather than the beta distribution.

$\boxed{Video:}$ Need to make an intuitive video here. One exists but is overly mathematical. $\boxed{Interactive:}$ again make.

8.3.6 Normal

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

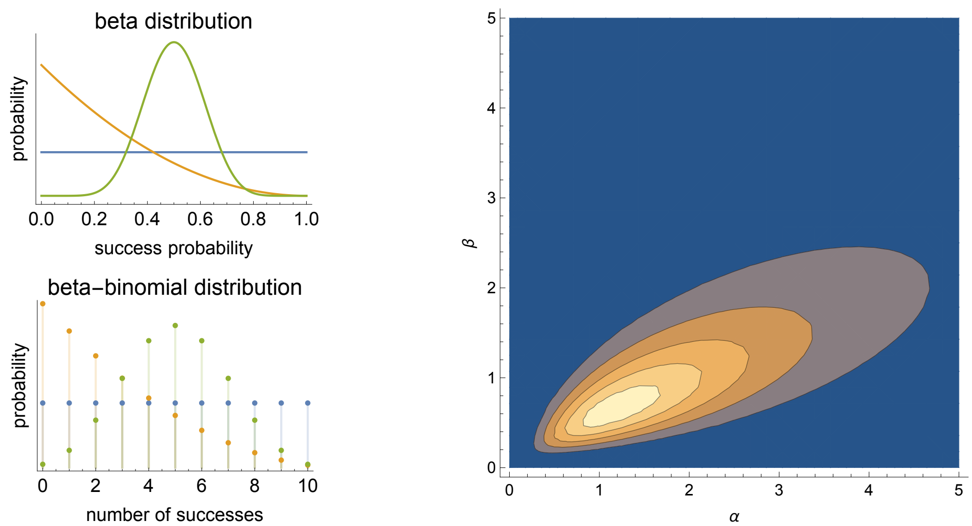


Figure 8.7: **Left, uppermost panel:** the beta distribution for three different sets of parameter values. Blue (1,1), Orange (1,3), Green (10,10). **Left, lowermost panel:** the beta-binomial distribution for the same parameter values, out of a sample size of 10. **Right**: likelihood contour plot for a sample of size studies 1, 10, 8, 3, 8, 9. Again the sample size is 10 in all cases. **Add legends**

1. Continuous data.

2. Unbounded outcomes; at least *practically* unbounded. For example weight of adults aged 30. This is bounded by zero, since weight cannot be negative, but this bound is never approached by data.

3. Outcome can be considered to be the outcome of a large number of additive factors.

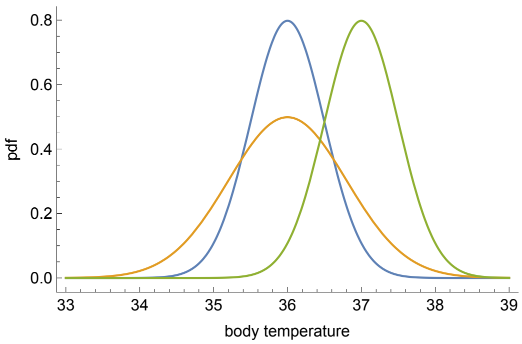
Example uses: error in regression models, the intelligence of individuals, the prevalence of genes within a population, an approximation to the binomial distribution when the sample size is large and the probability of success is close to $\frac{1}{2}$, an approximation to the poisson distribution when its mean is large.

The normal is the most commonly occurring distribution in nature. This is a result of the various central limit theorems (see section 3.6) which predict its occurrence whenever there are a number of additive factors which result in an outcome. These factors needn't necessarily be independent, although this often helps.

As an example, imagine that we are interested in explaining the distribution of body temperatures of people. In this example, we might imagine that measurements of body temperature depends on a number of factors: the amount of exercise taken before the appointment, the outside temperature, how much water the person has drunk, as well as genetic and other physiological characteristics. When we have no information on these multitude of factors, we might invoke the use of a central limit theorem, and suppose that the normal distribution might be a reasonable choice.

This distribution is specified by two parameters: μ, its mean; and σ its standard deviation. Changing these has the expected outcome of translating, and widening/narrowing of the probability density (remember that since the distribution is continuous, we are dealing with a density - see section 3.3.2). Figure 8.8 shows a plot of these densities for different sets of its mean and standard deviation.

Suppose we obtain the following readings for the body temperature across 10 different participants: 36.4, 37.2, 35.8, 36.4, 37.1, 35.6, 36.4, 37.6, 37.5, 37.1. Now that we have a distribution specified by two unknown parameters, the likelihood function is actually a surface, and it is easiest to understand by looking at contour plots (see figure 8.8).



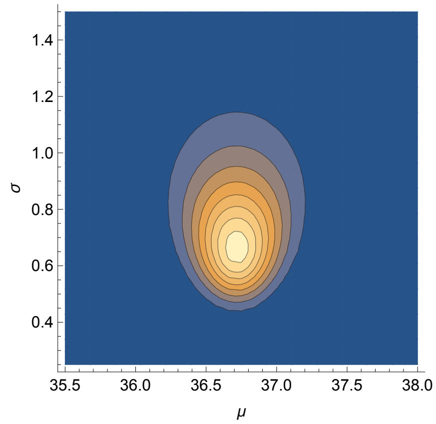


Figure 8.8: **Left**: the normal probability density functions for three different sets of (μ, σ): (36, 0.5), (36, 0.8), (37, 0.5). **Right:** a contour plot of the normal likelihood for a sample of body temperatures: 36.4, 37.2, 35.8, 36.4, 37.1, 35.6, 36.4, 37.6, 37.5, 37.1.

$\boxed{Properties:}$

$$notation : X \sim N(\mu, \sigma^2) \tag{8.17a}$$

$$pdf : f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8.17b}$$

$$mean : \mathrm{E}[X] = \mu \tag{8.17c}$$

$$variance : var[X] = \sigma^2 \tag{8.17d}$$

$\boxed{Video:}$ Need to make an intuitive video here. One exists but is overly mathematical. $\boxed{Interactive:}$ again make.

8.3.7 Student t

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous data.

2. Unbounded outcomes; at least *practically* unbounded. For example weight of adults aged 30. This is bounded by zero, since weight cannot be negative, but this bound is never approached by data.

3. Outcome can be considered to be the outcome of a large number of additive factors.

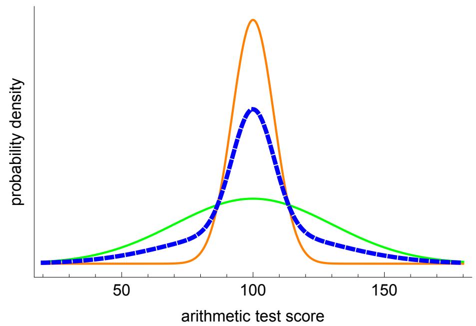
4. Data come from heterogeneous groups, with the same mean, but different variances.

Example uses: same uses as normal (it is a robust/over-dispersed model), literacy scores amongst adults of different ages, stock returns.

Like the binomial and poisson, the normal distribution has an over-dispersed cousin - the t distribution - which can be used to handle datasets with greater variability. Fortunately, like the other cases, this distribution can be derived in a similar manner; by imagining a process resulting in a mixture of normal distributions with different variances. The following example endeavours to make this a bit clearer.

Suppose that we were measuring the mean level of a standardised arithmetic test for two neighbouring schools: Privilege High, where the times are easy, and the fees extortionate; and Normal (forgive its other connotations) Comprehensive, where there is no fees, and they take kids of all abilities. We suppose that controversially, a local newspaper has done some research and worked out that the mean test scores are the same across both schools. However, Privilege High has a lower variance of test scores Normal Comp. Since, a student's test score is likely the result of another of factors: tiredness, work ethic, as well as their brilliance in math; we think that a normal distribution may be a reasonable distribution for the test scores of a number of individuals (see section 8.3.6).

Suppose that we are employed by the county, and are hence interested in developing a single model to explain the test scores across both schools. We can think of the resultant distribution as an average of the other two distributions; its head will rise, its shoulders fall, and its tails grow (see the left hand side of figure 8.9). This distribution will, in some senses, have a wider variability of results than a normal distribution of the same characteristics would allow (see the right hand side of figure 8.4). This variability isn't borne out in the variance of the distribution, but will be borne out in higher order moments (the fourth moment, called the kurtosis, here).



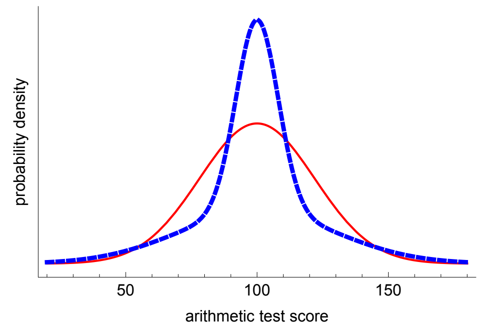


Figure 8.9: **Left**: the distribution covering test scores across both schools can be made from an average of the two individual distributions. **Right:** a comparison of the resultant distribution with a normal distribution of the same mean and variance.

We now imagine we are combining a large number of schools, each with the same the mean arithmetic score, but different variances, which we do not know. If we suppose that it is feasible that the variances varies according to: $\frac{1}{\sigma^2} \sim Inv - \Gamma(\frac{\nu}{2}, \frac{\nu}{2})$, then it turns out that an overall distribution across all schools can be proved to be a *Student t*[10] distribution with ν degrees of freedom (see the problem set at the end of the chapter). This distribution would have a larger variability than an equivalent normal with a variance equal to that of the mean of the variance distribution. The extra uncertainty has caused a widening of the distribution!

The higher the number of schools over which we are summing, the more the overall distribution starts to look normal (this is a consequence of the central limit theorem discussed in section 3.6), and we see a reduction in the tails. Figure 8.10 illustrates how manipulations of the parameters σ and ν which characterise the distribution affect its shape.

Now suppose that we have a sample of 5 test scores from across the schools, and are found to be: 74, 96, 108, 98, 101. Since we have three unknown parameter in our distribution the likelihood functions is actually a 3-dimensional surface, which is not possible to plot in our paltry 3 spacial dimensions (we need an extra dimension to show the height of the surface, meaning that the plot would be 4-D). However, we can proceed as before

[10]Named after the pseudonym used by William Sealy Gosset; a Guiness Brewery worker. Remarkable as this is perhaps the first time that statistics and alcohol were found to mix

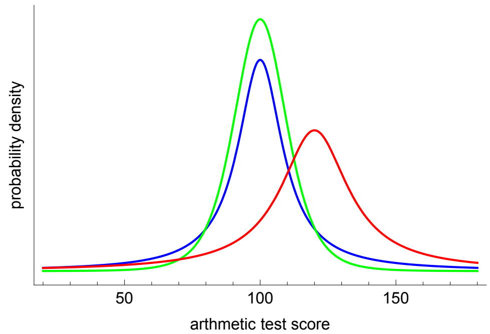


Figure 8.10: Student t distributions drawn for three parameter sets of (μ, σ, ν): (100,10,1), (100,10,5), and (120,15,1).

through pairwise contour plots of the variables (see figure 8.11). These are interesting in that they show that there is little correlation between the estimates of the mean of the distribution with the parameters σ and ν which dictate its shape. However, since both of these latter parameters dictate the overall variability of the distribution, the likelihood surface shows ridges of correlation between these two.

In the case when $\nu = 1$ this distribution has very fat tails, and it is sometimes called the *cauchy* distribution (see section 8.5.2). However, this choice of ν has a cost that the mean and variance are no longer calculable. Although, in circumstances where there are occasionally very extreme observations, this distribution provides a robust alternative. In fact, in his book 'The Black Swan' Nassim Taleb strongly advocates the use of this type of sampling distribution to account for rare (black swan) events [19].

$\boxed{Properties:}$

$$pdf : f(x|\mu,\sigma,\nu) = \frac{\left(\frac{\nu}{\nu+\frac{(x-\mu)^2}{\sigma^2}}\right)^{\frac{\nu+1}{2}}}{\sqrt{\nu}\sigma BetaFn\left(\frac{\nu}{2},\frac{1}{2}\right)} \tag{8.18a}$$

$$mean : \mathrm{E}[X] = \mu, if \nu > 1; Otherwise\ undefined \tag{8.18b}$$

$$variance : var[X] = \frac{\nu\sigma^2}{\nu-2}, if \nu > 2; Otherwise\ undefined. \tag{8.18c}$$

well.

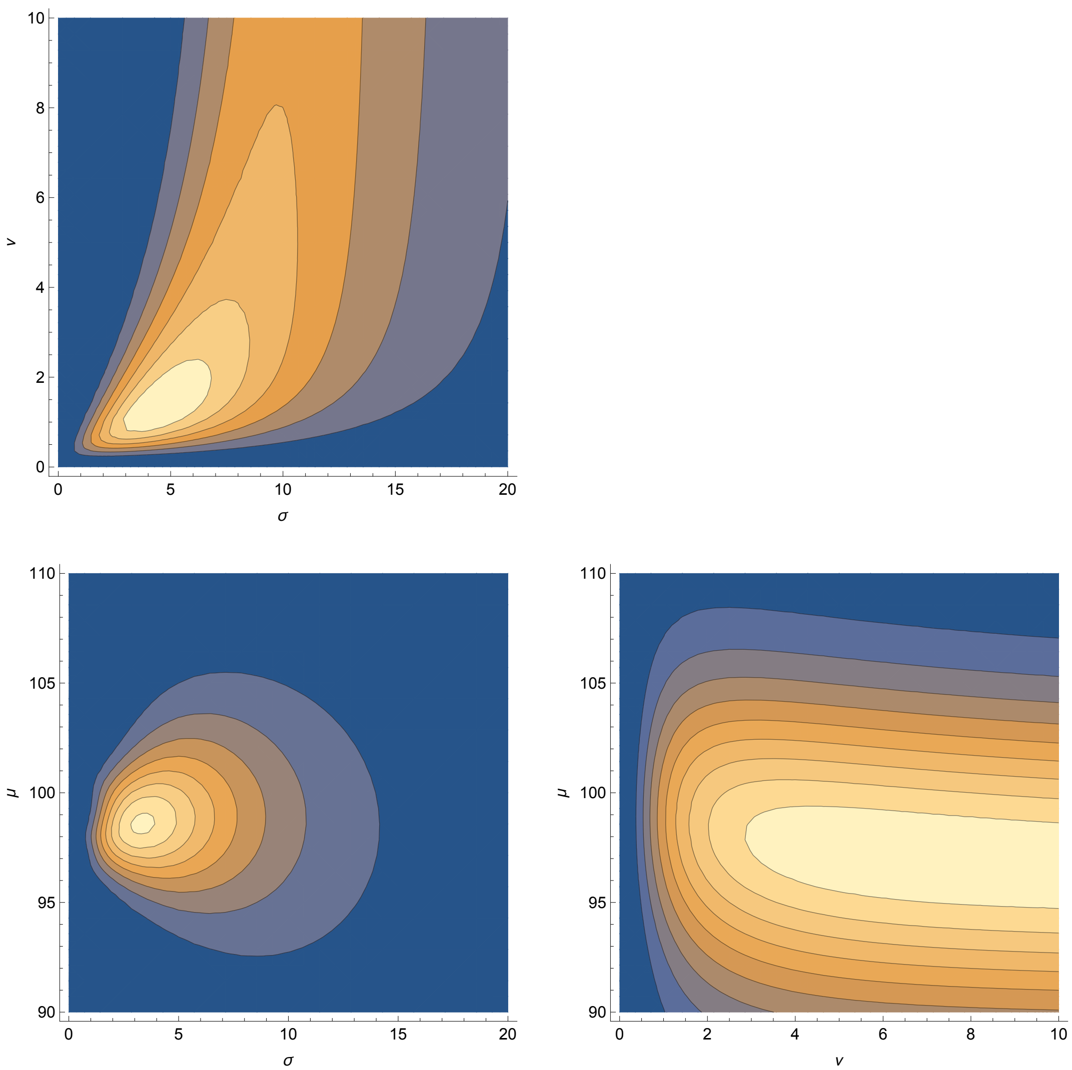


Figure 8.11: Pairwise contour plots of the likelihood surface. In each of the plots, the remaining variable is set at its maximum likelihood estimated value.

$\boxed{\textit{Video}:}$ Need to make an intuitive video here. $\boxed{\textit{Interactive}:}$ again make. These could be of two types: firstly showing how the sum of a user-selected number of normal distributions with variances given by an inverse-gamma distribution, results in a t. Another could be to allow the student to manipulate the parameters of the t distribution.

8.3.8 Exponential

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous, non-negative data.
2. Often used to measures times or space between events which occur independently, and at a constant rate through time or space respectively.

Example uses: failure times for artificial heart valves, the distance between the appearance of a new type of plant in a study area.

Suppose that you work for the World Health Organisation, and are interested in modelling the time between new epidemics of Ebola. If we define these epidemics as those *not* caused by spread from existent outbreaks, then we might suppose that each new case is a the result of a single new transfer of the virus from animal (most likely bats) to human. With this definition we might suppose that these crossover events occur independently of one another, and on a sufficiently long time scale (to neglect any seasonal effects), so that they might be considered to occur at a constant rate. In this circumstance, we might think that an *exponential* model for the times between outbreaks might be reasonable. This distribution has the following pdf:

$$f(x|\lambda) = \lambda e^{-\lambda x} \tag{8.19}$$

As can be seen from above, this distribution only depends on a single parameter λ - which is actually the same parameter that we have already seen measures the mean of a poisson process over some predefined time length

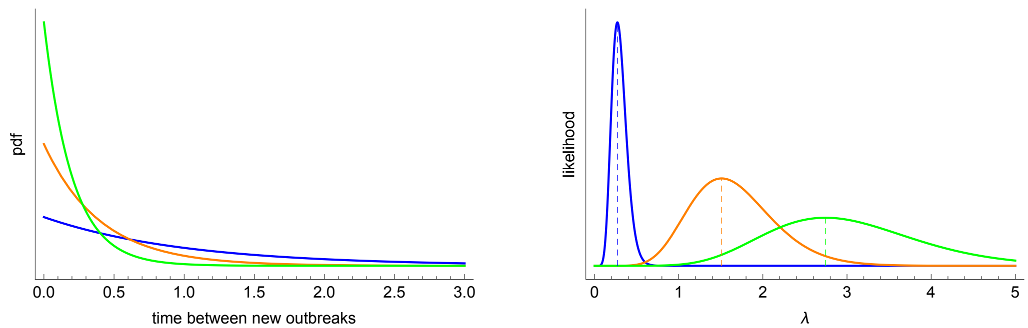


Figure 8.12: **Left:** the exponential probability distribution for three values of λ: 1 (blue), 2.5 (orange) and 5 (green). **Right:** the likelihood surface for three different datasets with mean times between outbreaks given by 3.1 (blue), 0.7 (orange) and 0.4 (green) years. The dashed vertical lines indicate the maximum likelihood estimates of the parameters in each case.**Add legends rather than state in text.**

(see section 8.3.3). Since this distribution depends only on a single parameter, it is relatively inflexible, and changing this parameter only governs the rate at which the distribution tails off to zero; the higher the λ, the faster the rate (see figure 8.12).

Suppose that we have the following data measuring the mean time, in years, between consecutive outbreaks across three different 10 year periods: 3.1, 0.7 and 0.4. Since the sample mean is a sufficient statistic (see section **??**) for λ, we can draw the likelihood for each of these three cases (see figure 8.12). Note that in each case, the maximum likelihood estimate of the parameter is straightforwardly $\frac{1}{\bar{x}}$.

$\boxed{Properties:}$

$$\boxed{\begin{aligned} notation &: X \sim Exp(\lambda) && (8.20a)\\ pdf &: f(x|\lambda) = \lambda e^{-\lambda x} && (8.20b)\\ mean &: \mathrm{E}[X] = \frac{1}{\lambda} && (8.20c)\\ variance &: var[X] = \frac{1}{\lambda^2} && (8.20d) \end{aligned}}$$

8.3.9 Gamma distribution

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous, non-negative data.
2. Greater flexibility than the exponential, but more complex.
3. Often used to model the time taken for the n independent events to occur.

Example uses: the time taken for the nth diode in a computer to fail, daily rainfall amounts in a variable environment.

Suppose that we now work for a company which organises lotteries within a particular city. Suppose that for a particular lottery, there is always one, and only one, jackpot winner. However, in one night in the holiday season, the company organises a 'bonanza' night, where there are 3 different games. People can enter all three, although they can win only one of them. Further suppose, that the time taken for individual jackpot winners to claim their prize is thought to be exponentially-distributed from the day of the game, with a rate $\lambda = 1\ day^{-1}$ that is roughly the same across people. It is of interest for the company to know how much time is likely to elapse before all three jackpot winners come forward, so that they can plan adequately.

Since we are interested in the time taken for all three lottery winners to come forward, we are essentially interested in the sum: $X \sim \sum_{i=1}^{3} T_i$; where T_i is the time taken for person i to claim their winnings. For one person, it is most likely that they come forward immediately, as the density is highest at 0. However, for all three people, it is highly unlikely that they all come forward immediately, and we expect the density to move away from 0 (see the left hand panel of figure 8.13). Also, because we are assuming that all three act independently of one another, the mean time taken for all three to come forward should be the sum of the mean time for each person (1 day), resulting in an expected time of 3 days. It turns out that in this circumstance, the time taken for all three people to come forward is described by a *gamma* distribution with *scale* parameter 3, and *shape* parameter 1; symbolically $X \sim \Gamma(3, 1)$.

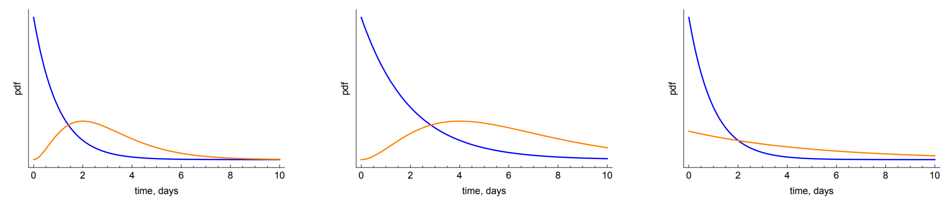


Figure 8.13: **Left:** the distribution of lottery claim times for 1 (blue) and 3 (orange) people, with $\lambda = 1$. **Middle:** the distribution of lottery claim times for 1 (blue) and 3 (orange) people, with $\lambda = \frac{1}{2}$. **Right:** the distribution of lottery claim times for 1 (blue) and 5 (orange) people, with $\lambda = 1$. **Add legends/titles.**

We can imagine what happens if the jackpot declines, resulting in people not being as driven to claim their prize. This can mathematically be represented by a rate parameter λ decrease to $\frac{1}{2}$. Individually, this means that there is greater uncertainty over the time taken to claim, meaning that across all three people, the distribution is significantly more spread out. Also, if λ declines, the overall mean will increase, which is seen as a shift rightwards of the density (see the middle panel of figure 8.13). Both of these are manifested in the gamma representation of the density $\Gamma(\alpha, \beta)$ by a larger value of the scale parameter $\beta = 2$.

Finally, if we consider a night where 5 different games are played, with 5 separate winners, with $\lambda = 1$. Now the density is shifted rightwards (see the right hand panel of figure 8.13), with a mean of 5, although our uncertainty has also increased vs the first case, since there are more people whose actions are uncertain!

Now imagine that we had the following data on the collective time taken (in days) for 3 people to claim their prizes: 5.1, 7.8, 3.4, 3.2, 4.3, 5.2, 8.1, 3.7, 2.3, 6.3. We now want to examine what the likelihood surface looks like for this sample, bearing in mind that we now have two parameters which can vary (see the left hand panel of figure8.14). We notice that since, for example, the mean of the distribution depends inversely on both parameters (with the $\Gamma(\alpha, \beta)$ parameterisation), that there is a strong positive correlation between these two parameters. Intuitively, to get the same mean, we can either have high α and β, or have both of them low, since the ratio will be the same. This strong correlation can be problematic for inference: both theoretically (since we have a problem of identifying the parameters of the model), and practically (many MCMC methods will be inefficient here).

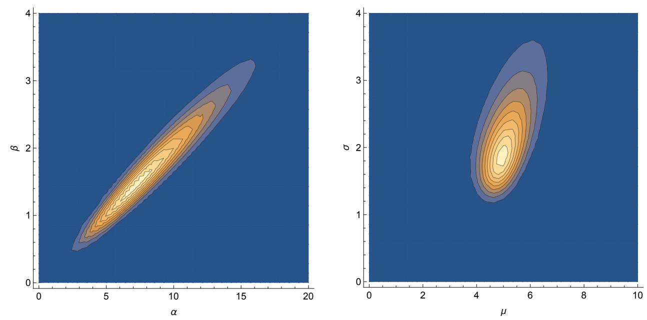


Figure 8.14: A contour plot of the likelihood surface across the parameters of a $\Gamma(\alpha, \beta)$ distribution modelling the time taken for 3 people to claim their prizes, with sample data: 5.1, 7.8, 3.4, 3.2, 4.3, 5.2, 8.1, 3.7, 2.3, 6.3. **Add legends/titles.**

On the right hand side of figure 8.14 we show the likelihood surface for the same distribution, but with the $\Gamma(\mu, \sigma)$ parameterisation. Note that this re-parameterisation has, to some extent, decreased the correlation between the two parameters; now the mean and variance. This might mean that in some circumstances the latter parameterisation is preferable (we actually prefer it since is more intuitive as well).

This distribution has three parameterisations, that we highlight below:

$\boxed{Properties:}$

$$notation: X \sim \Gamma(\alpha, \beta) \tag{8.21a}$$

$$pdf: f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{8.21b}$$

$$mean: \mathrm{E}[X] = \frac{\alpha}{\beta} \tag{8.21c}$$

$$variance: var[X] = \frac{\alpha}{\beta^2} \tag{8.21d}$$

$\boxed{Properties:}$

$$\boxed{\begin{aligned} notation &: X \sim \Gamma(k,\theta) && (8.22a)\\ pdf &: f(x|k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} && (8.22b)\\ mean &: \mathrm{E}[X] = k\theta && (8.22c)\\ variance &: var[X] = k\theta^2 && (8.22d) \end{aligned}}$$

$\boxed{Properties:}$

$$\boxed{\begin{aligned} notation &: X \sim \Gamma(k,\theta) && (8.23a)\\ pdf &: f(x|\mu,\sigma) = \frac{\left(\frac{\sigma^2}{\mu}\right)^{-\frac{\mu^2}{\sigma^2}} e^{-\frac{\mu x}{\sigma^2}} x^{\frac{\mu^2}{\sigma^2}-1}}{\Gamma\left(\frac{\mu^2}{\sigma^2}\right)} && (8.23b)\\ mean &: \mathrm{E}[X] = \mu && (8.23c)\\ variance &: var[X] = \sigma && (8.23d) \end{aligned}}$$

8.3.10 Multinomial

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Discrete data representing a finite set of categories.
2. Multiple trials.
3. Each trial has $k \geq 2$ outcomes.
4. Individual outcome probabilities are not determined by any other factor that varies systematically across individuals for which we have data.
5. Probability of obtaining a particular category is the same in each trial.
6. Overall data we measure is the aggregate numbers of outcomes in each category.

7. Trials are independent.

8. The generalisation of the binomial to handle an arbitrary number of categories.

Example uses: modelling the political party affiliation across a group of people (for example, Republican, Democrat, or Independent), the choice of treatment sought for back pain (for example osteopath, acupuncture, physiotherapy, surgery, or none), or the type of inferential framework used (for example, Frequentist, Bayesian, or rule-of-thumb).

Suppose we work for the Department of Health, and want to build a model that can explain the prevalence of blood types within a sample of people. For simplicity's sake, we suppose that the only categories of blood for which we want to build a model are: A, B or O. We suppose further that our test group of n people is randomly sampled from the larger US population. This latter assumption means that we can view the blood-type of one individual as being conditionally independent from another's.

Let's first of all consider how we might design a distribution to cope with the four categories of blood types. The probabilities of obtaining each of the four categories are equal to the proportions of each type in the population, which we call $\boldsymbol{p} = \{p_A, p_B, p_O\}$. We now create a set of binary random variables: $\{X_A, X_B, X_O\}$; each of which if equal to 1, means that the person has that bloody type. Note that here it is not possible for more than one of these set to be 1 for a given person (since they only have one blood type). We would like our distribution to return the relevant probabilities for each of the cases, in other words: $Pr(X_i = x_i, X_{-i} = 0|\boldsymbol{p}) = p_i$. Here $i \in \{A, B, O\}$ and X_{-i} represents the other three random variables. It turns out that we can get this behaviour from the following *multi-bernoulli* distribution:

$$Pr(X_A = x_A, X_B = x_B, X_O = x_O|\boldsymbol{p}) = p_A^{x_A} p_B^{x_B} p_O^{x_O} \tag{8.24}$$

Now that we have formulated a distribution for 1 person, we can fairly straightforwardly extend this to allow for a sample of n persons. Since we assumed that the group is randomly-sampled, then the overall distribution will be proportional to the product of the individual *multi-bernoulli* distributions, of the form above. We now define new random variables which hold the aggregate numbers of people in each of our categories: $Z_i = \sum\limits_{j=1}^{n} X_j^i$, where $i \in \{A, B, O\}$. From then we can reason (see the problem set at the end of the chapter) that the overall distribution has the form below:

$$Pr(Z_A = z_A, Z_B = z_B, Z_{AB} = z_{AB}, Z_O = z_O|\boldsymbol{p}) = \frac{n!}{z_A!z_B!z_O!} p_A^{z_A} p_B^{z_B} p_O^{z_O} \tag{8.25}$$

The factor at the front is the multiple-category version of the nCr term that we had for the binomial case, and allows for the fact that there are a large number of ways of obtaining the same aggregate numbers of people in each blood type (for example person 1 could be A, and person 2 B or vice versa).

We will see in section 8.5.1 that there is a trick that allows us to graph this distribution (see figure 8.15), and its likelihood for all three types. However, here we note that since there are only three categories, only two of them are free to vary, since the last category is given by $Z_i = n - Z_{-i}$, and for probabilities $p_i = 1 - p_{-i}$. This simplification means that we can draw the distribution in 2D, since the other variable is fully determined by the first two.

Figure 8.15 shows the probability distribution across Z_A and Z_B, with Z_O implicit for a sample $n = 5$, with $p_A = \frac{1}{2}, p_B = \frac{1}{3}, p_O = \frac{1}{6}$. Note that the distribution does not exceed the line $Z_A + Z_B = 5$, since it is not possible to have more people of either category than the sample size.

Figure 8.16 shows the likelihood across (p_A, p_B) for the following samples, each of size 10: 4, 2, 4, 5, 2, 3, 2, 2, 6. The plot is only the lowermost triangle, as to the right of the main diagonal the probabilities sum exceeds 1.

$\boxed{Properties:}$

$$\boxed{\begin{aligned} pmf &: Pr(X_1 = x_1, X_2 = x_2, ..., X_k = x_k|p_1, p_2, ..., p_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k} \\ mean &: \mathrm{E}[X_i] = np_i \\ variance &: var[X] = np_i(1 - p_i) \end{aligned}}$$

(8.26a)

(8.26b)

(8.26c)

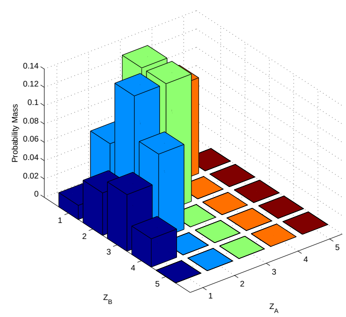


Figure 8.15: The multinomial probability distribution for probabilities ($p_A = \frac{1}{2}, p_B = \frac{1}{3}, p_O = \frac{1}{6}$).)**Add legends/titles, and combine with other - although one in Matlab, the other in Mathematica.**

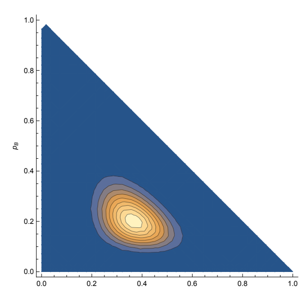


Figure 8.16: A contour plot of the multinomial likelihood across ($p_A, p_B, p_O = 1 - p_A - p_B$).**Add legends/titles, and combine with other - although one in Matlab, the other in Mathematica.**

8.3.11 Multivariate normal and multivariate t

The following is somewhat more (at least mathematically) advanced than the preceding distributions. Unless you are in need of multivariate models, and have a grasp of vectors, and matrices, then it can be left until you start Part V, after which we will start to use these concepts.

Distribution checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Multivariate distribution: used to specify likelihoods for *vectors*.
2. Continuous, unconstrained data.
3. Parameterised by two distinct parts: a vector of means, and a matrix of parameters which describes the covariance between elements of the vector.
4. Use multivariate t distribution to allow for a wider spread of data.

Example uses: gene expression data across multiple genes, test scores amongst members of family, temperature in neighbouring villages.

Imagine you get a job at a prestigious hedge fund as an analyst. Your manager gives you your first job: to model the risk of a portfolio containing the derivatives (complex financial instruments that allow for disproportionate profit *and* loss) on stocks of 2 different companies within a particular industry sector. You are told that if both stocks have a daily loss of 10% or more, then the portfolio goes bust. Further, your boss tells you that a risk of $\frac{1}{1000}$ is acceptable, but no more. You are given the historical daily returns of the stocks for the past year, (shown in the left hand panel of figure 8.17), and told to get modelling!

Excited about your first pay check, you get to work. Looking at the historical returns you notice that when one series tends to go up, the other does as well. So you plan to use a model that will take into account this covariance structure. Bearing this in mind, you go away and do some research, and decide that a reasonable model for the returns of the two stocks, $\boldsymbol{r}$, is a *multivariate normal*:

$$\boldsymbol{r} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8.27}$$

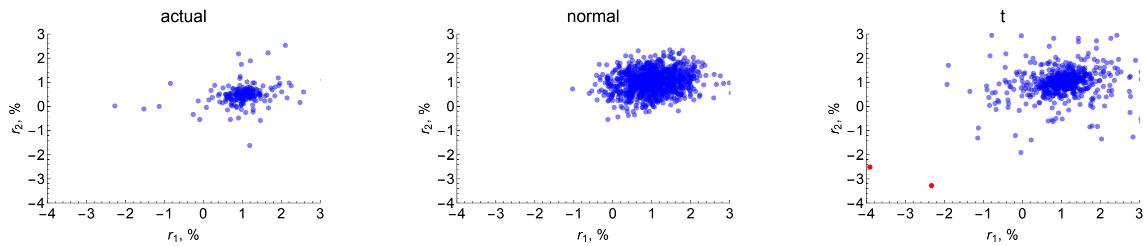


Figure 8.17: **Left:** the actual data from all days of trading last year. **Centre:** 1000 data points from a normal fit to the data. **Right:** 1000 data points from a t distribution fit to the data.

Note that in the above, the bold typeface indicates that we are dealing with vectors, and matrices rather than scalars. Thus here, $\boldsymbol{\mu}$ is a vector of length 2, where the individual elements correspond to the mean returns for each of the stocks (which are allowed to be different from one another). The matrix $\boldsymbol{\Sigma}$ is the *covariance* matrix, which can for our case be explicitly represented by:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{8.28}$$

Here σ_1^2, σ_2^2 represents the variances of the first and second stock daily returns respectively; ρ is the correlation between the daily stock returns of the two companies. So overall in our model there are 5 parameters: the two elements of $\boldsymbol{\mu}$ representing the mean return of each stock; and the three aforementioned parameters of the covariance matrix. The plots of the PDF for different values of ρ are shown in the top row of figure 8.18.

You fit your multivariate normal to the data using maximum likelihood (you haven't read this book yet), and then use your model to simulate 1000 replicate daily returns, resulting in the graph shown in the middle panel of figure 8.17. Feeling happy that none of your points are within the danger zone, you go to your boss, and tell her to "invest". Your boss being an analytical and sceptical person, asks you to show her your model basis and reasoning. Showing her the graph, and the model, she quickly realises that something is amiss, "Why does there appear to be more variability in the actual returns than we see in your fake data? I think you should think again!"

Feeling ashamed, you retreat to your desk, and realise the error of your ways. How could you have fallen into the trap of show many before, and used a distribution without sufficient allowance for extreme points? You dive back onto the web, and quickly find that the *multivariate t* distribution is a more robust alternative to the multivariate normal (see the bottom row of figure 8.18). You fit the distribution's parameters to the data (which is slightly harder since there is one more than originally - the degrees of freedom ν), then simulate the same number of replicates. To your shock, there are now a number of points in the danger area (see the right hand plot of figure 8.17). What's more, because there are more than 1 point in the danger area, the modelled risk of the portfolio going bankrupt is more than the threshold of $\frac{1}{1000}$. You slink over to your boss, and tell her, "Sorry - here are the results. We shouldn't invest." She recognises that people make mistakes, and that your new work is of much better quality than before, and says, "That's better. Go to lunch. We need you in a meeting in half an hour."

This cautionary tale should have conveyed the seriousness of ensuring that your modelling choices sufficiently take account of extreme data. In situations where there is evidence of this type of behaviour in the real data, and some would argue even if there isn't (see Mandelbrot's "Misbehaviour of markets" [11] and Taleb's "Black Swan" [19]), there is an argument in the multivariate world, in similar fashion to the univariate, that you should replace normals with t distributions. The multivariate t gives much more weight to the tails of the distribution, and is much better equipped to handle extreme variation in the data.

The multivariate normal has the properties shown below. Here k is the number of elements of X - in our example, the number of stocks under consideration.

$\boxed{\textit{Properties}:}$

$$notation: \boldsymbol{X} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{8.29a}$$

$$pdf: f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} \tag{8.29b}$$

$$mean: \mathrm{E}[X] = \boldsymbol{\mu} \tag{8.29c}$$

$$variance: var[X] = \boldsymbol{\Sigma} \tag{8.29d}$$

The multivariate t distribution's properties are also shown below.

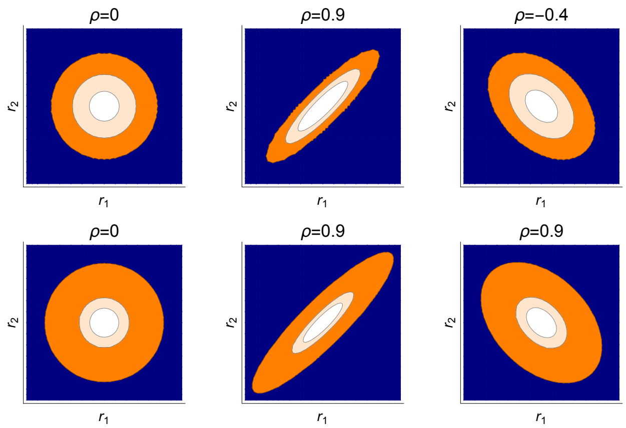


Figure 8.18: **Top:** contour plot of the 2D multivariate normal distribution function. **Bottom:** contour plot of the 2D multivariate t distribution, with $\nu = 3$ and the same covariance matrix as the normal.**Add legends/titles.**

$\boxed{Properties:}$

$$\boxed{\begin{aligned} notation &: \boldsymbol{X} \sim t_\nu\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ pdf &: f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma(\nu/2)\nu^{\frac{\nu}{2}}\pi^{\frac{\nu}{2}}|\Sigma|^{-\frac{1}{2}}\left[1 + -\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{(\nu+k)/2}} \\ mean &: \mathrm{E}[X] = \boldsymbol{\mu} \\ variance &: var[X] = \boldsymbol{\Sigma} \end{aligned}}$$

(8.30a)

(8.30b)

(8.30c)

(8.30d)

8.4 Table of common likelihoods, their uses, and reasonable priors

Show over-dispersed, or robust distributions for each of the cases.

8.5 Prior distributions

After a likelihood model is specified Bayesian analyses require that all specified parameters are allocated prior distributions, which represent our pre-analysis prejudices (see chapter 6). Much like the sampling distributions we have already met, there are a large number of possibilities available here, and it would be counter-productive to give space to all of them. Instead, we focus on some of the most common choices here from which, if necessary, more complex prior distributions can be constructed.

There are a number of different categories of parameters in probability models: probabilities and proportions, which are constrained to lie between 0 and 1; location parameters, encompassing for example, the mean; shape parameters, for example variances, which are non-negative; and distributions for discrete variables.

We will already have encountered some of these distributions in section 8.3, however the difference in focus here is sufficient to merit another (albeit short) discussion of each of these.

8.5.1 Distributions for probabilities, proportions and percentages

It is common to deal with variables that are naturally constrained to lie between 0 and 1. Examples include: probabilities, proportions, and percentages. In these cases, it is important to use priors that are fit for the purpose, since a poor choice here can lead to nonsensical values outside of this range.

Uniform

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous variable.
2. Variable constrained to lie between a and b; where for the case of probabilities, proportions, and percentages $a = 1$ and $b = 1$.

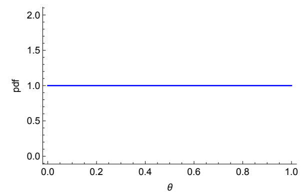


Figure 8.19: The continuous uniform distribution on $[0, 1]$.

3. Weakly informative prior.

A common choice of non-informative (although in section 6.8.1 we reasoned it was better to call these weakly informative) prior for variables constrained to lie between 0 and 1 is the continuous uniform prior (see figure). This choice of prior might be warranted if the analysis puts a premium on objectivity. This might be the case for clinical trials of drugs for example. It is easy to extend this distribution to handle situations where a variable is continuous and constrained to lie between two bounds $a < b$.

We argue that this distribution is actually superseded by the Beta distribution (see section 8.5.1), since the uniform is a special case of this more flexible and general distribution. However, since many analyses still use this distribution, it is worth knowing about it.

$\boxed{\textit{Properties}:}$

$$notation : \theta \sim unif(a, b) \tag{8.31a}$$

$$pdf : f(\theta|a, b) = \frac{1}{b-a} \tag{8.31b}$$

$$mean : \mathrm{E}[X_i] = \frac{1}{2}(a+b) \tag{8.31c}$$

$$variance : var[X] = \frac{1}{12}(b-a)^2 \tag{8.31d}$$

Beta

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous variable.

2. Variable constrained to lie between 0 and 1 (inclusive).

3. Encompasses a wide range of priors, ranging from weakly informative to strongly informative.

4. $Beta(1,1)$ equivalent to a $unif(0,1)$ distribution.

A more flexible distribution, compared to the uniform, for variables constrained to lie between 0 and 1, is the *Beta* distribution. There are a number of reasons why this distribution is preferable over the former. Firstly, this distribution encompasses a large range of potential priors, ranging from weakly informative (blue line of left-hand-panel of figure 8.20), to more strongly informative (green line). This easy ability to manipulate the distribution, allows for the analyst to quickly see the effects of changing priors on posterior conclusions. Secondly, the distribution allows a pre-experimental prejudice towards *any* value between 0 and 1 (see the right-hand-panel of figure 8.20), by changing its inputs. This is useful where there are pre-experimental reasons to suggest that some parameter values are more likely than others. As an example, consider obesity rates in the UK. We know that the percentage of people who are obese is less than 50%, so it doesn't make sense to use a vague prior here that gives equal weight to all values between 0 and 1. In this case we might be better going with a prior that gives more weight to some values less than 50% (orange line in right-hand-panel of figure 8.20 for example). Lastly, this distribution has the property that it is *conjugate* to a number of practically-useful distributions, and so allows for analytic calculation of the posteriors in these cases. Do not worry if this talk of conjugacy has gone over your head, as we devote the entirety of chapter 9 to this purpose.

One potential issue for Beta distributions is their use in hierarchical models. In these circumstances it can be difficult to formulate appropriate priors over the parameters of this distribution (see [5] for a way round this).

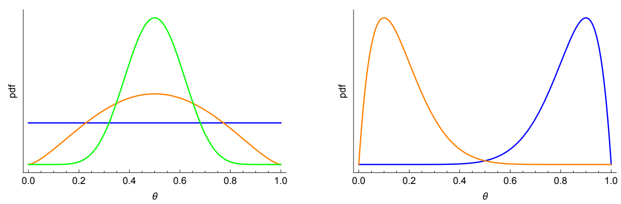


Figure 8.20: **Left:** three symmetric Beta distributions about $\frac{1}{2}$; $Beta(1,1)$ (blue), $Beta(2.5,2.5)$ (orange), and $Beta(10,10)$ (green). **Right:** two non-symmetric Beta distributions; $Beta(10,2)$ (blue), and $Beta(2,10)$ (orange). **Add legends.**

$\boxed{Properties:}$

$$notation : \theta \sim Beta(\alpha,\beta) \tag{8.32a}$$

$$pdf : f(\theta|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{BetaFn(\alpha,\beta)} \tag{8.32b}$$

$$mean : \mathrm{E}[X] = \frac{\alpha}{\alpha+\beta} \tag{8.32c}$$

$$variance : var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{8.32d}$$

Logit-normal

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous variable.

2. Variable constrained to lie between 0 and 1 (inclusive).

3. Encompasses a wide range of priors, ranging from weakly informative to strongly informative.

4. Perhaps more straightforward to parameterise in hierarchical models than a Beta distribution, although the jury is out on this one!

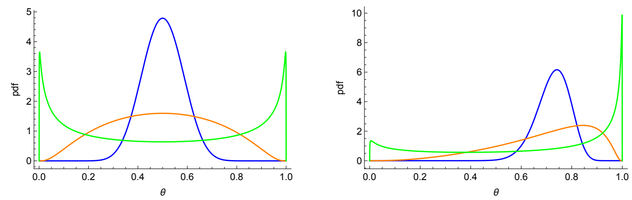


Figure 8.21: **Left:** $logit - normal(0, \sigma)$ distributions for $\sigma = \frac{1}{3}$ (blue),$\sigma = 1$ (orange), and $\sigma = 3$ (green). **Right:** $logit - normal(1, \sigma)$ distributions for the same values of σ.**Add legends.**

In life there are often multiple options that result in a similar outcome. In statistical inference, this is no different. An example of this is the relatively common use of the *logit-normal* distribution to model probabilities. The idea behind this choice is to allow an unconstrained variable to vary according to a normal distribution, then transform it to lie between 0 and 1, using a *logit* transform. Concretely:

$$log\left(\frac{\theta}{1-\theta}\right) \sim N(\mu, \sigma^2) \tag{8.33}$$

where the left hand side represents $logit(\theta)$. We can then use the *logistic* transformation to invert the expression, and get a distribution for θ. A large set or prior distributions are possible with this parameterisation (see figure 8.21). However, care must be taken to ensure that the prior does not place too much weight near 0 or 1, which can happen if we choose σ to be high (green lines).

This distribution is useful for two reasons: firstly, it perhaps extends slightly more naturally to hierarchical models than the Beta distribution; secondly, its multivariate analogue is a generalisation of the Dirichlet distribution to allow for correlation between probabilities. Do not worry about these two points too much, as we will cover hierarchical models fully in Part V, and the Dirichlet distribution next.

$\boxed{Properties:}$

$$\boxed{\begin{aligned} notation &: \theta \sim logit - normal(\mu, \sigma) \\ pdf &: f(\theta|\mu,\sigma) = \frac{1}{\sqrt{2\pi}(1-\theta)\theta\sigma} exp\left(-\frac{\left(\log\left(\frac{p}{1-p}\right)-\mu\right)^2}{2\sigma^2}\right) \\ mean &: \mathrm{E}[X] = no\ analytic\ expression \\ variance &: var[X] = no\ analytic\ expression \end{aligned}}$$

(8.34a)

(8.34b)

(8.34c)

(8.34d)

Dirichlet

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous variables.
2. Sum of variables constrained to 1.
3. Encompasses a wide range of priors, ranging from weakly informative to strongly informative.

We are frequently required to conduct inference over a range of discrete categories (see section 8.3.10). In particular, we may be interested in inferring the probabilities of obtaining each of the categories. For example, consider the example where there are *only* three political parties for which people can vote: Republican, Democrat and Liberals. In this circumstance, we know that the proportions of voting individuals choosing each party must sum to 1. Hence, inferences that we make about the three parameter values must be constrained to satisfy this condition.

We start by considering the case where there are only two political parties: Republican and Democrat. We can represent a particular parameter set as a point along a line (see the left-hand-panel of figure 8.22) of length 1. If we consider the leftmost point on the line, this represents the case where nobody votes for Republican, and everyone for Democrat, with the rightmost point being the exact opposite. At the mid-point, 50% vote Republican, and 50% Democrat.

When we generalise to the case of three categories, we can actually represent all feasible points as those lying within an equilateral triangle (see right-hand-panel of figure 8.22), where the axes corresponding to the three category probabilities are drawn from each vertex to the opposite midpoint. This triangular representation can be justified by imagining a 3D space defined by axes indicating (p_R, p_D, p_L). The set of allowable points corresponds to the plane defined by $p_R + p_D + p_L = 1$, with all probabilities being greater or equal to 0. It turns out (see the problem set), that this set of points that corresponds to an equilateral triangle-shaped plane in this 3D space (see figure 8.23).

More generally, this distribution is parameterised by k parameters: $\alpha_1, \alpha_2, ..., \alpha_k$, which can be thought of as weights towards that category; the higher the weight of one category relative to others, the greater the distribution will be skewed towards giving a high probability to that segment (compare the rightmost panel to the leftmost of figure 8.23).

This distribution is quite flexible, and allows for a large range of prior distributions - from fairly vague across the individual probabilities, to an informative orientation towards one of the categories. However, it should be said that this distribution isn't fully flexible, as we do not allow for correlation between the categories. An example might be analysing a survey which asked participants' percentage of travel by train, bike, car or other. We might expect there to be a correlation in preferences between train and bike, since they are both more carbon-friendly relative to a car. In this circumstance, the dirichlet distribution is not adequate, since it is unable to handle correlations between category probabilities. However, it transpires that the multivariate equivalent of the logit-normal is capable of allowing such relationships between category probabilities[11].

[11]See `https://en.wikipedia.org/wiki/Logit-normal\_distribution`.

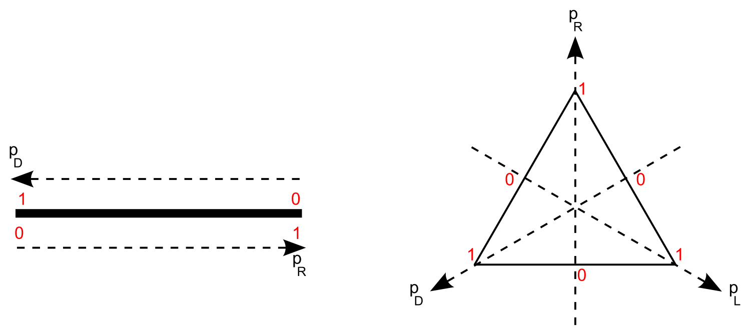


Figure 8.22: **Left:** feasible region for probabilities for two categories. **Right:** feasible region for probabilities for three categories.**Add legends.**

$\boxed{Properties:}$

$$notation: (p_1, p_2, ..., p_k) \sim Dir(\alpha_1, \alpha_2, ..., \alpha_k) \tag{8.35a}$$

$$where: \sum_{i=1}^{k} p_i = 1 \tag{8.35b}$$

$$pdf: f(\theta|\alpha_1, \alpha_2, ..., \alpha_k) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \tag{8.35c}$$

$$where: B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} \tag{8.35d}$$

$$and: \boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k) \tag{8.35e}$$

$$mean: \mathrm{E}[p_i] = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i} \tag{8.35f}$$

$$variance: var[p_i] = \frac{\alpha_i(\sum_{j=1}^{k} \alpha_j - \alpha_i)}{(\sum_{j=1}^{k} \alpha_j)^2(\sum_{j=1}^{k} \alpha_j + 1)} \tag{8.35g}$$

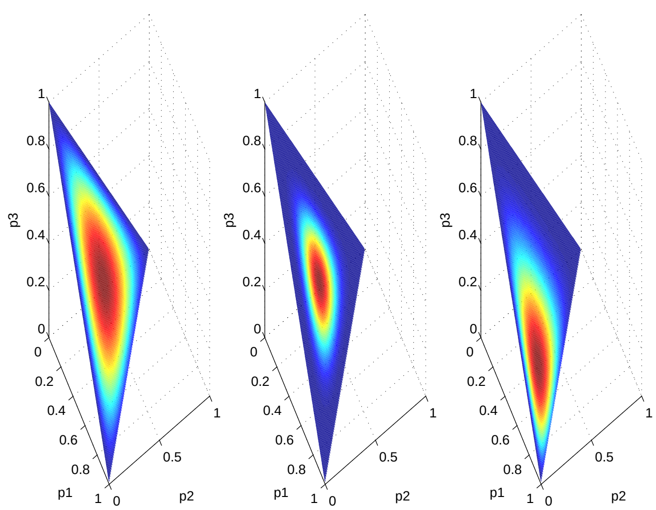


Figure 8.23: The dirichlet distribution for $(\alpha_1, \alpha_2, \alpha_3)$ equal to (2,2,2), (5,5,5) and (4,2,2), for the left, middle and right plots respectively.**Add legends, also change labelling to be pr, pd etc.**

8.5.2 Distributions for means and regression coefficients

In some continuous sampling models, the normal being the most frequently-used, the distribution is characterised by a location parameter - the mean - and a scale parameter - the variance, or standard deviation. Interest usually focuses on the location parameter, for example, estimating the mean proportion of the electorate that vote Republican, the average success rate for a new drug, or the median stock returns. Alternatively, in linear regression models the interest usually centres on estimation of the regression coefficients, which are multiplied by the independent variables to yield the mean (usually of a normal). As such, this section is devoted to showcasing some of the more popular choices for priors for these type of parameters.

Normal

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous unconstrained, or practically unconstrained, variables.
2. Encompasses a wide range of priors, ranging from weakly informative to strongly informative.

We already saw in section 8.3.6, that the use normal distribution for likelihoods under a wide range of situations, however, here we seek to justify its use for an altogether different purpose, for specifying prior distributions.

For example consider the problem of trying to estimate the global mean sea temperature, via a sample of many readings worldwide. We might suppose that there are a range of factors contributing to the temperature that we measure θ_i; time of year, geography, measurement error are some examples. These multitude of factors might justify the use of a normal distribution for the likelihood (see section 8.3.6):

$$\theta_i \sim N(\mu, \sigma^2) \tag{8.36}$$

This distribution is characterised by two parameters - its mean μ, and its standard deviation σ. Suppose that σ was known to a high degree of confidence for our sea temperature example. This might happen if we imagine we have compiled readings over a number of years, before our recent experiment, and found that the variance across locations does not change much over time. However, we suppose that there is an inherent variance in the mean temperatures measured across time. In this situation we might think that we are justified is treating σ as fixed, and assigning a prior for μ (actually, we would argue it is almost always worth estimating these unknown parameters). From experience we know that the mean sea temperature is likely in the range 10-22 degrees Celsius, although it varies from year to year. In this circumstance we may assign a fairly weakly informative normal prior for μ that captures the essence of this knowledge.

It is possible to examine the impact of changing the prior on the distribution for θ_i (see figure 8.24). It is important to note that this is *before* we carry out any Bayesian inference, and these distributions will often be very diffuse, and hard to describe. Whilst this does essentially represent a prior on the scale of θ_i, it is not that important to ensure this is located *only* around areas of the most likely parameter values, since the data will ensure that this is true for the posteriors. However, nonetheless it is worth bearing in mind the effect that changing priors can have on lower-down parameters.

As the variance of the prior distribution increases, it is unsurprising that this results in a higher variance for θ_i (see figure 8.24), although we do not derive the analytic results for the mean of θ_i and its variance here, as this is not our focus (see the chapter problem set). Similarly, the mean of θ_i exactly tracks that of the prior mean.

$\boxed{\textit{Properties}:}$ See section 8.3.6.

Student t

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous unconstrained, or practically unconstrained, variables.
2. Robust version of the normal, allowing for a greater range of parameter values.

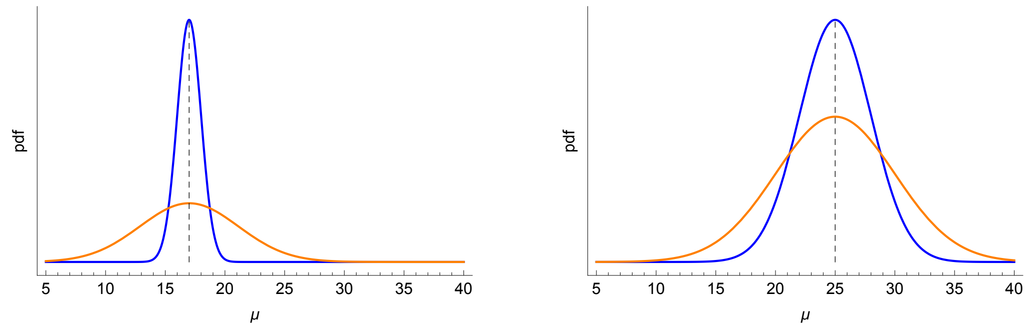


Figure 8.24: **Left:** the effect of a $\mu \sim N(17, 1)$ prior (blue) for $\theta_i \sim N(\mu, 4)$ distribution (orange). **Right:** the effect of a $\mu \sim N(25, 3)$ prior (blue) for $\theta_i \sim N(\mu, 4)$ distribution (orange). The mean of each distribution is shown as a dotted line.**Add legends.**

We suggested using the t-distribution for a sampling model when there was too much variation to be accounted for with the normal. Similarly, we advocate using the t-distribution for a prior when we wish to allow a wider range of parameter values, *a priori*, than would be possible with a normal.

To be clear, it is possible to set the parameters of a normal distribution such that it has the same amount of variance as that of a t distribution. However, the nature of that variance will not be the same (see figure). Whereas the normal is more evenly spread about its centre, the t distribution is sharply peaked, with less weight on its shoulders, and more on its tails. This difference between the two distributions is controlled by a parameter, ν, called the degree of freedom of this distribution (see section 8.3.7 for a better explanation of this parameter). The higher this parameter, the greater the correspondence between the normal and the t distributions, with in the limit that $\nu \to \infty$ they are the same.

$\boxed{\textit{Properties}:}$ See section 8.3.7.

Cauchy

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

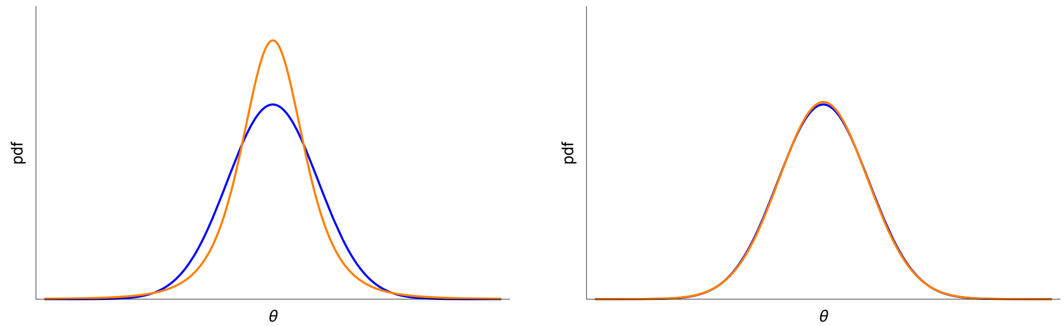


Figure 8.25: **Left:** normal (blue) and t (orange) distributions with the same variance and mean. Here $\nu = 4$ for the latter. **Right:** $\nu = 60$ for the t distribution.**Add legends.**

1. Continuous unconstrained, or practically unconstrained, variables.

2. Robust version of the t and normal distributions, allowing for a greater range of parameter values.

Imagine that you are building a model to help understand the effect of a new micro finance policy on the wage that an individual obtains in a particular location in a unspecified developing country. There are a number of factors which will also help determine a person's wage: education, social status and health to name but a few. To a first approximation we suppose that the relationship of interest can be assumed to be linear:

$$wage_i = \alpha + \beta mfinance_i + \epsilon_i \tag{8.37}$$

where $\epsilon \sim N(\mu, \sigma^2)$ represents the myriad of other factors influencing wage, $mfinance_i$ represents the amount borrowed through the scheme over the sampled period, and $wage_i$ is the amount of non-borrowed income obtained over the same period. We are assuming that both the variables - dependent and independent - are standardised (have mean 0 and s.d. 1); meaning that β represents the number of s.d. increase for a 1 s.d increase in $mfinance_i$. In the case where we are quite uncertain about the impact of the scheme, we may wish to allow a wide range of possibilities - potentially both positive and negative - for β before we carry out the investigation. Some proponents argue that the micro financing can exert disproportionately positive effects

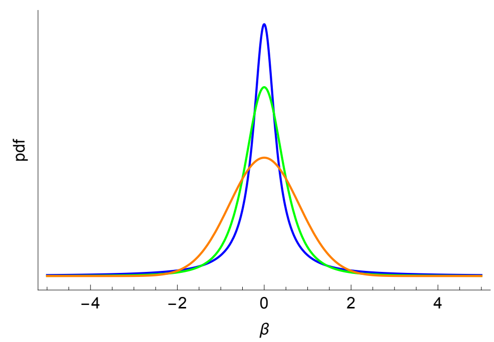


Figure 8.26: A normal (orange), a t (green) with 3 degrees of freedom, and a cauchy with scale parameter 0.3. The normal and t distributions drawn have variances of 0.64.**Add legends.**

on wage, by allowing the person the freedom to devote their time to producing a specialist good or service. Those in opposition perhaps argue that the moral hazard it induces can lead people to make worse choices, and hence detrimentally effects income.

In this case we may choose to specify a prior distribution that is wider than that which is given by a normal, or t distribution (with more than 1 degree of freedom); instead choosing to use a distribution that has fatter tails like the *cauchy* (see figure 8.26). This distribution actually corresponds to a t distribution with 1 degree of freedom.

Whilst there are benefits to this distribution, it does have its costs. In particular, the tails are so fat, that the distribution has no defined mean nor variance. This may look confusing, especially since the distribution is symmetric about its median. However, drawing repeated samples from this distribution is illuminating, as the running mean does not converge! Whether these costs are more than compensated for depends on the circumstances. It is recommended that this distribution only be used for those situations where we want to allow for significant prior range in the parameter value. In other, less extreme cases, the t distribution (with degrees of freedom greater than 1), or the normal will most likely suffice.

Properties :

$$pdf : f(\theta|\theta_0,\gamma) = \frac{1}{\pi\gamma\left(\frac{(\theta-\theta_0)^2}{\gamma^2}+1\right)} \tag{8.38a}$$

$$mean : \mathrm{E}[\theta] = undefined \tag{8.38b}$$

$$variance : var[\theta] = undefined \tag{8.38c}$$

$$median : median[\theta] = \theta_0 \tag{8.38d}$$

Multivariate normal, and multivariate t

The following is somewhat more (at least mathematically) advanced than the preceding distributions. Unless you are in need of multivariate models, and have a grasp of vectors, and matrices, then it can be left until you start Part V, after which we will start to use these concepts.

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Multivariate distribution: used to specify priors for a number of parameters simultaneously.
2. Continuous unconstrained, or practically unconstrained variables.
3. Mostly useful in hierarchical models, although can be used in non-hierarchical settings.

Imagine you are interested in evaluating the factors which contribute to the price of a house, P. You have two factors of interest which you want to investigate: the size of the house and gardens S, and the number of bedrooms B. We might choose to assume a linear relationship between the predictors and the price at which a house is sold:

$$P_i = \alpha + \beta_1 S_i + \beta_2 B_i + \epsilon_i \tag{8.39}$$

where $i \in \{1,...,N\}$. Before we carry out this analysis we might imagine that there is some correlation between the variables S_i and B_i. This dependence will most likely affect the estimates that we obtain of the parameters β_1 and β_2. Intuitively, since larger houses tend to have more bedrooms, we expect that if we choose to assign more weight to size, then we will have to assign less to bedrooms. Whilst in the simple example above, we could choose independent priors for the parameters β_1 and β_2, these would not represent our true pre-analysis conceptions of their interrelation. Forgetting α (we assume that we have standardised our data, and hence do not have to worry about correlations between the estimates of α and the other two parameters), in this circumstance, we might assume a prior of the following form:

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \sim N\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right]$$

Here $\rho < 0$ captures our pre-experimental prejudices as to the correlation in estimates of these two parameters, μ_1, μ_2 are the prior means, and σ_1^2, σ_2^2 are the prior variances. If we wanted to allow for greater freedom in the estimates of the parameters, we could instead choose to use a multivariate t distribution instead (see section 8.3.11).

Whilst in this non-hierarchical example, we could have easily assigned independent priors on the coefficient estimates, and let the data speak for itself, in hierarchical models, it is more important to take parameter dependence into account. In multi-level modelling this allows for more sensible estimates of parameters to be obtained, as well as allows estimation of overall variances and covariances in parameter estimates.

$\boxed{\textit{Properties}:}$ See section 8.3.11.

8.5.3 Distributions for non-negative parameters

There are a wide class of parameters which are naturally constrained to be non-negative in value. Examples include variances, as well as a range of other parameters including the mean of a poisson distribution, the shape parameter of a gamma, and so on. Whilst with Stan it is feasible to use some of the aforementioned distributions by setting bounds on parameter values, it is still preferable in many circumstances to set priors for parameters that naturally have support (also called positive density) on the correct range of parameter values. This section not only introduces a wide range of these distributions, but also provides some discussion as to which are preferred under particular circumstances.

Gamma

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous variables constrained to be non-negative.
2. Frequently used for parameters where there is lower uncertainty than those for the inverse-gamma, half cauchy or inverse-χ^2.

Suppose that we are tasked with building a statistical model for the count, X_i, of road accidents occurring on a given day, i, in a particular (small) geographic area to help to allocate emergency service resources. Supposing that the weather is sufficiently similar over all days for which data collection took place, then a reasonable model here might be the *poisson*: $X_i \sim poiss(\lambda)$. Historically, we know that the number of incidents has a mean around 10, although there is some variance around this value. In this circumstance, we might choose to use a gamma prior for λ, not least because of its mathematical convenience (due to its conjugacy - see chapter 9), but also because we have a fair degree of confidence from previous studies as to its most likely values. Some example priors for this circumstance are shown in figure 8.27, along with the resultant posteriors for a sample of 5 days. We can see that even though we only have a modest amount of data, the choice of prior distribution here (within reasonable bounds), only has a slight effect on the posterior.

$\boxed{\textit{Properties}:}$ See section 8.5.3.

Half-Cauchy, inverse-gamma, inverse-χ^2 and uniform

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

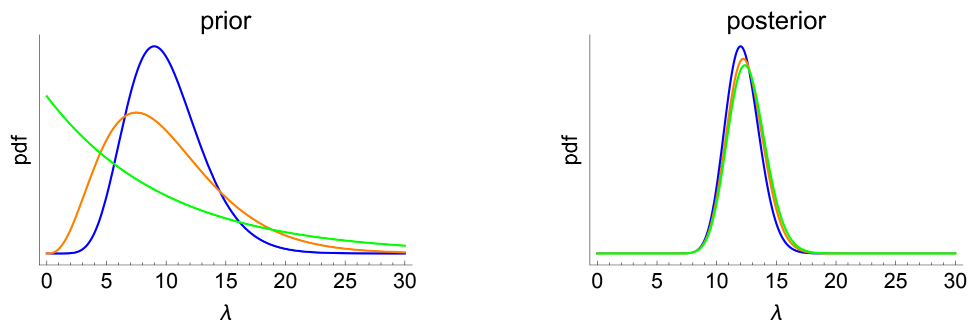


Figure 8.27: **Left:** three prior distributions for $\lambda \sim \Gamma(\alpha, \beta)$ where $\frac{\alpha}{\beta} = \mu$: $\Gamma(10, 1)$ (blue), $\Gamma(4, 2.5)$ (orange), and $\Gamma(1, 10)$ (green). **Right:** the resultant posteriors for the sample data {17, 6, 11, 13, 16} for each of the prior cases. **Add legends**.

1. Continuous variables constrained to be non-negative.
2. Often used for variances, as well as shape parameters (parameters which govern the shape of a distribution rather than its centre or its spread.)
3. Variants are typically used for specifying non-informative priors.
4. For non-informative priors it is recommended that analyses use a half-cauchy over the alternatives.

Imagine that we are tasked with finding a statistical model to represent the length, L, of fully-grown adult male crocodiles. We suppose that since size is determined by a whole range of factors (genetics, climate, local competitors), that a normal sampling model may be appropriate: $L \sim N(\mu, \sigma^2)$. Further imagine that the study that we are conducting is fairly novel, and is a first attempt at describing the lengths of these animals, so that previous data is scarce.

Here our primary purpose of conducting the inference might be to focus on the parameter μ representing the mean length of the male members of the species, thought to be around 4.5 from previous observational evidence. The other parameter σ^2 gives the spread of the values about this mean, and might in this circumstance be regarded as a nuisance parameter, since it is not the primary purpose of estimation.

Rather than set σ^2 manually, we would rather it be determined by the data. Hence, we want to set the parameter a prior. There are two distributions that are used here that have nice properties (in particular, conjugacy - see section 9.5): the inverse-gamma and the inverse-χ^2 distributions. The latter is a special case of the former here, although since both are sometimes used in the literature without recourse to the other, we consider them separately here.

Both of these distributions are derived from the bit after the 'inverse' in their names. An inverse-gamma(α,β) is simply the distribution of ζ^{-1} where $\zeta \sim \Gamma(\alpha,\beta)$. In words, we can imagine sampling a value, ζ from the $\Gamma(\alpha,\beta)$, then taking $\frac{1}{\zeta}$. If we took enough samples we would obtain a histogram that has the shape of a inverse-gamma(α,β). The inverse-χ^2 distribution, as the name suggests, is derived from the χ^2 distribution (not discussed in this book, but it is a special case of the Γ distribution) in the same way, although is simpler than the inverse-gamma; being parameterised by a single input only. Both of these distributions are often used to specify noninformative priors for parameters like our σ^2. For the inverse-gamma, this is obtained by letting each of its inputs tend to zero. For the inverse-χ^2, this is obtained by letting its sole input tend to zero.

Another distribution that has been popularised by Gelman amongst others, is the *half-cauchy* distribution, which is a truncated version of a cauchy-distribution (DO WE DISCUSS THIS?), where it is forced to have non-zero density only on the non-negative values of the sampled parameter.

A final alternative sometimes used is a continuous uniform distribution over some non-negative and finite set. The trouble with specifying this distribution is setting its upper-bound, as otherwise the distribution would be improper.

In circumstances where there is considerable prior knowledge as to the parameter value under consideration, then it may be useful to use the two 'inverse' distributions due to their conjugacy in certain situations. However, since we are very uncertain as to the value of the parameter σ^2 in our current example, we would like to specify the distribution which best represents our ignorance. We show in figure 8.28 some typical non-informative priors for the parameter σ; one for each of our distributions. We have chosen to specify priors for σ rather than σ^2 for easier interpretation below. We see that whereas the two 'inverse' distributions show relatively fast changes in density near 0, this is not the case for the half-cauchy which is much

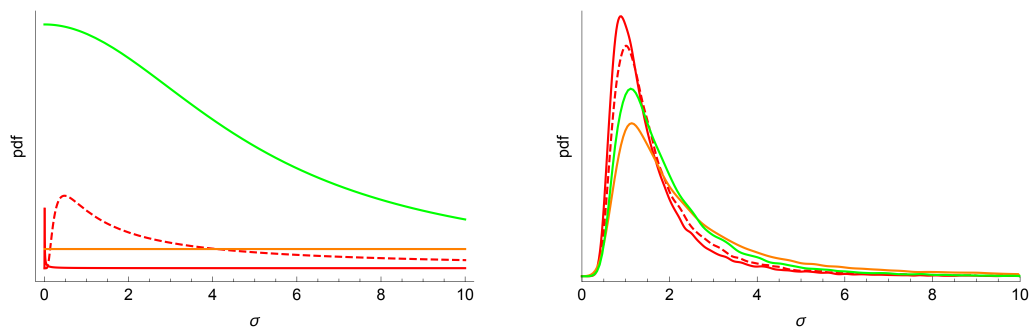


Figure 8.28: **Left:** the prior distributions for half-cauchy(0,5) (green), uniform(0,100) (orange), inverse-χ^2(0.1) (red-dashed) and an inverse-gamma(0.0001,0.0001) (red). **Right:** estimated posteriors using the left priors for a data sample of size 3 from an MCMC sample of 100,000 observations using Stan (see web for program to exactly replicate samples).**Add legends**.

smoother.

Importantly, we also show posterior distributions that we have estimated using MCMC (do not worry if you do not understand what this means, as we devote the entirety of Part IV to this purpose). We see that the two inverse distributions have distributions which are generally to the left of the half-cauchy and the uniform; particularly evident for values of $\sigma \leq 5$. We also see that whereas the uniform distribution allows values of σ close to 6, the other three do not allow this.

Here the data sample of size 3 were simulated from a $N(4.5, 3^2)$ distribution; meaning that the true value of $\sigma = 3$. In this case the distribution that yields the most sensible estimates of this parameter is the half-cauchy. It is midway between the extremes of the two densities that are too peaked at zero (the inverse-gamma and -χ^2 respectively), and the uniform that gives too much credence to unrealistic parameter values. These results are similar to those presented in Gelman [6], and although are partly due to judicious choice of prior parameters, nevertheless highlight some weaknesses of three out of the four distributions. These weaknesses become even more important in hierarchical models, and hence we advocate strongly for the use of half-cauchy distributions for these types of parameters.

$\boxed{Properties:}$ We include only properties of the preferred distribution - the half-cauchy - and refer the readers to the table contained in section 8.4 for a few of the properties of the inverse-gamma and inverse-χ^2. The continuous uniform distribution is discussed in section 8.5.1.

$$pdf : f(\theta|a,b) = \frac{2b}{(a^2 - 2a\theta + b^2 + \theta^2)\left(2\tan^{-1}\left(\frac{a}{b}\right) + \pi\right)} \quad If\theta > 0 \tag{8.40a}$$

$$mean : \mathrm{E}[\theta] = not\ well\ characterised \tag{8.40b}$$

$$variance : var[\theta] = not\ well\ characterised \tag{8.40c}$$

Log-normal

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Continuous variables constrained to be non-negative.
2. Parameter can potentially lie across range of width of an order of magnitude, or more.

Imagine that you are tasked with devising a probability model for the counts of traffic accidents, X, within a particular week of the year, in a location with considerable variability in weather. In this circumstance, we might think that a negative binomial likelihood distribution would be appropriate (see section 8.3.4): $X \sim NB(\mu, \kappa)$. We choose to use the specification of the distribution such that $var[X] = \lambda + \frac{\lambda^2}{\kappa}$. In this circumstance, κ represents the inverse of the dispersion, and in the limit $\kappa \rightarrow \infty$ this distribution becomes the poisson.

If we were not privy to much historical information on the traffic incidents for this particular location, at this time of year, then we may feel somewhat uninformed on κ, and ranges from near 0 to reasonably large values are all

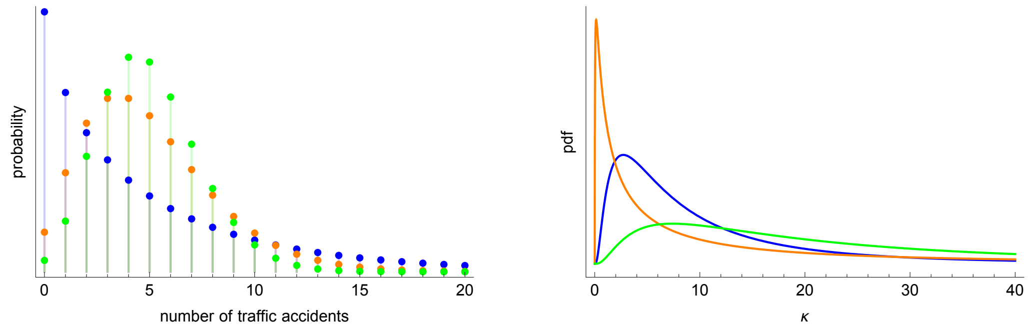


Figure 8.29: **Left:** the negative binomial sampling distribution for $\mu = 5$ with $\kappa = 0.8, 5, 40$ in blue, orange and green respectively. **Right:** log-normal prior distributions for κ, with (μ, σ) given by (2,1), (2,2) and (3,1) in blue, orange and green. Note that the μ used in the right-hand graph is different to that one in the left.**Add legends**.

feasible. In figure we show that up until $\kappa = 40$, increases in κ can have considerable effect on the resultant sampling distribution. Hence, in this situation we would like to use a prior distribution which allows this degree of variation in the parameter.

A *log-normal* distribution satisfies this requirement. The 'log' part of the name just specifies that we ascribe a distribution for the *log* of the variable of interest; the 'normal' part says that this distribution is *normal*. Mathematically if we use this prior distribution for κ, this means $log(\kappa) \sim N(\mu, \sigma^2)$. However, since this distribution is sufficiently popular, we create a distribution for κ in this circumstance, and call it the 'log-normal'.

This distribution can easily span across a range of different scales, although care must be taken when using it, as it is extremely sensitive to its parameterisation (see the variance dependence on μ and σ^2 in the properties section below). As such, if this distribution is necessary, we very strongly advocate that you draw the prior distribution using mathematical/statistical software each time to check that it looks sensible. In hierarchical models, this prior may become even more problematic, and it may be advisory to use a less sensitive distribution for top level hierarchical parameters, such as the half-cauchy, uniform, or altogether different distribution dependent on circumstance.

$\boxed{\textit{Properties}:}$

$$notation: \theta \sim log - N(\mu, \sigma^2) \tag{8.41a}$$

$$pdf: f(\theta|\mu, \sigma^2) = \frac{e^{-\frac{(\mu - \log(\theta))^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma\theta} \tag{8.41b}$$

$$mean: \mathrm{E}[\theta] = e^{\mu + \frac{\sigma^2}{2}} \tag{8.41c}$$

$$variance: var[\theta] = \left(e^{\sigma^2} - 1\right) e^{2\mu + \sigma^2} \tag{8.41d}$$

8.5.4 Distributions for covariance and correlation matrices

There are circumstances, particularly for hierarchical models, where it is necessary to specify priors for covariance and correlation matrices. These entities are quite different to most of the priors that we have discussed thus far, in that they describe objects that are highly constrained (see the discussion in section below). As such, the following section is a little more advanced than that which we have covered thus far, and can be left until after Part V if you are less comfortable with matrices.

There are two basic choices for covariance and correlation matrices. The old choice was to use Wishart and, more frequently, inverse-Wishart distributions for these matrices due to their nice conjugacy properties (see chapter 9). However, Stan makes this choice unnecessary, and we are free to use better alternatives. A particularly important recently created default prior is the LKJ distribution, due to its flexibility, particularly when compared to the aforementioned [10]. As such we cover this distribution first, with the implicit understanding that this distribution should act as a replacement for most circumstances where the Wishart-family were used before. However, there are still analyses that persist with the old guard, and to make these understandable, we spend the second half examining the two types of Wishart distributions.

LKJ

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Distribution for covariances, or correlation matrices.
2. By their nature the parameters of these matrices are constrained (the matrices must be positive definite - see below).
3. Mostly useful in hierarchical models, although can be used in non-hierarchical settings.
4. A better alternative to the Inverse-Wishart, and Wishart for almost all practical circumstances.

Suppose that we are interested in modelling the comorbidity (the presence of the condition in a patient) between a number of distinct psychological conditions: major depression, schizophrenia, obsessive compulsive disorder, and anorexia. Before we carry out the investigation, we are unsure whether the presence of any one of these disorders in a patient makes it more likely that they suffer from another. We are also unsure as to the underlying rates of these conditions, and would also like to model these. In this circumstance, a reasonable starting point might be to assume a multivariate normal likelihood for our sample of 10,000 people (see section 8.3.11):

$$\boldsymbol{d} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8.42}$$

Here $\boldsymbol{d}$ represents a vector of length 4, where the elements refer to the numbers of individuals within our sample who are diagnosed as having major depression, schizophrenia, obsessive compulsive disorder, and anorexia respectively. The 4-vector $\boldsymbol{\mu}$ indicates their underlying prevalences within the population, and $\boldsymbol{\Sigma}$ is a 4x4 covariance matrix representing their variances (on the diagonals) and covariances (off diagonals).

If we choose to use priors for the mean vector and covariance matrix independently, then we can consider the latter on its own. For the covariance matrix, this means that we need to come up with a prior distribution of covariance matrices; meaning that we need to associate a probability density with each possible covariance matrix. It is important to emphasize the word *possible* in the previous sentence. There are an infinity of 4x4 matrices, but not many of these will be admissible as covariance matrices. So what properties do we require of such a matrix?

- The diagonal terms are variances, and so must be positive.

- The off-diagonal terms can be positive or negative, representing positive or negative covariance respectively.
- There is a symmetry between the off-diagonal elements since, for example, the covariance between depression and anxiety must be the same as vice versa.
- The square of the covariances must be less than the product of the variances, forcing some structure on the relationships between the off-diagonal elements and the diagonal.

These requirements can be represented by assuming that all allowable covariance matrices are *positive definite* and symmetric. Do not worry about the first definition now, it does not serve our purpose to dive into its meaning in this book, just know that it is an algebraic condition meaning that all the above are satisfied when combined with symmetry. Requiring positive definiteness for our covariance matrix reduces the number of possible objects considerably, and tells us that we need to be very careful in constructing them.

Fortunately, we do not need to go about constructing the covariance matrices, since this job has already been done very well for us by [10]. However, to use their method we firstly follow [1], in decomposing our covariance matrix:

$$\boldsymbol{\Sigma} = \boldsymbol{\tau}\boldsymbol{\Omega}\boldsymbol{\tau} \tag{8.43}$$

where $\boldsymbol{\tau}$ is a vector representing the variance scales of each variable, and $\boldsymbol{\Omega}$ is a *correlation* matrix. Correlation matrices are similar to covariance matrices, but with the additional requirement that the diagonal elements are all 1 (the correlation of anything with itself). We can use a weakly informative non-negative prior (see section 8.5.3) for each of the elements of $\boldsymbol{\tau}$, and to keep things simple we assume it is independent of the prior on $\boldsymbol{\Omega}$. We are now in a position to use the *LKJ* distribution (named in honour of the names of the authors of [10]):

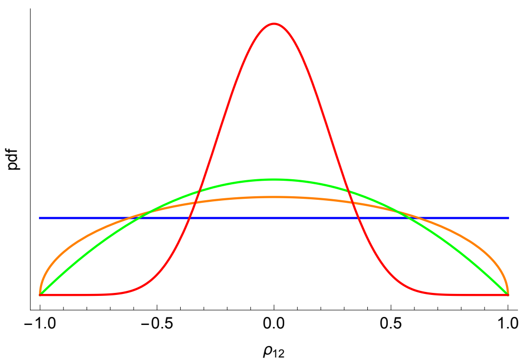
$$\boldsymbol{\Omega} \sim LKJ(\eta) \propto |\boldsymbol{\Omega}|^{\eta-1} \tag{8.44}$$

where $|\boldsymbol{\Omega}|$ represents the determinant of the correlation matrix $\boldsymbol{\Omega}$. Importantly this is a distribution only over *possible* correlation matrices, and hence only applies if the matrices are symmetric, positive definite and have unit diagonals.

But what does this distribution actually represent? And, how does the parameter η determine its shape? Here we will introduce some of the ways to visualise these distributions, but refer the interested reader to [21] for a more thorough exposition of ways to visualise covariance/correlation matrices (from which we borrow here a few suggestions).

It can be helpful to ascribe some meaning to the determinant of a covariance matrix. This is sometimes referred to as the *generalised variance*, which captures the overall freedom of the system to vary. If the off-diagonal elements are zero, then the system is very free to vary. If by contrast, the off-diagonal elements increase, the variance of the system in one direction is not free of another (since they covary), meaning there is less overall free variance (and a lower determinant). Similarly, we can think of the determinant of a correlation matrix as representing a kind of *generalised correlation*, having a maximum of 1 (where all elements are independent), and a minimum of 0 (where everything is the same). So essentially, the LKJ distribution gives a density over all possible values of generalised correlation.

We can think about what the above distribution looks like for different values of η [1]. Starting with $\eta = 1$, this means that all possible correlation matrices (of different determinants) are equally likely to be picked. Here we say that the distribution is *uniform* over all possible correlation matrices. For a 2x2 correlation matrix, the only choice is for the off-diagonal element, ρ_{12} which must be between -1 and 1. Since, all values of ρ_{12} result in valid correlation matrices, then its distribution is uniform over these bounds (see figure 8.30). For a 3x3 matrix, we are more constricted in our choice of ρ_{12}, since there are two other parameters we must also choose: ρ_{13} and ρ_{23}, requiring that the resultant matrix must be symmetric positive definite. It transpires that there are more allowable correlation matrices if we choose ρ_{12} to be close to 0, since this places fewer restrictions on the other two parameters; meaning that the distribution is peaked at 0 (see figure 8.30). For higher dimensional matrices (like our 4 element example), the restrictions which we place on one element affect other elements more, meaning that there are fewer allowable configurations if the value of ρ_{12} is extreme. It transpires that the distribution of allowable ρ is in general given by a $Beta\left(\frac{d}{2}, \frac{d}{2}\right)$ distribution, where d is the dimension of the matrix [10]. This results in a joint distribution of the form shown in figure 8.30 for



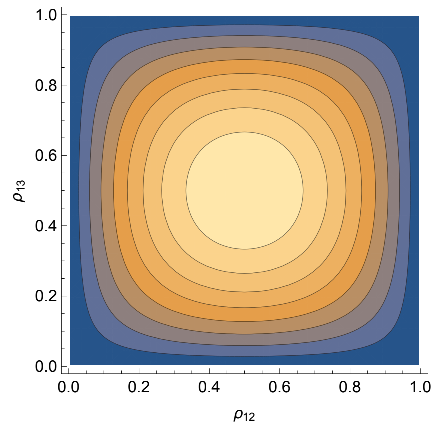


Figure 8.30: **Left:** the distribution of ρ_{12} for allowable correlation matrices, for a dimension of 2 (blue), 3 (orange), 4 (green) and 20 (red). **Right:** a contour plot of the joint distribution of ρ_{12} and ρ_{13} for a correlation matrix of dimension 4. **Add legends**.

the parameters ρ_{12} and ρ_{13}.

For all other values of $\eta \neq 1$, the determinants of the matrices matter. However, before attacking the problem of what the distribution of matrices looks like now, it is worth pausing a minute to think about what the distribution of determinants looks like for all possible correlation matrices (i.e. for when $\eta = 1$). For the 2x2 case, we know that the determinant is given by:

$$|\mathbf{\Omega}| = 1 - \rho_{12}^2 \tag{8.45}$$

Which has a maximum value of 1 (the identity matrix), and a minimum of 0 (matrix of 1s). We have already reasoned that for a 2x2 correlation matrix, the distribution of ρ_{12} is uniform between -1 and 1, meaning that we expect the distribution of ρ_{12}^2 to be peaked at zero (see section 6.8.1). This means that the distribution of $|\mathbf{\Omega}|$ will be peaked at 1, and decay away to 0 (see figure). This means that the majority of admissible correlation matrices will be towards the identity matrix, meaning that the correlations are quite small.

For a dxd correlation matrix, the distribution of determinants of allowable correlation matrices is more complex, but it can be derived analytically, and transpires to be given by [8]:

$$|\mathbf{\Omega}| \sim \prod_{j=1}^{d-1} Beta\left(\frac{j+1}{2}, \frac{d-j}{2}\right) \tag{8.46}$$

where $Beta\left(\frac{j+1}{2}, \frac{d-j}{2}\right)$ is shorthand for a Beta distributed random variable with those values of its inputs. The distribution for the determinants of all allowable correlation matrices is shown in figure 8.31, and illustrates that for $d > 2$ the number of these is increasing as $|\mathbf{\Omega}|$ decreases. This is because as the matrix gets closer to the identity, there is less latitude in the variability in the off-diagonal parameters.

If we then allow for a preference of values of $|\mathbf{\Omega}|$ which are greater, by letting $\eta > 1$, we find that the distribution shifts more mass towards higher values of the determinant (see figure 8.31), in other words more towards the identity correlation matrix. By contrast, if we choose a value of $\eta < 1$, then we prefer to sample those correlation matrices with determinants closer to 0. These matrices are a limit of the correlation matrix, when the off-diagonals approach the value of the diagonals (although do not quite get there), and can be thought of the case where there is almost a correlation of 1 between the variables.

In summary, if we want to allow for a relatively unrestricted correlation structure between our variables, as we might in our current example, then using an LKJ prior with $\eta \sim 1$ seems reasonable. By contrast, if we thought before experiment that the variables were likely independent, then we might choose a higher value of η. However, if we choose a value $\eta > 10$, then the prior distribution is very similar to an identity, and we might be better off computationally as specifying independent priors for each of the variances, and forgetting the covariances.

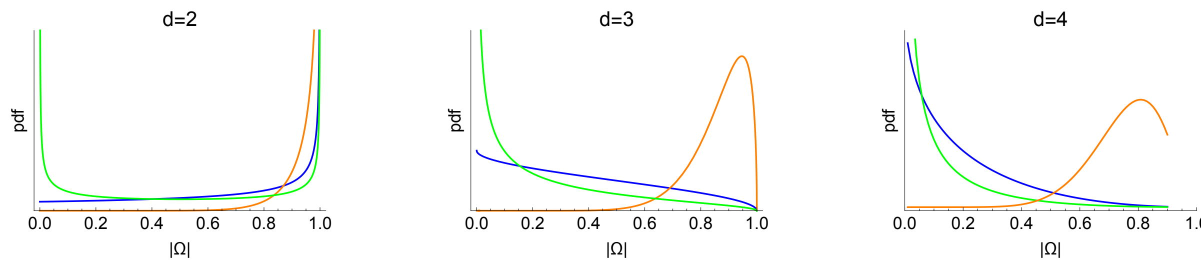


Figure 8.31: The LKJ distributions over allowable correlation matrices for left: $d = 2$, centre $d = 3$, and right: $d = 4$ when $\eta = 1$, $\eta = 10$ (orange), and $\eta = 0.5$ (green). **Add legends. Also need to draw last figure from 0.9-1, as it was too difficult to find pdf here.**.

$\boxed{Properties:}$

$$\boxed{\begin{aligned} decomposition &: \boldsymbol{\Sigma} = \boldsymbol{\tau\Omega\tau} \\ pdf &: f(\boldsymbol{\Omega}|\eta) \propto |\boldsymbol{\Omega}|^{\eta-1} \; where \; \boldsymbol{\Omega} \in \; corr\ matrices \\ marginal &: \rho_{ij} \sim Beta(\eta - 1 + \frac{d}{2}, \eta - 1 + \frac{d}{2}) \; over \; (-1, 1) \end{aligned}}$$

(8.47a)
(8.47b)
(8.47c)

Wishart and inverse-Wishart

Prior checklist (**want this in a different colour/shading. Not sure whether to have it at the end or the beginning. It would be good to have ticks next to these points.**)

1. Distribution for covariances matrices.
2. By their nature the parameters of these matrices are constrained (the matrices must be positive definite - see below).
3. Mostly useful in hierarchical models, although can be used in non-hierarchical settings.

4. Conjugate to multivariate normal likelihoods.

5. Represent quite restrictive relationships; better to use LKJ in most applied settings.

6. Gamma distribution is a univariate special case of the inverse-Wishart.

Before the invention of the LKJ distribution, the most frequently used prior distributions for covariance matrices were the inverse-Wishart and the Wishart distributions, in order of importance. Whilst these distributions are now somewhat defunct, they are still encountered in the literature, and can be useful for finding pen-and-paper posterior distributions for simple problems due to their conjugacy with the multivariate normal.

We choose to carry on with our example from section 8.5.4, on the prevalence and covariance of a number of conditions under the banner of 'mental health'. Recall that we had assumed a multivariate normal for the vector $\boldsymbol{d}$ where each element gives the prevalence of a particular disorder:

$$\boldsymbol{d} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8.48}$$

We postpone until chapter 9 a discussion of the conjugacy properties of the inverse-Wishart, and Wishart, and instead focus our efforts on what it would mean to ascribe such priors for $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} \sim inv - Wishart\,(\boldsymbol{\Psi}, \nu) \tag{8.49}$$

We see that this distribution is parameterised by two parameters: $\boldsymbol{\Psi}$, a scale matrix; and ν, the degrees of freedom. The probability distribution here is complex, and does not serve a purpose in explaining the functioning of this distribution, so we leave it till the end.

For simplicity we illustrate the behaviour of the inverse-Wishart, by assuming an identity scale matrix $\boldsymbol{\Psi}$, which makes all variables exchangeable [21]. Instead of exactly drawing the distribution, we instead summarise it by drawing a large number of samples. In the present case, we consider a 4-dimensional distribution, and illustrate the behaviour as we vary ν.

Here we will introduce some of the ways to visualise these distributions, but refer the interested reader to [21] for a more thorough exposition of

Panels shows sample-based estimates of variance and correlation distributions for the inverse-Wishart distribution**Top:** $\nu = 5$, and **Bottom:** $\nu = 50$. In both cases, the scale matrix is 4-dimensional, and estimates are based on sample sizes of 4,000,000. **Add legends. Figure too big to keep in.**.

Panels shows sample-based estimates of variance and correlation distributions for the Wishart distribution **Top:** $\nu = 5$, and **Bottom:** $\nu = 50$. In both cases, the scale matrix is 4-dimensional, and estimates are based on sample sizes of 4,000,000. **Add legends. Figure too big to keep in.**.

ways to visualise covariance/correlation matrices (from which we borrow here a few suggestions).

We firstly see that as we increase ν, the mean variance decreases (compare top- and bottom-left panels). However, this change of ν also has other consequences - the distribution of the correlation ends up becoming more and more concentrated towards 0, meaning the matrices become more and more diagonal. Furthermore, we can see that for low values of ν, the joint distributions of (ρ_{12}, ρ_{23}) become concentrated at the corners of the distribution. This is very restrictive, and usually undesirable [21]. The problem when we increase ν is that whilst the joint distribution (ρ_{12}, ρ_{23}) starts to adopt a more desirable symmetry, there is too much density in the centre (compare this with the density plot of figure 8.30).

Even if we decompose $\boldsymbol{\Sigma}$ into scale vectors and a correlation matrix, it turns out that we still run into the same problem with the correlation distributions when we use the inverse-Wishart [21].

Instead we might choose to use the Wishart distribution as a prior:

$$\boldsymbol{\Sigma} \sim Wishart\,(\boldsymbol{\Psi}, \nu) \tag{8.50}$$

Although usually this distribution is more commonly used as a prior for the inverse of the correlation matrix, we can still describe its properties here. Again we assume that the scale matrix, $\boldsymbol{\Psi}$ is the identity, and examine the effects of changes in ν. As an aside, the inverse Wishart distribution is obtained by first drawing $\boldsymbol{X} \sim Wishart\,(\boldsymbol{\Psi}, \nu)$, then inverting $\boldsymbol{C} = \boldsymbol{X}^{-1}$ (hence the name *inverse*). The Wishart distribution also turns out to be the sampling distribution for covariance matrices, where the data have been generated by multivariate normals.

We see now that for low ν, the joint correlation distribution of (ρ_{12}, ρ_{23}) is fairly diffuse, and symmetrical around 0, although not as diffuse as for an LKJ - see figure 8.30. Also, at low ν, the marginal distribution of ρ_{12} looks similar to that of the uniform distribution over all allowable correlation matrices (again see figure 8.30). However, if we seek to increase the level of the variances by increasing ν, we see that the correlations become increasing concentrated around 0, which is restrictive.

Essentially, both distributions - the inverse-Wishart, and the Wishart - are too simple because they only depend on a single parameter ν. This parameter affects not only the width of allowable correlations, but also the shape of the distribution. Whilst the LKJ distribution similarly only depends on a single parameter, η, parameter has a much simpler (monotonic) effect on the correlation distribution. Also by decomposing $\boldsymbol{\Sigma}$ into a correlation matrix and scale vectors, it allows for simple separation of prior views on the correlation structure and the variances.

$\boxed{\textit{Inverse} - \textit{Wishart properties}:}$

$$pdf: f(\boldsymbol{X}|\boldsymbol{\Psi},\nu) = \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu d}{2}}\Gamma_\nu\left(\frac{\nu}{2}\right)}|\boldsymbol{X}|^{-\frac{\nu+d-1}{2}}e^{\frac{tr(\boldsymbol{\Psi}\boldsymbol{X}^{-1})}{2}} \tag{8.51a}$$

$$mean: \mathrm{E}(\boldsymbol{X}) = \frac{1}{\nu-d-1}\boldsymbol{\Psi},\ \nu > d+1 \tag{8.51b}$$

$\boxed{\textit{Wishart properties}:}$

$$pdf: f(\boldsymbol{X}|\boldsymbol{\Psi},\nu) = \frac{|\boldsymbol{X}|^{\frac{\nu-d-1}{2}}e^{\frac{-tr(\boldsymbol{\Psi}^{-1}\boldsymbol{X})}{2}}}{2^{\frac{\nu d}{2}}|\boldsymbol{\Psi}|^{\frac{\nu}{2}}\Gamma_\nu\left(\frac{\nu}{2}\right)} \tag{8.52a}$$

$$mean: \mathrm{E}(\boldsymbol{X}) = \nu\boldsymbol{\Psi} \tag{8.52b}$$

8.6 Chapter summary

The reader should now have a good idea as to the array of likelihood and prior distributions from which to choose for a particular circumstance, however this chapter was not meant to be all-encompassing. It is almost inevitable to come across situations where none of the distributions discussed are appropriate, but it is hoped that this chapter should provide some guidance as to where to look in the literature. It is also frequently the case that a suitable distribution can be constructed from the basic building blocks discussed.

One of the most important decisions when choosing a likelihood and prior is how much variability to allow. We have tried to stress the benefits of the robust versions of sampling distributions (for example the negative binomial to the poisson, and the t distribution to the normal), as well as those choices for priors (again the t distribution to the normal, and the LKJ to the Wishart distributions). However, how should one decide on *how* robust to make a model? We shall postpone some of this discussion until chapter 10, where we will see there is a simple, but very general methodology for making these decisions. However, it would be remiss not to mention in passing some of the costs of more robust models: they are more complex than their simple alternatives, and can significantly lengthen the time taken for modern computational samplers to run (see part IV). These costs must be balanced against the needs of the situation, although we advise to always err on the side of robustness.

We have in passing mentioned the property of *conjugacy*. This is a condition which makes a narrow range of posterior distributions calculable by hand, without the need to appeal to a computer for sampling. Whilst this narrowing necessarily limits the complexity, and realism, of models we can build, we shall see that, equipped with our knowledge of distributions, we can use these toy problems to gain a better understanding of Bayesian statistics. It will also allow us to be better informed when it comes to designing models that are more appropriate to a given situation.

8.7 Chapter outcomes

The reader should now be familiar with the following concepts:

1. The various forms of sampling distribution that can be used, encompassing discrete and continuous probabilities, and multivariate outcomes.

2. The robust versions of sampling distributions to allow for greater sampling variability.

3. Prior distributions to use for:

Probabilities, proportions and percentages.

Means and regression coefficients.

Non-negative parameters for scale and shape.

(Extra): covariance and correlation matrices.

4. The robust versions of priors to allow for a wider possible range of parameter values.

8.8 Problem set

8.8.1 Drug trials

We suppose that we are testing the efficacy of a certain drug which aims to cure depression, across two groups, each of size 10, with varying levels of the underlying condition: *mild* and *severe*. We suppose that the success rate of the drug varies across each of the groups, with $\theta_{mild} > \theta_{severe}$. We are comparing this with another group of 10 individuals, which has a success rate equal to the mean of the other two groups $\theta = \frac{\theta_{mild}+\theta_{severe}}{2}$.

1. Calculate the mean number of successful trials in each of the three groups.

2. Compare the mean across the two heterogeneous groups, with that of the single group of 10 homogeneous people.

3. Calculate the variance of outcomes across each of the three groups.

4. How does the variance across the combination of both heterogeneous studies compare with that of an equivalent 10-person homogeneous group?

No consider the extension to a large number of trials, with the depressive status of each group unknown to the experimenter, but follows $\theta \sim Beta(\alpha, \beta)$.

1. Calculate the mean value of the Beta distribution.

2. What combinations of α and β would make the mean the same as that of a single study with success probability θ? How does the variance change, as the parameters of the beta distribution are changed, so as to keep the same mean of θ?

3. How does the variance of the number of disease cases compare to that of the a single study with success probability θ?

8.8.2 100m results across countries

We suppose that we would like to build a model for the 100m times for the fastest 10 runners across a number of different countries. We suppose that, neglecting Usain Bolt, the times for the runners in each country can be assumed to be approximately normal around a mean of 10 seconds. `http://www.johndcook.com/t_normal_mixture.pdf`

8.8.3 Triangular representation of simplexes

Show that the triangular representation of a 3 parameter Dirichlet is correct.

Hint: consider the points at particular values of the remaining probability for $p_1 + p_2 + p3 = 1$.

8.8.4 Normal distribution with normal prior

Prove that the mean of the lower distribution is the same as that of the prior

Find the mean of the lower distribution

Prove that the distribution of the lower distribution is unconditionally normal

Chapter 9

Conjugate priors and their place in Bayesian analysis

9.1 Chapter mission statement

At the end of this chapter the reader should understand what is meant by a conjugate prior, and how they can be used to exactly derive posterior distributions for particular circumstances.

9.2 Chapter goals

Bayesian analysis requires us to evaluate the denominator of Bayes' rule in order to exactly derive the posterior. For realistic models of most phenomena, this is usually a bridge too far, since we need to calculate a highly-dimensional integral, which is practically-intractable. However, there are a class of models - pairs of likelihoods and priors - where this calculation is possible. Furthermore, researchers have tabulated the formulae for the particular posteriors in these circumstances, which we can use to avoid having to actually calculate the denominators! The choice of priors that result in these cases, are specific to the particular choice of likelihood, and are called *conjugate*. Through this judicious choice, the resultant posteriors in these circumstances are actually within the same family of distributions as the priors themselves, making it even easier to remember and use this class of models.

Whilst the use of conjugate priors is usually overly-restrictive, they can nevertheless be a useful base from which to start an analysis, before moving onto more realistic models which require computational sampling to avoid having to actually calculate the posterior (see part IV). Additionally, since many texts and research literature assumes a basic understanding of conjugate priors, it is important to be familiar with their use.

This chapter is deliberately kept short for three reasons. Firstly, I believe that a few indicative examples can provide sufficient insight into the workings of conjugate priors; specifically how we can construct conjugate priors, and how we can use these to derive the algebraic form of the posterior parameters. Secondly, the whole point of using conjugate priors is that you don't need to do the maths - so why learn it! Thirdly, as we shall see the use of conjugate priors is very restrictive, and quite rightly now becoming less important with new computational methods.

9.3 What is a conjugate prior and why are they useful?

Suppose you run a restaurant and are interested in building a model for the number of people, X, who have food allergies within a particular sitting, in order to inform your buying of ingredients. If we make the assumption that the allergy-status of a given person is independent of everyone else's (see section 5.6.2), then we might choose a binomial sampling model $X \sim B(n, \theta)$, where n is the number of people eating, and θ is the probability that a randomly-chosen individual has an allergy. We can write the binomial likelihood as:

$$\begin{aligned} p(X = k|\theta, n) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \\ &\propto \theta^k(1-\theta)^{n-k} \end{aligned} \tag{9.1}$$

We choose to use a beta prior to represent our pre-data knowledge about the proportion of people that have an allergy to a miscellaneous type of model due to its flexibility, and since it is defined only over the [0,1] interval (see section 8.5.1). This distribution can be written as:

$$
\begin{aligned}
p(\theta|\alpha,\beta) &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}
\end{aligned}
\tag{9.2}
$$

where $B(\alpha,\beta)$ is a beta function, which is not dependent on θ.

We notice a similarity in the expressions for the sampling distribution, and the prior, in that they both contain a term $\theta^a(1-\theta)^b$. When we use Bayes' rule to calculate the posterior, we are required (in the numerator) to multiply together the likelihood and the prior, resulting in:

$$
\begin{aligned}
p(\theta|data) &\propto p(data|theta) \times p(\theta) \\
&\propto \theta^k(1-\theta)^{n-k} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \\
&= \theta^{\alpha'-1}(1-\theta)^{\beta'-1}
\end{aligned}
\tag{9.3}
$$

where $\alpha' = \alpha + k$ and $\beta' = \beta + n - k$. Notice that the posterior in this case is proportional to a term containing θ which is of exactly the same form as the beta prior. Furthermore, since the prior is a proper probability distribution, the posterior must also be one. This means that, because the posterior has a dependence on θ that looks like a beta distribution, it must actually be one! Put another way, if you go through and actually carry out the calculation, you will find that the normalising factor of the posterior is $B(\alpha',\beta')$, meaning that the functional form of the posterior is a beta distribution, although with different parameters to the prior. Here we say that the beta prior is *conjugate* to the binomial likelihood, since the posterior is also a beta distribution.

Ok, let's step away from the maths for a minute, and examine what happens to the beta posterior as we collect samples with differing numbers of people with allergies (see the left hand panels of figure 9.1). As we obtain a higher number of individuals with allergies within our sample, k, this leads to a shift rightwards in the likelihood, which is mirrored by $Beta(\alpha + k, \beta + n - k)$ posteriors. Alternatively, if we hold the number of people with allergies in our sample constant, and alter the priors, we again see that the $Beta(\alpha + k, \beta + n - k)$ posteriors are moved to reflect the priors.

So we have seen that using binomial likelihood with a beta prior leads to beta posterior. We have also noted that the behaviour of the posterior described by the formula we obtained behaves as we would expect it to

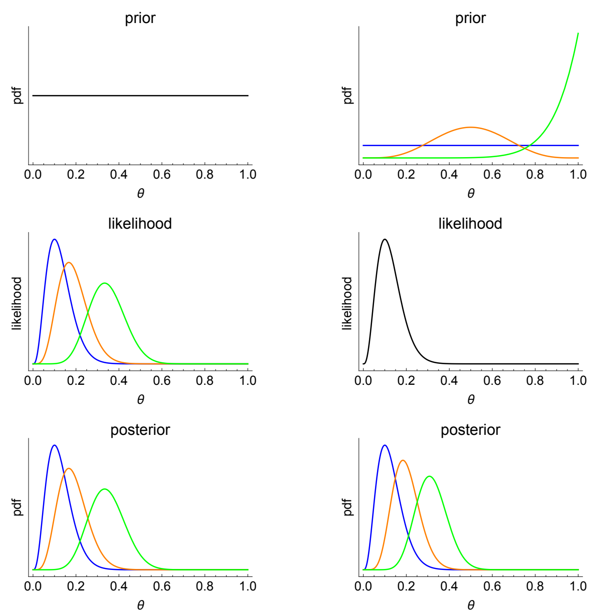


Figure 9.1: The priors, likelihoods and posteriors for across left: samples with different numbers of people with allergies, and right: the same sample with different priors. **Add legends. Left assumes a beta(1,1) prior, with samples of (3,5,10) (blue, orange, green) out of 30 with allergies. Right assumes a sample of 3/30 was obtained, with priors given by (1,1), (5,5), (10,1), (blue, orange, green). We should actually label the posterior curves with their posteriors, since this is the point of this diagram. Left: Beta(4,28), Beta(6,26), Beta(11,21). Right: Beta(4,28), Beta(8,32), Beta(13,28) with colours (blue, orange, green).**

visually. However, we have not yet adequately defined what it means to be a conjugate prior. We can think about the property of conjugacy through the flow diagram below, which represents the Bayesian inference process:

$$prior \xrightarrow{likelihood} posterior \tag{9.4}$$

In our example we have found the following flow diagram:

$$beta \xrightarrow{binomial} beta' \tag{9.5}$$

In other words, we specified a beta prior, and the data (through the binomial likelihood) updated our beliefs, and resulted in a beta posterior (albeit with different parameters, hence why *beta′* is used above). Conjugate priors are always defined relative to a particular likelihood, and should (like our example) mean that both the prior and the posterior come from the same *family* of distributions. Diagrammatically, we have that for a specified likelihood L, and a prior distribution F:

$$F \xrightarrow{L} F' \tag{9.6}$$

For example if we feed a gamma prior into our Bayesian updating rule, we should get a gamma posterior out (as is the case for a poisson likelihood). If we feed a normal prior in, we should get a normal posterior out (for particular normal likelihoods) etc.

Furthermore, for all conjugate prior examples we can actually write down a mechanistic rule using our flow diagram for how we should update our prior beliefs in light of data. In our example above:

$$beta(\alpha, \beta) \xrightarrow{binomial\ \ likelihood\ with\ (n,k)} beta(\alpha + k, \beta + n - k) \tag{9.7}$$

This simple rule negates the need to actually do any maths, since we can just substitute our numbers into the above formula (which we can look up in table 9.6), to yield the posterior. This really is Bayesian statistics made easy!

9.4 Gamma-poisson example

As another example imagine that you work for the department for transport, and are interested in building a model that predicts the number of cars, Y, which approach a particular intersection over a certain period of the day. Here we might choose to use a poisson likelihood if we believe that the arrival of cars are independent of one another (see section 8.3.3). We can write down the likelihood in this case as:

$$\begin{aligned} p(Y = \boldsymbol{y}|\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{k!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} y_i} e^{-n\lambda}}{k!} \\ &\propto \lambda^{n\bar{y}} e^{-n\lambda} \end{aligned} \tag{9.8}$$

where $\boldsymbol{y}$ represents a data vector with the numbers of cars that approached the intersection over n samples, and we have used $n\bar{y} = \sum_{i=1}^{n} y_i$.

We might also choose to use a gamma prior to describe our prior preferences (see section 8.5.3), which we can write in the following form:

$$\begin{aligned} p(\lambda|\alpha,\beta) &= \frac{\beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \\ &\propto \lambda^{\alpha-1} e^{-\beta\lambda} \end{aligned} \tag{9.9}$$

Again we notice a common form in both the likelihood and prior expressions (this time $\lambda^{a} e^{-b\lambda}$), meaning that when we multiply the two together in the numerator of the Bayes' rule we obtain a posterior given by:

$$\begin{aligned} p(\lambda|k) &\propto \lambda^{n\bar{y}} e^{-n\lambda} \times \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \lambda^{n\bar{y}+\alpha-1} e^{-(\beta+n)\lambda} \\ &= \lambda^{\alpha'-1} e^{-\beta'\lambda} \end{aligned} \tag{9.10}$$

where $\alpha' = \alpha + n\bar{y}$, and $\beta' = \beta + n$. Again, we can use the argument that since the dependence on λ is of the same form in both the posterior and prior distributions, and since the prior is a proper probability distribution, the posterior must also be a gamma density. Writing down the updating rule for this example, we find that:

$$gamma(\alpha, \beta) \xrightarrow{poisson\ \ likelihood\ with\ (n, \bar{y})} gamma(\alpha + n\bar{y}, \beta + n) \tag{9.11}$$

Note that this rule is only correct for the form of gamma density parameterised as in (9.9). If we use a different parameterisation of the gamma prior we would need to adjust the rule accordingly.

Let us now apply this rule to the case of cars arriving at an intersection. Suppose that we have two data vectors representing the numbers of cars arriving at two different intersections over a particular period of time. We can use the above rule to calculate the posteriors in each of these cases (see the left hand plots of figure 9.2). We see that increases in the average number of cars in the sample, leads to an increase in α' resulting in a shift rightwards in the posterior, in accordance with the likelihood. Similarly, if we alter the prior parameters (α, β), this leads to changes in (α', β') for the posterior, resulting in a shift in the posterior (see the right hand panel of figure 9.2).

9.5 Normal example: extra

Suppose that we are modelling the average height of adult male giraffes in a particular nature reserve. Since the height of a giraffe is the result of a multitude of different factors - both genetic and environmental - we might choose to use a normal likelihood (see section 8.3.6). Furthermore, we assume that the heights of the individual giraffes in our sample, $\boldsymbol{z} = \{z_1, z_2, ..., z_n\}$ are independent of one another, resulting in a likelihood of the form:

$$\begin{aligned} p(\boldsymbol{z}|\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(z_i - \mu)^2}{2\sigma^2}\right) \\ &\propto \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - \mu)^2\right) \end{aligned} \tag{9.12}$$

If we now think about constructing a prior that will be conjugate to this

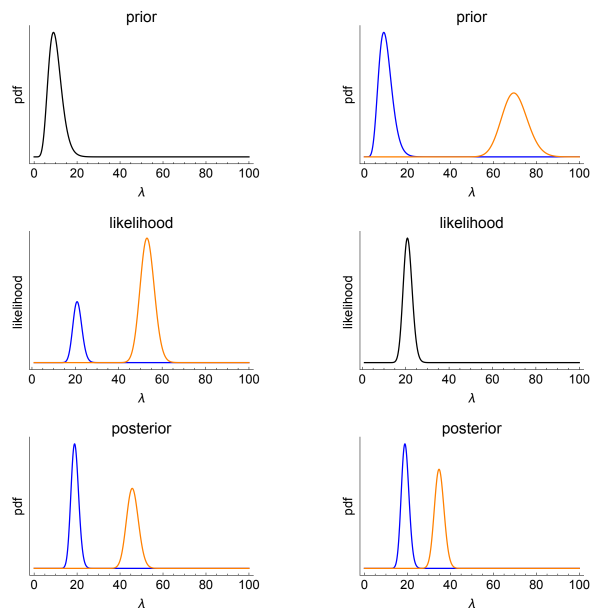


Figure 9.2: The effect of left: different samples, and right: different priors, on the posterior. **Add legends. LH prior has** $(\alpha, \beta) = (10, 1)$ **with samples of size 5, with means of (20.8,53) for (blue,orange). RH priors have (10,1) and (140,2) respectively for (blue,orange), with a sample of mean 20.8 and size 5. Add the labels to the posterior curves, importantly, with Gamma(114,1/6) and Gamma(275,1/6) for the left hand side (blue,orange), and Gamma(114,1/6) and Gamma(405,1/7), again for (blue,orange).**

likelihood, we know from our previous examples that its overall structure should be of the same form. This means that we require a prior distribution of something similar to the form:

$$p(\mu, \sigma^2) \propto (\sigma^2)^a \times exp\left(-\frac{1}{2\sigma^2}c(\mu - b)^2\right) \tag{9.13}$$

Notice there that we could separate the prior using the law of conditional probability into $p(\mu, \sigma^2) = p(\sigma^2) \times p(\mu|\sigma^2)$, where we have that:

$$p(\sigma^2) \propto (\sigma^2)^a \tag{9.14}$$

$$p(\mu|\sigma^2) \propto exp\left(-\frac{1}{2\sigma^2}c(\mu - b)^2\right) \tag{9.15}$$

This dependent-prior structure turns out often to be quite reasonable, where we want the prior variance in the mean σ^2 to be on the same scale as the sampling variation of heights z. These two relations can be satisfied using the following priors [5] (see sections 8.5.3 for a discussion of the first):

$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2) \tag{9.16}$$

$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0) \tag{9.17}$$

Gelman et al. (2013) label the overall prior specified by these assumptions as a $N - Inv - \chi^2(\mu_0, \sigma_0^2\kappa_0; \nu_0, \sigma_0^2)$ distribution. If we then multiply together the likelihood and the prior we obtain the numerator of Bayes' rule, which dictates the form of the posterior. We won't go into the maths here, but it is possible because of our construction of the prior to derive the parameters (see [5]) of the posterior $N - Inv - \chi^2$ distribution, and we obtain:

$$\mu' = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{z} \tag{9.18}$$

$$\kappa' = \kappa_0 + n \tag{9.19}$$

$$\nu' = \nu_0 + n \tag{9.20}$$

$$\nu'\sigma'^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{z} - \mu_0)^2 \tag{9.21}$$

Whilst the maths above is perhaps a tad bewildering, there is intuition in the results. The posterior mean giraffe height μ' is a weighted average of the prior mean, and the mean from the data, with more weight being given to the data the larger the sample size. The posterior variance parameter σ'^2 is a weighted average of the prior variance for σ^2, the data sample variance s^2, and a third term which represents the uncertainty described by the difference between the prior mean and the sample mean.

Note that here we have assumed that *both* parameters of the likelihood are unknown beforehand. In some circumstances, one of (μ, σ^2) may be known, in which case it is simpler to devise conjugate priors (see table 9.6) since we do not need to worry about the dependence between the two.

9.6 Table of conjugate priors

9.7 The lessons and limits of a conjugate analysis

Using conjugate priors means that there is no need to actually do any of the maths oneself, since we can stand on the shoulders of past giants, and use their tabulated results. All we need to do is use the relevant updating rule for our parameters of interest, and we are able to write down an exact form of the posterior distribution. We have also seen that if we follow these rules we end up with posterior parameters that are somewhere between those of the prior parameters, and values suggested from considering the likelihood. This is exactly as we expect, since Bayesian posteriors can be seen as a trade-off between prior expectations and those dictated by the data, manifest through maximum likelihood estimates.

Whilst the use of conjugate priors does make Bayesian statistics easy, it is limiting. These limits are quickly found to be inadequate when we need greater flexibility in our model, build models from a number of basic components, or are using hierarchical models (see part V).

As an example, imagine we are interested in evaluating the variance of the amount of iron, F, found in blood samples from a group of patients about a known mean μ (given the history of this test, this may be well-defined from the large number of experiments conducted). We suppose that we have a model of the form:

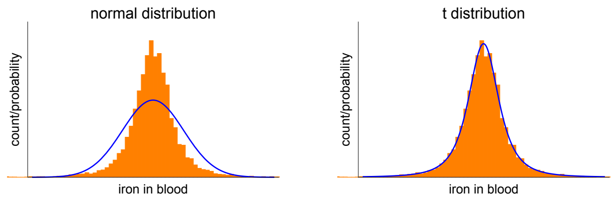


Figure 9.3: Fitting blood iron data (histogram), using left: a normal distribution, and right: a t distribution.

$$F_i = \mu + \epsilon_i \tag{9.22}$$

where $\epsilon_i \sim N(0, \sigma^2)$. Since we are assuming that the mean is known, we look up the relevant conjugate prior for a normal in the table in section 9.6, and we find it is a inverse-gamma distribution (see section 8.5.3). So we choose, $\sigma^2 \sim Inv - gamma(\alpha, \beta)$, as our prior, and we apply the rules in the table to find the posterior:

$$\sigma^2 | \boldsymbol{F} \sim Inv - gamma\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^{n} (F_i - \mu)^2}{2}\right) \tag{9.23}$$

where $\boldsymbol{F} = \{F_1, F_2, ..., F_n\}$ is a data vector of measurements of blood iron for n patients. This intuitively makes sense, since the mean of an inverse gamma distribution increases in β and decreases in α.

However, imagine that you examine your data sample and try to fit a normal distribution to it, and find that it is not adequately able to account for the variation you see in the sample (see figure 9.3). Instead you choose to use the more robust t distribution (see section 8.3.7), finding that it is much better fit to the data (see the right hand plot of figure 9.3).

As such, you go to the table in section 9.6, and look for the conjugate prior to a t distribution, where its mean is known. Alas! There is no corresponding entry. This means that simply by choosing a more appropriate sampling distribution, this has made it impossible to use a conjugate prior. This leaves you with a choice. Either you stay with a normal likelihood, which you know to be deficient, and wholly inappropriate, or you abandon the conjugate-ship, and use computational methods (see part IV). There is only one correct answer here - you should not let your analysis be determined by complexity, and you should punt for MCMC. Of course, you can use the normal distribution, with its inverse-gamma conjugate prior to gain some understanding of the problem, but using the t distribution ultimately will be much more satisfactory.

9.8 Chapter summary

The reader should now understand how choosing a *conjugate* prior can help us to do *easy* Bayesian inference. However, this simplicity comes at a cost! In most real life examples of inference, the constraints of choosing likelihood-prior conjugate pairs is too restrictive, and can lead us to use models that inadequately capture the degree of variability within the data. That is not to say that there aren't example circumstances where conjugate pairs are appropriate, it is just that we need to be sure to check the suitability of the modelling framework used before drawing any conclusions.

Bayesians happen to have at their hands powerful and flexible ways of checking the adequacy of a particular model. Whilst these have been alluded to in passing, it is now time to concentrate on this under-used, and perhaps most important aspect of Bayesian modelling.

9.9 Chapter outcomes

The reader should now be familiar with the following concepts:

1. The definition of a conjugate prior.
2. How to use the table in section 9.6 to find a conjugate prior for a particular likelihood, as well as write down (not calculate!) the posterior distribution.
3. The limits of conjugate priors, and the need for computational methods.

Chapter 10

Evaluation of model fit

10.1 The classical methodology

10.2 Posterior predictive checks

10.2.1 Graphical examples

10.2.2 Bayesian p values

10.2.3 A number of examples

10.3 Deviance, WAIC and LOO

See the two papers by Gelman on this subject. They are on my iPad. Suggests that deviance isn't a good way of evaluating model fit.

10.4 Sensitivity analysis

Chapter 11

Objective Bayesian analysis

11.1 The illusion of uniformed uniform prior

11.2 Jeffrey's priors

11.3 Reference priors

11.4 Zellner's g-priors

11.5 Empirical Bayes

11.6 A move towards weakly informative priors

Part IV

A practical guide to doing real life Bayesian analysis: Computational Bayes

Chapter 12

Discrete approximation of continuous posteriors

Unsure whether as it stands to keep this chapter, or lose it.

Chapter 13

Leaving the conjugates behind: Markov Chain Monte Carlo

13.1 The difficulty with real life Bayesian inference

13.2 Integrating using independent Monte Carlo samples

13.3 Moving from independent to dependent samples

Include subsection on effective sample size.

13.3.1 The burn-in/warm-up phase

13.4 Practical computational inference

13.4.1 The importance of pre-simulation MLE

13.4.2 Fake data simulation

Chapter 14

Gibbs sampling

14.1 The intuition behind the Gibbs algorithm

Imagine wanting to sample the heights of all the points on earth as a posterior distribution. One could do a random walk a la MH, but if the proposal distribution is too narrow, then the sampler could get stuck on a mountain range or desert; alternatively large swathes of landscape may be ignored if the step size is too large. This means it can take require a long burn-in, and may or may not be representative of the underlying landscape. The Gibbs algorithm works by cutting through the earth along lines of latitude or longitude, and allowing all steps along this conditional distribution. Since arbitrary step lengths are allowed, the algorithm is quicker, and there is no need to fine tune it to the degree of MH.

14.2 Simple examples

14.3 DAG models

14.4 The benefits and difficulties with Gibbs

Chapter 15

Metropolis-Hastings

15.1 Intuition behind Metropolis

The analogy of walking around a landscape. Moving to a new spot probabilistically if it is lower, and moving to a higher spot definitely. The idea is that you will visit positions at a rate proportional to their height. Rain analogy: absolute height is unimportant, only the relative difference between parts of the landscape.

15.2 The importance of the proposal distribution

The annoyance of having to pick a step length.

15.3 As a subset of Gibbs and vice versa

Chapter 16

Hamiltonian Monte Carlo

16.1 Introduction to Hamiltonian Monte Carlo

16.2 Avoiding manual labour: the No-U-turn sampler

Reference Hoffman and Gelman 2014 paper.

16.3 Riemannian MCMC

Reference downloaded Girolami article. Remember the comment from Ben Goodrich here: `https://groups.google.com/forum/?fromgroups=#!topic/stan-users/0F0O4hfHA8g` - in other words, using the second derivative is not free computationally! Ideal to design programs that don't need to use RMHMC.

Chapter 17

Stan and JAGS

17.1 Motivation for Stan and JAGS

17.1.1 Their similarities and differences

17.1.2 The future

17.1.3 The nuts and bolts of Stan and JAGS: access through R

How to get packages

How to package data

17.2 Stan

17.2.1 How to download, and install

17.2.2 A first program in Stan

Sample from a simple prior. Sample from a hierarchical prior system.

17.2.3 The building blocks of a Stan program

Include table showing when, and how often each block of the program is called.

Parameters

Model

Transformed parameters

Transformed data

Generated quantities

Sample from posteriors. Very useful to build in PPCs directly, rather than doing it post-hoc.

17.2.4 A simple model with data

17.2.5 Diagnostics

$\hat{R}$ and effective sample size

Divergent iterations

Tree depth exceeding maximum

Leapfrog steps

17.2.6 More complex models with array indexing

17.2.7 Essential Stan reading

The idea behind this section is to introduce the reader to a wide, and somewhat unconnected, range of tricks that I have found very useful in writing Stan programs. It also will contain some common error messages which I have come across; illustrating their meaning, and ways to alleviate the problem (if it *really* is a problem).

The importance of log probability incrementing

Marginalising of discrete parameters

If model has many discrete parameters consider moving to JAGS.

17.3 JAGS

17.3.1 How to download, and install

17.3.2 A first program in JAGS

Sample from a simple prior. Sample from a hierarchical prior system.

17.3.3 The literal meaning of a for loop

17.3.4 A simple model

17.3.5 Diagnosis

In-built using vanilla R

coda

More complex model

17.3.6 Essential JAGS reading

Controlling the flow without if conditions

How to control flow without explicit if using Step.

Storing values in matrices

Part V

Regression analysis and hierarchical models

17.4 Part mission statement

17.5 Part goals

Chapter 18

Hierarchical models

18.1 The spectrum from pooled to heterogeneous

18.1.1 The logic and benefits of partial pooling

18.1.2 Shrinkage towards the mean

Increases robustness naturally, and prevents overfitting.

18.2 Meta analysis example: simple

18.3 The importance of fake data simulation for complex models

18.3.1 The importance of making 'good' fake data

Chapter 19

Linear regression models

Include Gelman recommendations here about standardising your data beforehand: `https://groups.google.com/forum/?fromgroups=#!topic/stan-users/RMgEEFwCsQk`

Chapter 20

Generalised linear models

20.1 Malarial example of complex meta-analysis

Bibliography

[1]

[2] Joshua D Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

[3] William M Bolstad. *Introduction to Bayesian statistics*. John Wiley & Sons, 2007.

[4] Joshua M Epstein. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12, 2008.

[5] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[6] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[7] Jeff Gill. *Bayesian methods: A social and behavioral sciences approach*. CRC press, 2007.

[8] Anca Hanea. The asymptotic distribution of the determinant of a random correlation matrix. *arXiv preprint arXiv:1309.7268*, 2013.

[9] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[10] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[11] Benoit B Mandelbrot and Richard L Hudson. *Misbehaviour of Markets*. Profile Books, 2008.

[12] Sharon Bertsch McGrayne. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press, 2011.

[13] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Technische Universit at Wien, 2003.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[15] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[16] Nate Silver. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[17] Stan Development Team. Stan: A c++ library for probability and sampling, version 2.5.0, 2014.

[18] Wayne Stewart and Sepideh Stewart. Teaching markov chain monte carlo: Revealing the basic ideas behind the algorithm. *PRIMUS*, 24(1):25–45, 2014.

[19] Nassim Nicholas Taleb. *The black swan:: The impact of the highly improbable fragility*, volume 2. Random House, 2010.

[20] Max Tegmark. *Our mathematical universe: my quest for the ultimate nature of reality*. Knopf, 2014.

[21] Tomoki Tokuda, Ben Goodrich, I Van Mechelen, Andrew Gelman, and F Tuerlinckx. Visualizing distributions of covariance matrices. *Columbia Univ., New York, NY, USA, Tech. Rep*, 2011.