Shravan Vasishth, Bruno Nicenboim, and Daniel Schad

An Introduction to Bayesian Data Analysis for Cognitive Science

This book is dedicated to the thousands, perhaps millions, of psycholinguists and psychologists struggling to understand what their data is/are trying to tell them.

Contents

List of Tables

List of Figures

Preface

This book is a relatively gentle introduction to carrying out Bayesian data analysis and cognitive modeling using the probabilistic programming language Stan (Carpenter et al., 2017), and the front-end to Stan called `brms` (Bürkner, 2019). Our target audience is cognitive scientists (e.g., linguists and psychologists) who carry out behavioral experiments, and who are interested in learning the Bayesian data analysis methodology from the ground up and in a principled manner. Our aim is to make Bayesian statistics a standard part of the data analysis toolkit for experimental linguistics, psycholinguistics, psychology, and related disciplines.

Many excellent introductory textbooks exist already for Bayesian data analysis. Why write yet another book? Our text is different from other attempts in two respects. First, our main focus is on showing how to analyze data from planned experiments involving repeated measures; this type of experimental data involves unique complexities. We provide many examples of data-sets involving eyetracking (visual world and reading), self-paced reading, event-related potentials, reaction time, acceptability rating judgements, speeded grammaticality judgements, and question-response accuracies. Second, from the very outset, we stress a particular workflow that has as its centerpiece simulating data; we aim to teach a philosophy that involves thinking hard about the assumed underlying generative process, **even before the data are collected**. The data analysis approach that we hope to teach through this book involves a cycle of prior predictive and posterior predictive checks, and model validation using simulated data. We try to inculcate a sense of how inferences can be drawn from the posterior distribution of theoretically interesting parameters without resorting to binary decisions like “significant” or “not-significant”. We are hopeful that this will set a new standard for reporting results of data analyses in a more nuanced manner, and lead to more measured claims in the published literature.

0.1 Prerequisites

Any rigorous introduction to Bayesian data analysis requires at least a passive knowledge of probability theory, calculus, and linear algebra. We do not require that the reader already has this background when they start the book. Instead, the relevant ideas are introduced informally just in time, as soon as they are needed. The reader is never required to have an active ability to solve probability problems, to solve integrals or compute derivatives, or to carry out matrix computations by hand. What we do expect is some relatively simple high school arithmetic and algebra; a quick look through chapter 1 of Gill (2006) before starting this book is highly recommended. We also expect that the reader is willing to learn enough of the programming language R (R Core Team, 2019) and Stan/brms to reproduce the examples presented. For newcomers to R, we provide a quick introduction in the appendix that covers all the constructs used in the book. There are many good online resources on R that the reader can consult. Examples are: R for data science[1], and Efficient R programming[2].

We also assume that the reader is familiar with basic frequentist data analysis methodology; in particular, the reader should know how to carry out one and two sample t-tests, both paired and unpaired, and should know how to interpret the t-score and p-value that are computed from such tests. We remind the reader of these basic ideas in chapter 1, but we don't use up too much space in this book on comparing frequentist and Bayesian methods. We do not try to convince the reader to use the Bayesian approach over the frequentist one; our goal is to focus on the *what* and the *how* of Bayesian data analysis, and not the *why*. Other books and articles discuss the latter aspect in detail; for example, Kruschke (2014) compares frequentist and Bayesian methods in detail.

note

provide comprehensive book recommendations

[1] https://r4ds.had.co.nz/

[2] https://csgillespie.github.io/efficientR/

0.2 How to read this book

The chapters in this book are intended to be read in sequence, but during the first pass through the book, the reader should feel free to completely skip the sections marked with an asterisk. These sections provide a more formal development that will be useful when the reader transitions to more advanced textbooks like Gelman et al. (2014).

to-do: add a Mackay type chapter ordering for different scenarios.

0.3 Online materials

The entire book, including all data and source code, is available online for free on `https://github.com/vasishth/Bayes_CogSci`. The solutions to exercises are provided there under the directory solutions.

to-do: provide solutions

0.4 Software needed

Before you start, please install

- R[3] (and RStudio[4], or any other Integrated Development Environment that you prefer)

[3] `https://cran.r-project.org/`

[4] `https://www.rstudio.com/`

- The R package `rstan` (please pay close attention to the installation instructions!):
  - Instructions for Windows[5]
  - Instructions for Mac or Linux[6]
- The R packages `MASS`, `dplyr`, `purrr`, `readr`, `extraDistr`, `ggplot2`, `brms`, and `bayesplot`:
  - They can be installed in the usual way: `install.packages(c("MASS", "dplyr", "purrr", "readr", "extraDistr", "ggplot2", "brms", "bayesplot"))`.

In every R session, we'll need to set a seed (this ensures that the random numbers are always the same when we re-run our code).

```r
set.seed(42)
library(MASS)
##be careful to load dplyr after MASS
library(dplyr)
library(purrr)
library(readr)
library(extraDistr)
library(ggplot2)
library(brms)
library(rstan)
## Save compiled models:
rstan_options(auto_write = TRUE)
## Parallelize the chains using all the cores:
options(mc.cores = parallel::detectCores())
library(bayesplot)
```

[5] https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Windows

[6] https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Mac-or-Linux

0.5 Acknowledgements

We are grateful to the many generations of students at the University of Potsdam, various summer schools at ESSLLI, the LOT winter school, other short courses we have taught at various institutions, and the annual summer school on Statistical Methods for Linguistics and Psychology (SMLP). The participants in these courses helped us considerably in improving the material presented here. We are also grateful to members of Vasishth lab for comments on earlier drafts of this book.

Vasishth acknowledges the University of Potsdam for granting a sabbatical semester during 2019-20, and the Zentrum für Interdisziplinäre Forschung (ZiF) at the University of Bielefeld, Germany, for providing time for writing during Septemer 2019; this stay at ZiF was part of the activities of the research group *Statistical Models for Psychological and Linguistic Data* (led by Reinhold Kliegl, Douglas Bates, and Harald Baayen).

Shravan Vasishth, Bruno Nicenboim, Daniel Schad, Potsdam, Germany

About the Authors

Shravan Vasishth (http://vasishth.github.io) is professor of Psycholinguistics at the University of Potsdam, Germany. After completing his Bachelors degree in Japanese from Jawaharlal Nehru University, New Delhi, India, he spent five years in Osaka, Japan, studying Japanese and then working as a translator in a patent law firm in Osaka. He completed an MS in Computer and Information Science and a PhD in Linguistics from the Ohio State University, Columbus, USA, and an MSc in Statistics from the School of Mathematics and Statistics, University of Sheffield, UK. His research is in the area of computational modeling of human sentence comprehension processes. He has published n articles and k books, with n and k both greater than one.

Bruno Nicenboim (http://www.ling.uni-potsdam.de/~nicenboim/) is a postdoctoral researcher at the University of Potsdam, Germany. He started studying Electronic Engineering in the National University of Rosario, Argentina, then transitioned to Human Sciences and spent eight years in Israel where he completed a Bachelors degree in Sociology and Linguistics and Masters degree in Linguistics in Tel Aviv University. During this time, he also worked in several IT companies. He is currently in Germany, where he completed a PhD in Cognitive Science in the University of Potsdam. His research interests are sentence comprehension, memory processes, decision making, and predictions.

Daniel Schad (https://danielschad.github.io/) is a postdoctoral researcher at the University of Potsdam, Germany. His research interests are… to-do

1

Introduction

The central idea we will explore in this book is: given some data, how to use Bayes' theorem to quantify uncertainty about our belief regarding a scientific question of interest. Before we get into the details of the underlying theory and its application, some familiarity with the following topics needs to be in place: the basic concepts behind probability, the concept of random variables, probability distributions, and the concept of likelihood. We therefore turn to these topics first.

1.1 Probability

Informally, we all understand what the term "probability" means. We routinely talk about things like the probability of it raining today. However, there are two distinct ways to think about probability. One can think of the probability of an event with reference to the frequency with which it might occur in repeated observations. Such a conception of probability is easy to imagine in cases where an event can, at least in principle, occur repeatedly. An example would be obtaining a 6 when tossing a die again and again. However, this frequentist view of probability is difficult to justify when talking about certain one-of-a-kind events, such as earthquakes. In such situations, probability is expressing our uncertainty about the event happening. Moreover, we could even be uncertain about exactly how probable the event in question is; for example, we might say something like "I am 90% certain that the probability of an earthquake happening in the next year is between 10 and 40%". In this book, we will be particularly interested in quantifying uncertainty in this way: we will always want to know how unsure we are of the estimate we are interested in.

Both the frequency-based and the uncertain-belief perspective have their place in statistical inference, and depending on the situation, we are going to rely on both ways of thinking. Regardless of these differences in perspective, the probability of an event happening is defined to be constrained in the following way.

- The probability of an event must lie between 0 and 1, where 0 means that the event is impossible and cannot happen, and 1 means that the event is certain to happen.
- For any two mutually exclusive events, the probability that one or the other occurs is the sum of their individual probabilities.
- Two events are independent if and only if the probability of both events happening is equal to the product of the probabilities of each event happening.
- The probabilities of all possible events in the entire sample space must sum up to 1.

The above defitions are based on the axiomatic definition of probability by Kolmogorov (2018).

In the context of data analysis, we will talk about probability in the following way. Consider some data that you might have collected. This could be discrete 0,1 responses in a question-response accuracy task, or continuous measurements of reading times in milliseconds from an eyetracking study, etc. In any such cases, we will say that the data are being generated from a so-called **random variable**, which we will designate with a capital letter such as Y.[1]

The actually observed data will be distinguished from the random variable that generated it by using lower case y. We can call y an instance of Y; every new set of data will be slightly different due to random variability.

[1] Here, we use Y, but we could have used any letter, such as $X, Z, ...$ Later on, in some situations we will use Greek letters like θ, μ, σ to represent a random variable; however, following onvention in statistics, we will always use lower-case Greek letters for these.

So what is a random variable? As a concrete example, consider an experiment where we ask participants to respond to 10 questions that can be can either have a correct or incorrect answer. We will say that the number of correct responses from a participant are instances of a random variable Y. Because only discrete responses are possible (the number of correct responses can be 0, 1, 2, ..., 10), this is an example of a **discrete random variable**.

This random variable will be assumed to have a parameter θ that represents the probability of producing a correct response. In statistics, given some observed data, typically our goal is to obtain an estimate of this parameter's true (unknown) value.

We will follow the convention that the actually observed number of correct responses is written as y, as opposed to the abstract random variable Y. As mentioned above, given that we have 10trials, y can have values 0,1,2,...,10.

This discrete random variable Y has associated with it a function called a **probability mass function** or PMF. This function, which is written $p(y)$, gives us the probability of obtaining each of these 11 possible correct responses. We will write that this PMF depends on, or is conditional on, a particular fixed but unknown value for θ; the PMF will be written $p(y|theta)$.

In frequentist approaches to data analysis, the observed data y are used to draw inferences about θ. A typical question that we ask in the frequentist paradigm is: does θ have a particular value θ_0? One can obtain estimates of the unknown value of θ from the observed data y, and then draw inferences about how different–or more precisely how far away–this estimate is from the hypothesized θ_0. This is the essence of null hypothesis significance testing. The conclusions from such a procedure are framed in terms of either rejecting the hypothesis that θ has value θ_0, or failing to reject this hypothesis. Here, rejecting the null hypothesis is the primary goal of the statistical test.

Bayesian data analysis begins with a different question. What is common to the frequentist paradigm is the assumption that the data are generated from a random variable Y and that there is a function $p(y|\theta)$ indexed by the parameter θ. Where the Bayesian approach diverges from the frequentist one is that the goal now is to express our uncertainty about θ. In other words, we treat the parameter θ itself as a random variable, which means that we assign a probability distribution $p(\theta)$ to this random variable. This distribution $p(\theta)$ is called the **prior distribution** on θ; such a distribution could express our belief about the probability of correct responses, before we observe any data.

In a later chapter, we will spend some time trying to understand how such a prior distribution can be defined for a range of different research problems.

Given such a prior distribution and some data y, the end-product of a Bayesian data analysis is the so-called **posterior distribution** of the parameter given the data: $p(\theta|y)$. This posterior distribution is the probability distribution of θ after conditioning on y, i.e., after the data has been observed and is therefore known. All our statistical inference is based on this posterior distribution of θ; we can even carry out hypothesis tests analogous to the frequentist one sketched above.

We already mentioned conditional probability above when discussing the probability of the data given some parameter θ, which we wrote as the PMF $p(y|\theta)$. Conditional probability is an important concept in Bayesian data analysis, not least because it allows us to derive Bayes' theorem. Let's look at the definition of conditional probability next.

1.2 Conditional probability

Suppose that A stands for some discrete event; an example would be "the streets are wet." Suppose also that B stands for some other discrete event; an example is "it has been raining." We can talk about the probability of the streets being wet given that it has rained; or more generally, the probability of A given that B has happened.

This kind of statement is written as $Prob(A|B)$ or more simply $P(A|B)$. This is the conditional probability of event A given B. Conditional probability is defined as follows.

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ where } P(B) > 0 \tag{1.1}$$

We can rearrange the above equation so that we can talk about the joint probability of both events A and B happening. This joint probability can be computed by first taking $P(B)$, the probability that event B (it has been raining) happens, and multipling this by the probability that A happens conditional on B, i.e., the probability that the streets are wet given it has been raining. This multiplication will give us $P(A,B)$, the joint probability of A and B, i.e., that it has been raining and that the streets are wet. We will write the above description as: $P(A,B) = P(A|B)P(B)$

to-do: include venn diagram?

Now, since the probability A and B happening is the same as the probability of B and A happening, i.e., since $P(B,A) = P(A,B)$, we can equate the expansions of these two terms:

$$P(A,B) = P(A|B)P(B) \text{ and } P(B,A) = P(B|A)P(A) \tag{1.2}$$

Equating the two expansions, we get:

$$P(A|B)P(B) = P(B|A)P(A) \tag{1.3}$$

to-do: do we need to spell this out more?

Dividing both sides by $P(B)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.4}$$

The above statement is Bayes' rule, and is the basis for all the statistical inference we will do in this book. For now, this is all the probability theory we need to know!

The next sections expand on the idea of a random variable, the probability distributions associated with the random variable, what it means to specify a prior distribution on a parameter, and how the prior and data can be used to derive the posterior distribution of θ.

To make the discussion concrete, we will use an example of a discrete random variable, the Binomial. Aftere discussing this discrete random variable, we present anotheer example, this time involving a continuous random variable, the Normal random variable.

The Binomial and Normal cases serve as the canonical examples that we will need in the initial stages of this book. We will introduce other random variables as needed: the Uniform, Beta, Poisson, Gamma, and the Exponential, among others. The properties of all the distributions we will eventually need are summarized in the appendix.

1.3 Discrete random variables: An example using the Binomial distribution

Consider the following sentence:

"It's raining, I'm going to take the ...."

Suppose that our research goal is to estimate the probability, call it θ, of the word "umbrella" appearing in this sentence, versus any other word. If the sentence is completed with the word "umbrella", we will refer to it as a success; any other completion will be referred to as a failure. This is an example of a Binomial random variable: there can be only two possible outcomes, a success or a failure, and there is some true unknown probability θ of success that we want to estimate.[2]

One way to empirically estimate this probability of success is to carry out a so-called cloze task. In a cloze task, participants are asked to complete a

[2] Technically, each single trial that can result in either a success or failure is called a Bernoulli random variable—but this random variable is just a special case of the Binomial when the number of trials is 1.

fragment of the original sentence, such as "It's raining, I'm going to take the …". The predictability or cloze probability of "umbrella" is then calculated as the proportion of times that the target word "umbrella" was produced as an answer by participants.

Assume for simplicity that 10 participants are asked to complete the above sentence; each participant does this task only once. This gives us independent responses from 10 trials that are either coded a success ("umbrella" was produced) or as a failure (some other word was produced). We can sum up the number of sucesses to calculate how many of the 10 trials had "umbrella" as a response. For example, if 8 instances of "umbrella" are produced in 10 trials, we would estimate the cloze probability of producing "umbrella" would be $8/10$.

We can repeatedly generate simulated sequences of the number of successes in R (later on we will demonstrate how to generate such random sequences of simulated data). Here is a case where we run the same experiment 20 times (the sample size is 10 each time).

```r
rbinom(10,n=20,prob=0.5)
```

```
##  [1] 7 7 4 7 6 5 6 3 6 6 5 6 7 4 5 7 8 3 5 5
```

The number of successes in each of the 20 simulated experiments above is being generated by a discrete random variable Y with a probability distribution $p(y|\theta)$ called the **Binomial distribution**.

For discrete random variables such as the Binomial, the probability distribution $p(y|\theta)$ is called a **probability mass function** (PMF). The PMF defines the probability of each possible outcome. In the above example, with $n = 10$ trials, there are 11 possible outcomes: $y = 0, 1, 2, ..., 10$ successes. Which of these outcomes is most probable depends on the parameter θ in the Binomial distribution that represents the probability of success.

The left-hand side plot in Figure 1.1 shows an example of a Binomial PMF with 10 trials, with the parameter θ fixed at 0.5. Setting θ to 0.5 leads to a PMF where the most probable outcome is 5 successes out of 10. If we had set θ to, say 0.1, then the most probable outcome would be 1 success out of 10; and if we had set θ to 0.9, then the most probable outcome would be 9 successes out of 10.

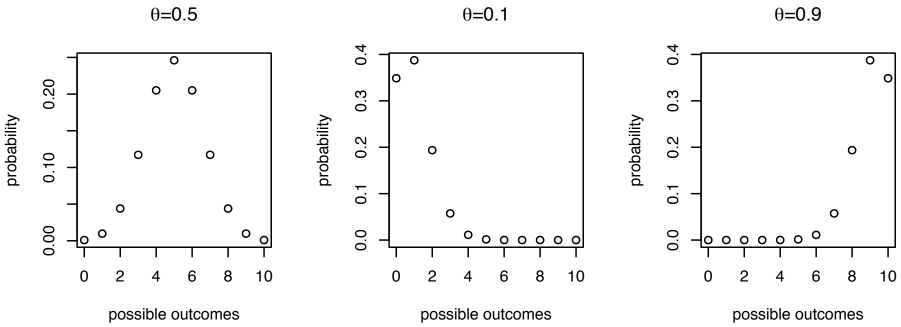

FIGURE 1.1: Probability mass functions of a binomial distribution assuming 10 trials, with 50%, 10%, and 90% probability of success.

to-do bar or line graphs above, instead of points

The probability mass function for the binomial is written as follows.

$$\text{Binomial}(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{1.5}$$

Here, n represents the total number of trials, k the number of successes (this could range from 0 to 10), and θ the probability of success. The term $\binom{n}{k}$, pronounced n-choose-k, represents the number of ways in which one can choose k successes out of n trials. For example, 1 success out of 10 can occur in 10 possible ways: the very first trial could be a 1, the secone trial could be a 1, etc. The term $\binom{n}{k}$ expands to $\frac{n!}{k!(n-k)!}$. In R, it is computed using the function `choose(n,k)`, with n and k representing positive integer values.

When we want to express the fact that the data is assumed to be generated from a Binomial random variable, we will write $Y \sim Binomial(n,\theta)$, where $\sim$ should be read as "is being generated from". If the data is generated from a random variable that has some other probability distribution $f(\theta)$, we will write $Y \sim f(\theta)$.

1.3.1 The mean and variance of the Binomial distribution

It is possible to analytically compute the mean and variance of the PMF associated with the Binomial random variable Y. Without getting into the details of how these are derived mathematically, we just state here that the mean of Y (also called the expectation, conventionally written $E[Y]$) and variance of Y (written $Var(Y)$) of a Binomial distribution with parameter θ and n trials are $E[Y] = n\theta$ and $Var(Y) = n\theta(1-\theta)$.

Of course, n is a fixed number because we decide on the total number of trials before running the experiment. In the PMF θ is also a fixed value; the only variable in a PMF is k. In real experimental situations we never know the true value of θ. But θ can be estimated from the data. From the observed data, we can compute the estimate of θ, $\hat{\theta} = k/n$. The quantity $\hat{\theta}$ is the observed proportion of successes, and is called the **maximum likelihood estimate** of the true (but unknown) expectation E[Y]. Once we have estimated θ in this way, we can also obtain an estimate (also a maximum likelihood estimate) of the variance by computing $n\theta(1-\theta)$. These estimates are then used for statistical inference.

What does the term "maximum likelihood estimate" mean? The term **likelihood** refers to the Binomial distribution function, i.e., the PMF we saw above, $p(k|n,\theta)$. Recall that the PMF assumes that θ and n are fixed, and k will vary from 0 to 10 when the experiment is repeated multiple times. The likelihood function is the same function as the PMF, $p(k|n,\theta)$, but assumes that the data is fixed and only θ varies (from 0 to 1).

For example, suppose you record $n = 10$ trials, and observe $k = 7$ successes. What is the probability of observing 7 successes out of 10? We need the Binomial distribution to compute this value:

$$\text{Binomial}(k = 7|n = 10, \theta) = \binom{10}{7}\theta^7(1-\theta)^{10-7} \tag{1.6}$$

Once we have observed the data (k=7 successes), both n and k are fixed. The only variable in the above equation now is θ: the above function is now only dependent on the value of θ.

When the data are fixed, the probability mass function is only dependent on the value of the parameter θ, and is called a **likelihood function**. It is therefore often expressed as a function of θ:

$$p(k = 7, n = 10|theta) = \mathcal{L}(\theta)$$

Since the PMF and the likelihood refer to the same function seen in two different ways, sometimes the likelihood is written $p(\theta|k = 7, n = 10)$ to distinguish it from the PMF, which has the data appearing first ($p(k|n, \theta)$). We will write both the PMF and the likelihood identically in this book; context will disambiguate what we are referring to.

If we now plot the likelihood function for all possible values of θ ranging from 0 to 1, we get the plot shown in Figure 1.2.

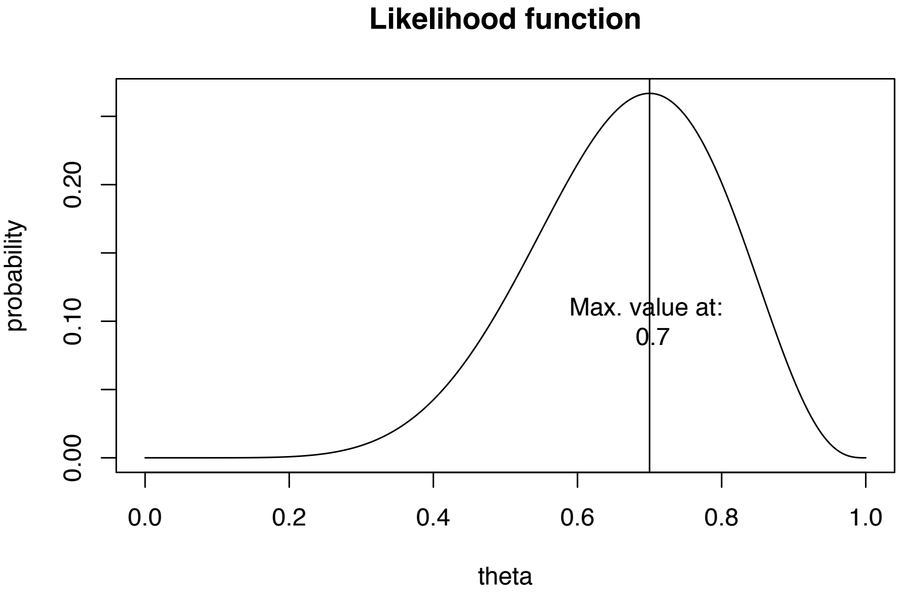

FIGURE 1.2: The likelihood function for 7 successes out of 10.

DS comment: do we want to show the code for computing all likelihood values? (maybe this comes later?)

What is important about this plot is that it shows that, given the data, the maximum point is at the point 0.7, which corresponds to the estimated mean using the formula shown above: $k/n = 7/10$. Thus, the maximum likelihood estimate (MLE) gives us the most likely value that the parameter θ has, given the data.

It is crucial to note here that the phrase "most likely" here does not mean that the MLE from a *particular* sample of data invariably gives us an accurate estimate of θ. For example, if we run our experiment for 10 trials and get 1 success out of 10, the MLE is 0.10. We could have happened to observe only one success out of ten even if the true θ were 0.5. The MLE would however give an accurate estimate of the true parameter as n approaches infinity.

1.3.2 What information does a probability distribution provide?

In Bayesian data analysis, we will constantly be asking the question: what information does a probability distribution give us? In particular, we will treat each parameter θ as a random variable; this will raise questions like: "what is the probability that the parameter θ lies between two values a and b;"what is the range over which we can be 95% certain that the true value of the parameter lies"? In order to be able to answer questions like these, we need to know what information we can obtain once we have a probability distribution, and how to extract this information. We therefore discuss the different kinds of information we can obtain from a probability distribution. For now we focus only on the Binomial random variable discussed above.

1.3.2.1 Compute the probability of a particular outcome (discrete case only)

The Binomial distribution shown in Figure 1.1 already shows the probability of each possible outcome under a different value for θ. In R, there is a built-in function that allows us to calculate the probability of k successes out of n, given a particular value of k (this number constitutes our data), the number of trials n, and given a particular value of θ; this is the `dbinom` function. For example, the probability of 5 successes out of 10 when θ is 0.5 is:

```
dbinom(5,size=10,prob=0.5)
```

```
## [1] 0.2461
```

The probabilities of success when θ is 0.1 or 0.9 can be computed by replacing 0.5 above by each of these probabilities. One can just do this by giving `dbinom` a vector of probabilities:

```
dbinom(5,size=10,prob=c(0.1,0.9))
```

```
## [1] 0.001488 0.001488
```

Note that the probability of a particular outcome is only computable in the discrete case; in the continuous case, this probability will always be zero (we discuss this when we turn to continuous probability distributions below).

1.3.2.2 Compute the cumulative probability of k or less (more) than k successes

Using the `dbinom` function, we can compute the cumulative probability of obtaining 1 or less, 2 or less successes etc. This is done through a simple summation procedure:

```
## the cumulative probability of obtaining
## 0, 1, or 2 successes out of 10,
## with theta=0.5:
dbinom(0,size=10,prob=0.5)+dbinom(1,size=10,prob=0.5)+
  dbinom(2,size=10,prob=0.5)
```

```
## [1] 0.05469
```

Mathematically, we could write the above summation as:

$$\sum_{k=0}^{2} \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{1.7}$$

An alternative to the cumbersome addition in the R code above is this more compact statement, which closely mimics the above mathematical expression:

```
sum(dbinom(0:2,size=10,prob=0.5))
```

```
## [1] 0.05469
```

R has a built-in function called `pbinom` that does this summation for us. If we want to know the probability of 2 or less successes as in the above example, we can write:

```
pbinom(2,size=10,prob=0.5,lower.tail=TRUE)
```

```
## [1] 0.05469
```

The specification `lower.tail=TRUE` ensures that the summation goes from 2 to numbers smaller than 2 (which lie in the lower tail of the distribution in Figure 1.1). If we wanted to know what the probability is of obtaining 2 or more successes out of 10, we can set `lower.tail` to `FALSE`:

```
pbinom(2,size=10,prob=0.5,lower.tail=FALSE)
```

```
## [1] 0.9453
```

The cumulative distribution function or CDF can be plotted by computing the cumulative probabilities for any value k or less than k, where k ranges from 0 to 10 in our running example. The CDF is shown in Figure 1.3.

1.3.2.3 Compute the inverse of the cumulative distribution function (the quantile function)

We can also find out the value of the variable k (the quantile) such that the probability of obtaining k or less than k successes is some specific probability value p. If we switch the x and y axes of Figure 1.3, we obtain another very useful function, the inverse CDF.

The inverse of the CDF (known as the quantile function in R because it returns the quantile, the value k) is available in R as the function `qbinom`.

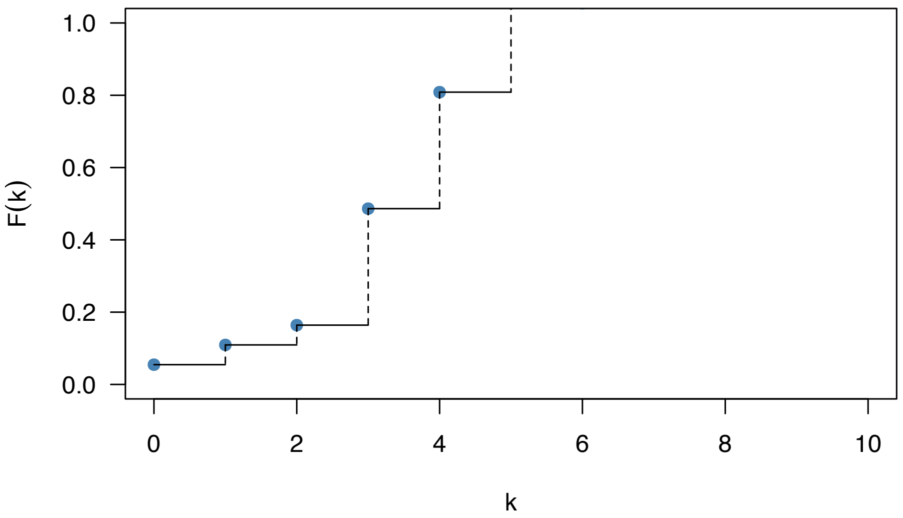

FIGURE 1.3: The cumulative distribution function for a Binomial distribution assuming 10 trials, with 50% probability of success.

The usage is as follows: to find out what the value k of the outcome is such that the probability of obtaining k or less successes is 0.37, type:

```r
qbinom(0.37,size=10,prob=0.5)
```

```
## [1] 4
```

to-do: explain why qbinom(0.77 gives 5 as an answer and not 4)

DS comment: maybe it's good to include an additional Figure for the inverse CDF and an example

1.3.2.4 Generate simulated data from a Binomial(n, θ) distribution

We can generate simulated data from a Binomial distribution by specifying the number of trials and the probability of success θ. In R, we do this as follows:

```
rbinom(1,size=10,prob=0.5)
```

```
## [1] 7
```

The above code generates the number of successes in an experiment with 10 trials. Repeatedly run the above code; you will get different sequences each time. For each generated sequence, one can calculate the number of successes by just summing up the vector, or computing its mean and multiplying by the number of trials, here 10:

```
y<-rbinom(10,size=1,prob=0.5)
mean(y)*10 ; sum(y)
```

```
## [1] 6
```

```
## [1] 6
```

1.4 Continuous random variables: An example using the Normal distribution

We will now revisit the idea of the random variable using a continuous distribution. Imagine that you have a vector of reading time data y measured in milliseconds and coming from a Normal distribution. The Normal distribution is defined in terms of two parameters: a mean value μ, which determines its center, and the variance σ^2, which determines how much spread there is around this center point.

The probability density function (PDF) of the Normal distribution is defined as follows:

$$Normal(y|\mu,\sigma) = f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{1.8}$$

Here, μ is some true, unknown mean, and σ^2 is some true, unknown variance of the Normal distribution that the reading times have been sampled from. There is a built-in function in R that computes the above function once we specify the mean μ and the standard deviation σ (in R, this parameter is specified in terms of the standard deviation rather than the variance).

Figure 1.4 visualizes the Normal distribution for particular values of μ and σ, as a PDF (using `dnorm`), a CDF (using `pnorm`), and the inverse CDF (using `qnorm`). It is clear from the figure that these are three different ways of looking at the same information.

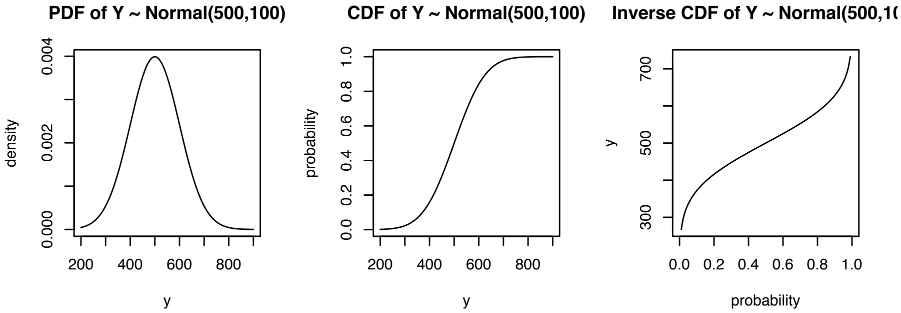

FIGURE 1.4: The PDF, CDF, and inverse CDF for the $Normal(\mu = 500, \sigma = 100)$.

to-do: Maybe this is the place to mention some interesting properties like: Normal(mu, sigma) = mu + Normal(0,1) * sigma (We'll use this property a lot later when we code in Stan)

As in the discrete example, the PDF, CDF, and inverse of the CDF allow us to ask questions like:

- **What is the probability of observing values between a and b from a Normal distribution with mean μ and standard deviation σ?** Using the above example, we can ask what the probability of observing values between 200 and 700 ms:

```
pnorm(700,mean=500,sd=100)-pnorm(200,mean=500,sd=100)
```

```
## [1] 0.9759
```

to-do: add figure illustrating the above

Notice here that the probability of any point value in a PDF is always 0. This is because the probability in a continuous probability distribution is the area under the curve, and the area at any point on the x-axis is always 0. The implication here is that we can only ask about probabilities between two different points; e.g., the probability that Y lies between a and b, or $P(a < Y < b)$.

- **What is the quantile q such that the probability is p of observing that value q or something less (or more) than it**? For example, we can work out the quantile q such that the probability of observing q or something less than it is 0.975, in the Normal(500,100) distribution. Formally, we would write this as $P(Y < a)$.

```
qnorm(0.975,mean=500,sd=100)
```

```
## [1] 696
```

The above output says that the probability that the random variable is less than $q = 695$ is 97.5%.

- **Generating simulated data**. Given a vector of n independent and identically distributed data y, i.e., given that each data point is being generated independently from $Y \sim Normal(\mu, \sigma)$ for some values of the parameters, the maximum likelihood estimates for the expectation and variance[3] are:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.9}$$

$$Var(y) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} \tag{1.10}$$

[3] R will compute variance by dividing by $n - 1$, not n; this is because dividing by n gives a biased estimate. This is not an important detail for our purposes, and in any case for large n it doesn't really matter whether one divides by n or $n - 1$.

For example, we can generate 10 data points using the `rnorm` function, and then compute the mean and variance from the simulated data:

```
y<-rnorm(10,mean=500,sd=100)
mean(y);var(y)
```

```
## [1] 559.6
```

```
## [1] 6778
```

Again, the sample mean and sample variance computed from a particular data-set need not necessarily be close to the true values of the respective parameters.

1.4.1 An important distinction: probability vs. density in a continuous random variable

In continuous distributions like the Normal discussed above, it is important to understand that the probability density function or PDF, $p(y|\mu,\sigma)$ defines a mapping from the y values (the possible values that the data can have) to a quantity called the density of each possible value. We can see this function in action when we use `dnorm` to compute, say, the density value corresponding to $y = 1$ in the $Normal(\mu = 0, \sigma = 1)$ distribution.

```
## density:
dnorm(1,mean=0,sd=1)
```

```
## [1] 0.242
```

The quantity above is *not* the probability of observing 1 in this distribution. As mentioned earlier, probability in a continuous distribution is the area under the curve, and this area will always be zero at any point value. If we want to know the probability of obtaining values between an upper and lower bound b and a, i.e., $P(a < Y < b)$ where these are two distinct values, we must use the `pnorm` function. For example, the probability of observing a value between +2 and -2 in a Normal distribution with mean 0 and standard deviation 1 is:

```
pnorm(2,mean=0,sd=1)-pnorm(-2,mean=0,sd=1)
```

```
## [1] 0.9545
```

Notice that the situation is different in discrete random variables. These have a probability mass function (PMF) associated with them—the Binomial distribution that we saw earlier is an example. There, the PMF maps the possible y values to the probabilities of those values occurring. That is why, in the Binomial distribution, the probability of observing exactly 2 successes when sampling from a $Binomial(n = 10, \theta = 0.5)$ can be computed using either `dbinom` or `pbinom`:

```
dbinom(2,size=10,prob=0.5)
```

```
## [1] 0.04395
```

```
pbinom(2,size=10,prob=0.5)-pbinom(1,size=10,prob=0.5)
```

```
## [1] 0.04395
```

In the second line of code above, we are computing the cumulative probability of observing two or less successes, minus the probability of observing one or less successes. This gives us the probability of observing exactly two successes. The `dbinom` gives us this same information.

1.5 An important concept: The marginal likelihood (integrating out a parameter)

Here, we introduce a concept that will turn up many times in this book. The concept we unpack here is called "integrating out a parameter". We will need this when we encounter Bayes' rule in the next chapter, and when we use Bayes factors later in the book.

Integrating out a parameter refers to the following situation. Suppose we have a Binomial random variable Y with PMF $p(Y)$. Suppose also that this PMF is defined in terms of parameter θ that can have only three possible values, $0.1, 0.5, 0.9$, each with equal probability. In other words, the probability that θ is $0.1, 0.5, or 0.9 is 1/3 each.

We stick with our earlier example of $n = 10$ trials and $k = 8$ successes. The **likelihood function** then is

$$p(k = 8|n = 10, \theta) = \binom{10}{8} \theta^8 (1 - \theta)^2 \tag{1.11}$$

There is a related concept of **marginal likelihood**, which we can write here as $p(k = 8, n = 10)$. Marginal likelihood is the likelihood computed by "marginalizing" out the parameter θ: for each possible value that the parameter θ can have, we compute the likelihood at that value and multiply that likelihood with the probability of that θ value occurring. Then we sum up each of the products computed in this way. Mathematically, this means that we carry out the following operation.

In our example, there are three possible values of θ, call them $\theta_1 = 0.1$, $\theta_2 = 0.5$, and $\theta_3 = 0.9$. Each has probability $1/3$; so $p(\theta_1) = p(\theta_2) = p(\theta_3) = 1/3$. Given this information, we can compute the marginal likelihood as follows:

$$\begin{aligned} p(k = 8, n = 10) = & \binom{10}{8} \theta_1^8 (1 - \theta_1)^2 \times p(\theta_1) \\ & + \binom{10}{8} \theta_2^8 (1 - \theta_2)^2 \times p(\theta_2) \\ & + \binom{10}{8} \theta_3^8 (1 - \theta_3)^2 \times p(\theta_3) \end{aligned} \tag{1.12}$$

Writing the θ values and their probabilities, we get:

$$
\begin{aligned}
p(k=8,n=10) =& \binom{10}{8} 0.1^8 (1-0.1)^2 \times \frac{1}{3} \\
&+ \binom{10}{8} \theta_2^8 (1-\theta_2)^2 \times \frac{1}{3} \\
&+ \binom{10}{8} \theta_3^8 (1-\theta_3)^2 \times \frac{1}{3}
\end{aligned}
\tag{1.13}
$$

We can simplify this summation by collecting together the common terms:

$$
\begin{aligned}
p(k=8,n=10) =& \frac{1}{3}[\binom{10}{8} 0.1^8 (1-0.1)^2 \\
&+ \binom{10}{8} \theta_2^8 (1-\theta_2)^2 \\
&+ \binom{10}{8} \theta_3^8 (1-\theta_3)^2] \\
=& 0.0582
\end{aligned}
\tag{1.14}
$$

Thus, a marginal likelihood is a kind of weighted sum of the likelihood, weighted by the possible values of the parameter.[4]

The above example was contrived, because we stated that the parameter θ has only three possible values. In reality, because the parameter θ can have all possible values between 0 and 1, the summation has to be done over a continuous space $[0,1]$. The way this summation is expressed in mathematics is through the integral symbol:

$$
p(k=8,n=10) = \int_0^1 \binom{10}{8} \theta^8 (1-\theta)^2 \, d\theta
\tag{1.15}
$$

This statement is saying exactly what we computed above, except that the summation is being done over a continuous space ranging from 0 to 1. We say that the parameter θ has been integrated out, or marginalized.

[4] Where does the above formula come from? It falls out from the law of total probability; see Blitzstein and Hwang (2014) for a detailed exposition.

Integrating out a parameter will be a very common operation in this book, but fortunately we will never have to do the calculation ourselves. For the above case, we can easily compute the integral in R:

```
BinLik<-function(theta){
  choose(10,8)*theta^8 * (1-theta)^2
}
integrate(BinLik,lower=0,upper=1)$value
```

```
## [1] 0.09091
```

This completes our discussion of random variables and probability distributions. We now summarize what we have learnt so far.

1.6 Summary of useful R functions relating to distributions

Table 1.1 summarizes the different functions relating to PMFs and PDFs, using the Binomial and Normal as examples.

TABLE 1.1: Important R functions relating to random variables.

	Discrete	Continuous
Example:	Binomial$(y\|n,\theta)$	Normal$(y\|\mu,\sigma)$
Likelihood function	dbinom	dnorm
Prob Y=y	dbinom	always 0
Prob $Y \geq y, Y \leq y, y_1 < Y < y_2$	pbinom	pnorm
Inverse CDF	qbinom	qnorm
Generate simulated data	rbinom	rnorm

Later on, we will use other distributions, such as the Uniform, Beta, etc., and each of these has their own set of d-p-q-r functions in R. The appendix summarizes the properties of the distributions that we will need in this book.

1.7 Summary of concepts introduced in this chapter

to-do: add summary

1.8 Further reading

A quick review of the mathematical foundations needed for statistics is available in the short book by Fox (2009). Morin (2016) and Blitzstein and Hwang (2014) are accessible introductions to probability theory.

1.9 Exercises

1.9.1 Practice using the `pnorm` function

1.9.1.1 Part 1

Given a normal distribution with mean 61 and standard deviation 101, use the pnorm function to calculate the probability of obtaining values between 217 and -95 from this distribution.

1.9.1.2 Part 2

Calculate the following probabilities. Given a normal distribution with mean 51 and standard deviation 4, what is the probability of getting

- a score of 48 or less
- a score of 48 or more
- a score of 56 or more

1.9.1.3 Part 3

Given a normal distribution with mean 53 and standard deviation 4, what is the probability of getting

- a score of 48 or less.
- a score between 50 and 56.
- a score of `mu+1` or more.

1.9.2 Practice using the `qnorm` function

1.9.2.1 Part 1

Consider a normal distribution with mean 1 and standard deviation 1.

Compute the lower and upper boundaries such that:

- the area (the probability) to the left of the lower boundary is 0.27.
- the area (the probability) to the left of the upper boundary is 0.91.

1.9.2.2 Part 2

Given a normal distribution with mean 56.932 and standard deviation 0.741. There exist two quantiles, the lower quantile q1 and the upper quantile q2, that are equidistant from the mean 56.932, such that the area under the curve of the Normal probability between q1 and q2 is 85%. Find q1 and q2.

1.9.3 Practice using `qt`

Take an independent random sample of size 144 from a normal distribution with mean 133, and standard deviation 54. Next, we are going to pretend we don't know the population parameters (the mean and standard deviation). We compute the MLEs of the mean and standard deviation using the data and get the sample mean 130.633 and the sample standard deviation 50.045.

- Compute the estimated standard error using the sample standard deviation provided above.
- What are your degrees of freedom for the relevant t-distribution?
- Calculate the **absolute** critical t-value for a 95% confidence interval using the relevant degrees of freedom you just wrote above.
- Next, compute the lower bound of the 95% confidence interval using the estimated standard error and the critical t-value.
- Finally, compute the upper bound of the 95% confidence interval using the estimated standard error and the critical t-value.

1.9.4 Maximum likelihood estimation 1

Given the data point `9.079`. The function `dnorm` gives the likelihood given a data point (or multiple data points) and a value for the mean and the standard deviation (sd). Using `dnorm`, compute

- the likelihood of the data point `9.079` assuming a mean of `12` and standard deviation 5.
- the likelihood of the data point `9.079` assuming a mean of `11` and standard deviation 5.
- the likelihood of the data point `9.079` assuming a mean of `10` and standard deviation 5.
- the likelihood of the data point `9.079` assuming a mean of `9` and standard deviation 5.

1.9.5 Maximum likelihood estimation 2

You are given 10 independent and identically distributed data points that are assumed to come from a Normal distribution with unknown mean and unknown standard deviation:

```
x
```

```
##  [1] 504 503 497 487 507 506 492 484 502 497
```

The function `dnorm` gives the likelihood given multiple data points and a value for the mean and the standard deviation. The log-likelihood can be computed by typing `dnorm(...,log=TRUE)`.

The product of the likelihoods for two independent data points can be computed like this: Suppose we have two independent and identically distributed data points 5 and 10. Then, assuming that the Normal distribution they come from has mean 10 and standard deviation 2, the joint likelihood of these is:

```
dnorm(5,mean=10,sd=2)*dnorm(10,mean=10,sd=2)
```

```
## [1] 0.001748
```

It is easier to do this on the log scale, because then one can add instead of multiplying. This is because $\log(x \times y) = \log(x) + \log(y)$. For example:

```
log(2*3)
```

```
## [1] 1.792
```

```
log(2) + log(3)
```

```
## [1] 1.792
```

So the joint log likelihood of the two data points is:

```
dnorm(5,mean=10,sd=2,log=TRUE)+dnorm(10,mean=10,sd=2,log=TRUE)
```

```
## [1] -6.349
```

Even more compactly:

```
sum(dnorm(c(5,10),mean=10,sd=2,log=TRUE))
```

```
## [1] -6.349
```

- Given the 10 data points above, calculate the maximum likelihood estimate (MLE) of the expectation.

- The sum of the log-likelihoods of the data-points x, using as the mean the MLE from the sample, and standard deviation 5.
- What is the sum of the log-likelihood if the mean used to compute the log-likelihood is `495.9`?
- Which value for the mean, the MLE or `495.9`, gives the higher log-likelihood?

2

Introduction to Bayesian data analysis

Recall Bayes' rule: When A and B are observable events, we can state the rule as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.1}$$

Given a vector of data y, Bayes' rule allows us to work out the posterior distributions of the parameters of interest, which we can repesent as the vector of parameters θ. This computation is achieved by rewriting (2.1) as (2.2). What is different here is that Bayes' rule is written in terms of probability distributions. Here, $p(\cdot)$ is a probability density, not the probability of a single event, which we represent above using $P(\cdot)$.

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} \tag{2.2}$$

The above statement can be rewritten in words as follows:

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Marginal Likelihood}} \tag{2.3}$$

The terms here have the following meaning. We elaborate on each point with an example below.

- The *Posterior*, $p(\theta|y)$ is the probability distribution of the parameters conditional on the data.
- The *Likelihood* is as described in chapter 1: it is the PMF (discrete case) or the PDF (continuous case) expressed a function of θ.
- The *Prior* is the initial probability distribution of the parameter, before seeing the data.

- The *Marginal Likelihood* was introduced in chapter 1 and standardizes the posterior distribution to ensure that the area under the curve of the distribution sums to 1, that is, it ensures that the posterior is a valid probability distribution.

An example will clarify all these terms, as we explain below.

2.1 Deriving the posterior using Bayes' rule: An analytical example

Recall our cloze probability example earlier. Participants are shown sentences like

"It's raining. I'm going to take the ..."

Ten participants are asked to complete the sentence. If 8 out of 10 participants complete the sentence with "umbrella," the estimated cloze probability or predictability (given the preceding context) would be $\frac{8}{10} = 0.8$. This is the maximum likelihood estimate of the probability of producing this word; we will designate the estimate with a "hat" on the parameter name: $\hat{\theta} = 0.8$.

Notice an important point here: one shortcoming of simply writing down the proportion in this way is that it ignores the uncertainty of our measurement: 0.8 could come from 10 participants ($\frac{8}{10}$), 100 participants ($\frac{80}{100}$), or $100,000$ participants ($\frac{80000}{100000}$). The uncertainty of the estimate 0.8 is different in each of these cases, and that is very relevant when drawing conclusions from data.

In the frequentist framework, the only thing we can characterize our uncertainty about is the **sampling distribution** of this parameter under imaginary repeated sampling; we can never talk about our uncertainty about the parameter's true value itself. Thus, for a sample size of 10, our uncertainty of the sampling distribution would be computed by calculating the sample variance σ^2 (here, $n \times \hat{\theta}(1-\hat{\theta}) = 10 \times 0.8 \times (1-0.8) = 1.6$), and then calculating the standard error: $\sigma/\sqrt{n} = 0.4$. Increasing the sample size will make this standard error smaller and smaller for the same estimated proportion of successes of 0.8. This increased precision is a statement about the uncertainty of the sampling distribution of θ under imaginary repeated sampling; it is not an estimate of the uncertainty of θ itself.

The Bayesian framework gives us the opportunity to talk directly about our uncertainty of the parameter itself, given the data. This is achieved by obtaining the posterior distribution of the parameter using Bayes' rule, as we show below.

2.1.1 Choosing a likelihood

Under the assumptions we have set up above, the responses follow a Binomial distribution, and so the PMF can be written as follows.

$$p(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{2.4}$$

where k indicates the number of times "umbrella" is given as an answer, and n the total number of answers given.

In a particular experiment that we carry out, if $n = 10$ and $k = 8$, these data are now a fixed quantity. The PMF above now becomes a function of θ, the likelihood function:

$$p(k=8|n=10,\theta) = \binom{n}{k}\theta^8(1-\theta)^2 \tag{2.5}$$

The above function is a now a continuous function of the value θ, which has possible values ranging from 0 to 1. Compare this to the PMF of the Binomial, which treats θ as a fixed value and defines a discrete distribution over the n+1 possible discrete values k that we can observe (the possible number of successes).

It is important to pause for a moment here to appreciate the fact that the PMF and the likelihood are the same function seen from different points of view. The only difference between the two is what is considered to be fixed and what is varying. The PMF treats data as varying from experiment to experiment and θ as fixed, whereas the likelihood function treats the data as fixed and the parameter θ as varying.

We now turn our attention back to our main goal, which is to find out, using Bayes' rule, the posterior distribution of θ given our data: $p(\theta|n, k)$. In order to use Bayes' rule to calculate this posterior distribution, we need to define a prior distribution over the parameter θ. In doing so, we are explicitly expressing our prior uncertainty about plausible values of θ.

2.1.2 Choosing a prior for θ

For the choice of prior for θ in the Binomial distribution, we need to assume that the parameter θ is a random variable that has a PDF whose range lies within [0,1], the range over which θ can vary (this is because θ represents a probability). The Beta distribution, which is a PDF for a continuous random variable, is commonly used as prior for parameters representing probabilities. One reason for this choice is that its PDF rangers over the interval $[0, 1]$. The other reason for this choice is that it makes the Bayes' rule calculation remarkably easy.

The Beta distribution has the following PDF.

$$p(\theta|a, b) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1} \tag{2.6}$$

The term $B(a, b)$ expands to $\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta$, and is a normalizing constant rhat ensures that the area under the curve sums to one.[1]

The Beta distribution's parameters a and b can be interpreted as expressing our prior beliefs about the probability of success; a represents the number of "successes", in our case, answers that are "umbrella" and b the

[1] In some textbooks, you may see the PDF of the Beta distribution with the normalizing constant $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ (the expression $\Gamma(n)$ is defined as (n-1)!):

$$p(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \tag{2.7}$$

These two statements for the Beta distribution are identical because $B(a, b)$ can be shown to be equal to $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

number of failures, the answers that are not "umbrella". Figure 2.1 shows the different Beta distribution shapes given different values of a and b.

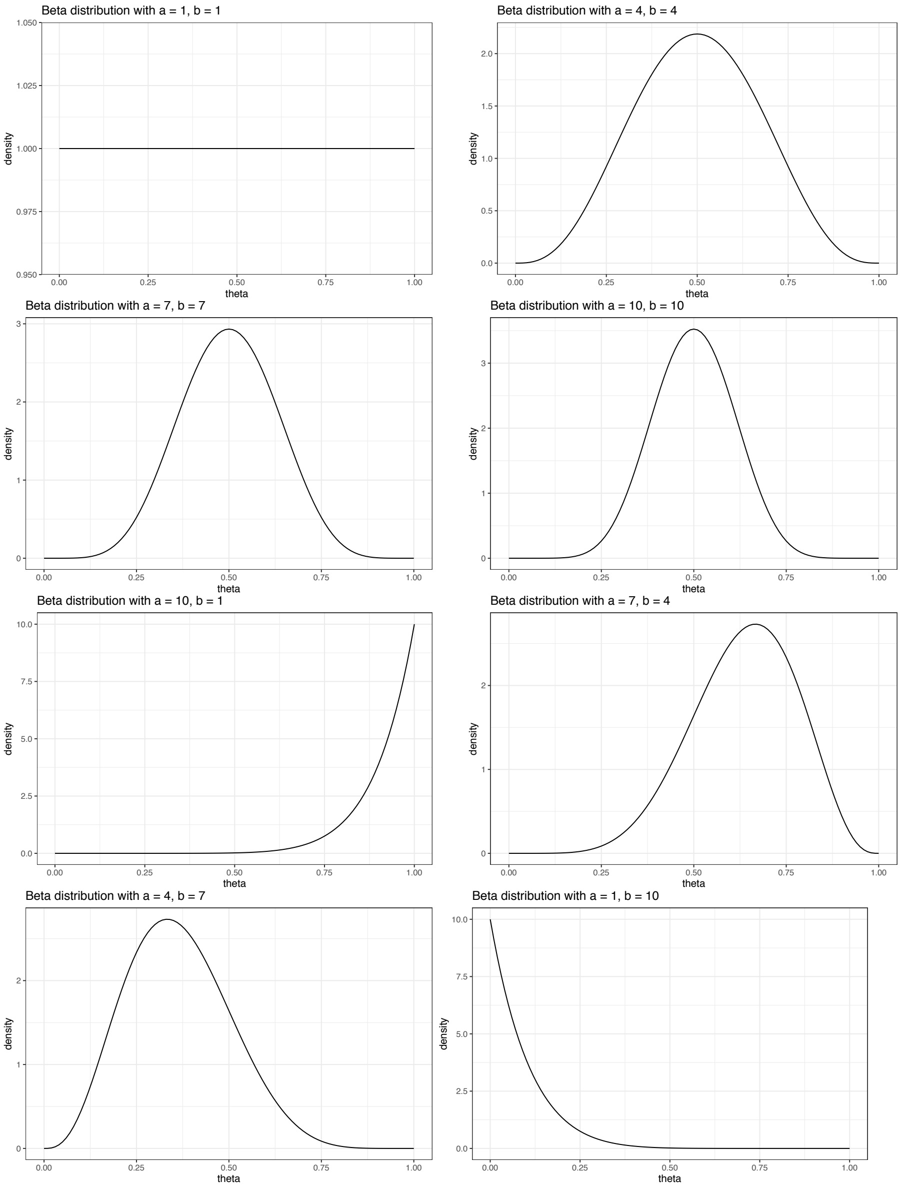

FIGURE 2.1: Examples of Beta distributions with different parameters.

As in the Binomial and Normal distributions that we saw in chapter 1, one can analytically derive the formulas for the expectation and variance of the Beta distribution. These are:

$$\mathrm{E}[X] = \frac{a}{a+b} \quad \mathrm{var}(X) = \frac{a \cdot b}{(a+b)^2(a+b+1)} \tag{2.8}$$

As an example, choosing $a = 4$ and $b = 4$ would mean that the answer "umbrella" is as likely as a different answer, but we are relatively unsure about this. We could express our uncertainty by computing the region over which we are 95% certain that the value of the parameter lies; this is the **95% credible interval**. For this, we would use the `qbeta` function in R; the parameters a and b are called `shape1` and `shape2` in `R`.

```
qbeta(c(0.025,0.975),shape1=4,shape2=4)
```

```
## [1] 0.1841 0.8159
```

If we were to choose $a = 10$ and $b = 10$, we would still be assuming that a priori the answer "umbrella" is just as likely as some other answer, but now our prior uncertainty about this mean is lower, as the 95% credible interval computed below shows.

```
qbeta(c(0.025,0.975),shape1=10,shape2=10)
```

```
## [1] 0.2886 0.7114
```

In Figure 2.1, we can see also the difference in uncertainty in these two examples graphically.

Which prior should we choose? In a real data analysis problem, the choice of prior would depend on what prior knowledge we want to bring into the analysis. If we don't have much prior information, we could use $a = b = 1$; this gives us a uniform prior. This kind of prior goes by various names: **non-informative prior**, or **weakly informative prior**. By contrast, if we have a lot of prior knowledge and/or a strong belief (e.g., based on a particular theory's predictions, or prior data) that θ has a particular range of plausible values, we can use a different set of a,b values to reflect our belief about the parameter. Notice in the above example that the larger our parameters a and b, the narrower the spread of the distribution; i.e., the lower our uncertainty about the mean value of the parameter.

For the moment, just for illustration, we choose the values $a = 4$ and $b = 4$ for the Beta prior. Then, our prior for θ is the following Beta PDF:

$$p(\theta) = \frac{1}{B(4,4)}\theta^3(1-\theta)^3 \tag{2.9}$$

Having chosen a likelihood, and having defined a prior on θ, we are ready to carry out our first Bayesian analysis to derive a posterior distribution for θ.

2.1.3 Using Bayes' rule to compute the posterior $p(\theta|n,k)$

Having specified the likelihood and the prior, we will now use Bayes' rule to calculate $p(\theta|n,k)$. Using Bayes' rule simply involves replacing the Likelihood and the Prior we defined above into the equation we saw earlier:

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Marginal Likelihood}} \tag{2.10}$$

Replace the terms for likelihood and prior into this equation:

$$p(\theta|n=10,k=8) = \frac{\left[\binom{10}{8}\theta^8 \cdot (1-\theta)^2\right] \times \left[\frac{1}{B(4,4)} \times \theta^3(1-\theta)^3\right]}{p(k=8)} \tag{2.11}$$

where $p(k=8)$ is $\int_0^1 p(k=8|n,\theta)p(\theta)\,d\theta$. This term will be a constant once the number of successes k is known; this is the marginal likelihood we encountered in chapter 1. In fact, once k is known, there are several constant values in the above equation; they are constants because none of them depend on the parameter of interest, θ. We can collect all of these together:

$$p(\theta|n=10,k=8) = \left[\frac{\binom{10}{8}}{B(4,4)\times p(k=8)}\right][\theta^8(1{-}\theta)^2{\times}\theta^3(1{-}\theta)^3] \tag{2.12}$$

The first term that is in square brackets, $\frac{\binom{10}{8}}{B(4,4)\times p(y)}$, is all the constants collected together, and is the normalizing constant we have seen before; it makes the posterior distribution $p(\theta|n=10,k=8)$ sum to one. Since it is a constant, we can ignore it for now and focus on the two other terms in the equation. Because we are ignoring the constant, we will now say that the posterior is proportional to the right-hand side.

to-do: introduce the idea of an unnormalized posterior here? see other suggestion elsewhere.

$$p(\theta|n=10,k=8) \propto [\theta^8(1-\theta)^2 \times \theta^3(1-\theta)^3] \tag{2.13}$$

A common way of writing the above equation is:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{2.14}$$

Resolving the right-hand side now simply involves adding up the exponents! In this example, computing the posterior really does boil down to this simple addition operation on the exponents.

$$p(\theta|n=10,k=8) \propto [\theta^{8+3}(1-\theta)^{2+3}] = \theta^{11}(1-\theta)^5 \tag{2.15}$$

The expression on the right-hand side corresponds to a Beta distribution with parameters $a=12$, and $b=6$. This becomes evidence if we rewrite the right-hand side such that it represents the core part of a Beta PDF. All that is missing is a normalizing constant which would make the area under the curve sum to one.

$$\theta^{11}(1-\theta)^5 = \theta^{12-1}(1-\theta)^{6-1} \tag{2.16}$$

This core part of any PDF or PMF is called the kernel of that distribution.

Without a normalizing constant, the area under the curve will not sum to one. Let's check this:

```
PostFun<-function(theta){
  theta^11 * (1-theta)^5
}
(AUC<-integrate(PostFun,lower=0,upper=1)$value)
```

```
## [1] 1.347e-05
```

So the area under the curve (AUC) is not 1—the posterior that we computed above is not a proper probability distribution.

All that is needed to make this into a proper probability distribution is to include a normalizing constant, which, according to the definition of the Beta distribution, would be $B(12, 6)$. This term is in fact the integral we computed above.

$$p(\theta|n = 10, k = 8) = \frac{1}{B(12, 6)}\theta^{12-1}(1 - \theta)^{6-1} \tag{2.17}$$

Now, this function will sum to one:

```
PostFun<-function(theta){
  theta^11 * (1-theta)^5/AUC
}
round(integrate(PostFun,lower=0,upper=1)$value,2)
```

```
## [1] 1
```

2.1.4 Summary of the procedure

To summarize, we started with a Binomial likelihood, multiplied it with the prior $\theta \sim Beta(4, 4)$, and obtained the posterior $p(\theta|n, k) \sim Beta(12, 6)$. The constants were ignored when carrying out the multiplication; we say that we computed the posterior **up to proportionality**. Finally, we showed how, in this simple example, the posterior can be rescaled to become a probability distribution, by including a proportionality constant.

The above example is a case of a **conjugate** analysis: the posterior on the parameter has the same form as the prior. The above combination of likelihood and prior is called the Beta-Binomial conjugate case. There are several other such combinations of Likelihoods and Priors that yield a posterior that has the same PDF as the prior on the parameter; some examples will appear in the exercises.

Formally, conjugacy is defined as follows:

DEFINITION Given the likelihood $p(y|\theta)$, if the prior $p(\theta)$ results in a posterior $y(\theta|y)$ that has the same form as $p(\theta)$, then we call $p(\theta)$ a conjugate prior.

For the Beta-Binomial case, we can derive a very general relationship between the likelihood, prior, and posterior. Given the Binomial likelihood up to proportionality (ignoring the constant) $\theta^k(1-\theta)^{n-k}$, and given the prior, also up to proportionality, $\theta^{a-1}(1-\theta)^{b-1}$, their product will be:

$$\theta^k(1-\theta)^{n-k}\theta^{a-1}(1-\theta)^{b-1} = \theta^{a+k-1}(1-\theta)^{b+n-k-1} \tag{2.18}$$

Thus, given a $Binomial(n,k|\theta)$ likelihood, and a $Beta(a,b)$ prior on θ, the posterior will be $Beta(a+k, b+n-k)$.

2.1.5 Visualizing the prior, likelihood, and the posterior

We established in the example above that the posterior is a Beta distribution with parameters $a = 12$, and $b = 6$. We visualize the likelihood, prior, and the posterior alongside each other in 2.2.

We can summarize the posterior distribution either graphically as we did above, or summarize it by computing the mean and the variance. The

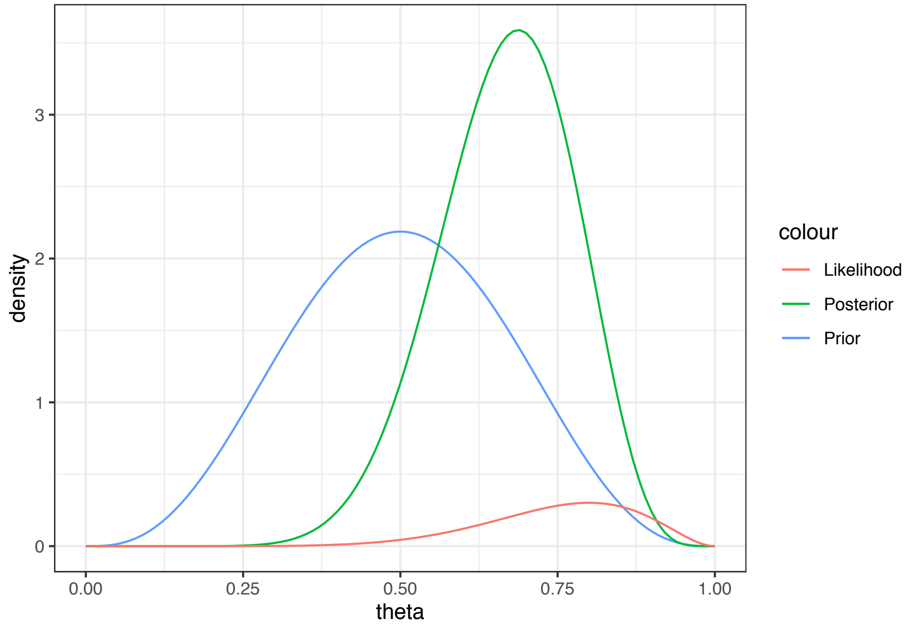

FIGURE 2.2: The likelihood, prior, and posterior in the Beta-Binomial example.

mean gives us an estimate of the Cloze probability of producing "umbrella" in that sentence (given the model, i.e., given the likelihood and prior):

$$\mathrm{E}[\hat{\theta}] = \frac{12}{12+6} = 0.67 \tag{2.19}$$

$$\mathrm{var}[\hat{\theta}] = \frac{12 \cdot 6}{(12+6)^2(12+6+1)} = .01 \tag{2.20}$$

We could also display the 95% credible interval, the range over which we are 95% certain the true value of θ lies, given the data and model.

```r
qbeta(c(0.025,0.975),shape1=12,shape2=6)
```

```
## [1] 0.4404 0.8579
```

Typically, we would summarize the results of a Bayesian analysis by displaying the posterior distribution of the parameter (or parameters) graphically, along with the above summary statistics: the mean, the standard deviation or variance, and the 95% credible interval. You will see many examples of such summaries later.

2.1.6 The posterior distribution is a compromise between the prior and the likelihood

Just for the sake of illustration, let's take four different Beta priors, each reflecting increasing certainty.

- Beta(a=2,b=2)
- Beta(a=3,b=3)
- Beta(a=6,b=6)
- Beta(a=21,b=21)

Each prior reflects a belief that $\theta = 0.5$, with varying degrees of (un)certainty. Given the general formula we developed above for the Beta-Binomial case, we just need to plug in the likelihood and the prior to get the posterior:

$$p(\theta|n,k) \propto p(k|n,\theta)p(\theta) \tag{2.21}$$

The four corresponding posterior distributios would be:

$$p(\theta \mid y,n) \propto [\theta^{8}(1-\theta)^{2}][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{10-1}(1-\theta)^{4-1} \tag{2.22}$$

$$p(\theta \mid y,n) \propto [\theta^{8}(1-\theta)^{2}][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{11-1}(1-\theta)^{5-1} \tag{2.23}$$

$$p(\theta \mid y,n) \propto [\theta^{8}(1-\theta)^{2}][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{14-1}(1-\theta)^{8-1} \tag{2.24}$$

$$p(\theta \mid y,n) \propto [\theta^{8}(1-\theta)^{2}][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{31-1}(1-\theta)^{23-1} \tag{2.25}$$

We can easily visualize each of these triplets of priors, likelihoods and posteriors. Use the Shiny app embedded below to visualize these different prior-likelihood combinations and look at the posterior in each case.

to-do: put in a shiny app that varies the a,b parameters and the amount of data, to show how the posterior is influenced by the data and the prior under different scenarios.

```
knitr::include_app("https://vasishth.shinyapps.io/AppTypeIPower",
  height = "500px")
```

If you vary the prior's certainty (held constant at $n = 10, k = 8$ in the above example), the posterior orients itself increasingly towards the prior. In general, we can say the following about the likelihood-prior-posterior relationship:

- The posterior distribution is a compromise between the prior and the likelihood.
- For a given set of data, the greater the certainty in the prior, the more heavily the posterior will be influenced by the prior mean.
- Conversely, for a given set of data, the greater the **un**certainty in the prior, the more heavily the posterior will be influenced by the likelihood.

Another important observation emerges if we increase the sample size from 10 to, say, $1,000,000$. Suppose we still get a sample mean of 0.8 here, so that $k = 800,000$. Now, the posterior mean will be influenced almost entirely by the sample mean. This is because, in the general form for the posterior $Beta(a + k, b + n - k)$ that we computed above, the n and k become very large relative to the a, b values, and dominate in determining the posterior mean.

Whenever we do a Bayesian analysis, it is good practice to check whether the parameter you are interested in estimating is sensitive to the prior specification. Such an investigation is called a **sensitivity analysis**. Later in this book, we will see many examples of sensitivity analyses in realistic data-analysis settings.

2.1.7 Incremental knowledge gain using prior knowledge

In the above example, we used an artificial example where we asked 10 participants to complete the sentence shown at the beginning of the chapter, and then we counted the number of times that they produced "umbrella" vs. some other word as a continuation. Given 8 instances of "umbrella", and using a $Beta(4,4)$ prior, we derived the posterior to be $Beta(12,6)$. We could now use this posterior as our prior for the next study. Suppose that we were to carry out a second experiment, again with 10 participants, and this time 6 produced "umbrella". We could now use our new prior (Beta(12,6)) to obtain an updated posterior. We have $a = 12, b = 6, n = 10, k = 6$. This gives us as posterior: $Beta(a+k, b+n-k) = Beta(12+6, 6+10-6) = Beta(18,10)$.

Now, if we were to pool all our data from the 20 participants that we have now, then we would have had as data $n = 20, k = 14$. Suppose that we keep our initial prior of $a = 4, b = 4$. Then, our posterior would be $Beta(4+14, 4+20-14) = Beta(18,10)$. This is exactly the same posterior that we got when first analyzed the first 10 participants' data, derived the posterior, and then used that posterior as a prior for the next 10 participants' data.

This toy example illustrates an important point that has great practical important for cognitive science. One can incrementally gain information about a research question by using information from previous studies and deriving a posterior, and then use that posterior as a prior. For practical examples from psycholinguistics showing how information can be pooled from previous studies, see Jäger et al. (2017) and Nicenboim et al. (2018). Vasishth and Engelmann (2020) illustrates an example of how the posterior from a previous study or collection of studies can be used to compute the posterior derived from new data. We return to this point in later chapters.

to-do: check that we do.

2.2 Summary of concepts introduced in this chapter

In this chapter, we learnt how to use Bayes' rule in the specific case of a Binomial likelihood, and a Beta prior on the θ parameter in the likelihood function. Our goal in any Bayesian analysis will follow the path we took in this simple example: decide on an appropriate likelihood function, decide on priors for all the parameters involved in the likelihood function, and using this model (i.e., the likelihood and the priors) derive the posterior distribution of each parameter. Then we draw inferences about our research question based on the posterior distribution of the parameter.

In the example discussed in this chapter, Bayesian analysis was easy. This was because we considered the simple conjugate case of the Beta-Binomial. In realistic data-analysis settings, our likelihood function will be very complex, and many parameters will be involved. Multiplying the likelihood function and the priors will become mathematically difficult or impossible. For such situations, we use computational methods to obtain samples from the posterior distributions of the parameters.

to-do: add summary

2.3 Further reading

2.4 Exercises

2.4.1 Deriving Bayes' rule

Let A and B be two observable events. P(A) is the probability that A occurs, and P(B) is the probability that B occurs. $P(A|B)$ is the conditional probability that A occurs given that B has happened. $P(A, B)$ is the joint probability of A and B both occurring.

You are given the definition of conditional probability:

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ where } P(B) > 0 \tag{2.26}$$

Using the above definition, and using the fact that $P(A, B) = P(B, A)$ (i.e., the probability of A and B both occurring is the same as the probability of B and A both occurring), derive an expression for $P(B|A)$. Show the steps clearly in the derivation.

2.4.2 Conjugate forms 1

2.4.2.1 Computing the general form of a PDF for a posterior

Suppose you are given data k consisting of the number of successes, coming from a $Binomial(n, \theta)$ distribution. Example data are shown below, generated with probability of success $\theta = 0.5$, just for illustration:

```r
## data:
k<-rbinom(n=1,size=10,prob=0.5)
k
```

```
## [1] 5
```

Here, n represents the number of trials, and k the number of successes. The above code and output is just an example, and is no longer relevant for the question below.

Given k successes in n trials coming from a Binomial distribution, we define a $Beta(a, b)$ prior on the parameter θ.

Write down the Beta distribution that represents the posterior, in terms of $a, b, n, and k$.

2.4.2.2 Practical application

We ask 10 yes/no questions from a participant, and the participant returns 0 correct answers. We assume a Binomial likelihood function for these data. Also assume a Beta(1,1) prior on the parameter θ, which represents the probability of success. Use the result you derived above to write down the posterior distribution of the θ parameter.

2.4.3 Conjugate forms 2

Suppose you have n independent and identically distributed data points from a distribution that has the likelihood function $f(x|\theta) = \theta(1-\theta)^{\sum_{i=1}^{n} x_i}$, where the data points x can have values 0,1,2,.... Let the prior on θ be Beta(a,b), a Beta distribution with parameters a,b. The posterior distribution is a Beta distribution with parameters a* and b*. Determine these parameters in terms of a, b, and $\sum_{i=1}^{n} x_i$.

2.4.4 Conjugate forms 3

The Gamma distribution is defined in terms of the parameters a, b: Ga(a,b). The probability density function is:

$$Ga(a,b) = \frac{b^a \lambda^{a-1} \exp\{-b\lambda\}}{\Gamma(a)} \tag{2.27}$$

We have data $x_1, \ldots, x_n$, with sample size n that is exponentially distributed. The exponential likelihood function is:

$$p(x_1, \ldots, x_n|\lambda) = \lambda^n \exp\{-\lambda \sum_{i=1}^{n} x_i\} \tag{2.28}$$

It turns out that if we assume a Ga(a,b) prior distribution and the above likelihood, the posterior distribution is a Gamma distribution. Find the parameters a' and b' of the posterior distribution.

2.4.5 Conjugate forms 4

2.4.5.1 a. Computing the posterior

This is a contrived example. Suppose we are modeling the number of times that a speaker says the word “I” per day. This could be of interest if we are studying, for example, how self-oriented a speaker is. The number of times x that the word is uttered in over a oarticular time period (here, one day) can be modeled by a Poisson distribution:

$$f(x \mid \theta) = \frac{\exp(-\theta)\theta^x}{x!} \tag{2.29}$$

where the rate θ is unknown, and the numbers of utterances of the target word on each day are independent given θ.

We are told that the prior mean of θ is 100 and prior variance for θ is 225. This information is based on the results of previous studies on the topic. We will use the Gamma(a,b) density (see previous question) as a prior for θ because this is a conjugate prior to the Poisson distribution.

- First, visualize the prior. a Gamma density prior for θ based on the above information.

[Hint: Note that we know that for a Gamma density with parameters a, b, the mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$. Since we are given values for the mean and variance, we can solve for a,b, which gives us the Gamma density.]

```r
x<-0:200
plot(x,dgamma(x,10000/225,100/225),type="l",lty=1,
     main="Gamma prior",ylab="density",
     cex.lab=2,cex.main=2,cex.axis=2)
```

- Next, derive the posterior distribution of the parameter θ up to proportionality, and write down the posterior distribution in terms of the parameters of a Gamma distribution.

2.4.5.2 b. Practical application

Suppose we know that the number of "I" utterances from a particular individual is $115, 97, 79, 131$. Use the result you derived above to obtain the posterior distribution. In other words, write down the a,b parameters of the Gamma distribution representing the posterior distribution of θ.

Plot the prior, likelihood, and the posterior alongside each other.

Now suppose you get one new data point: 200. Write down the updated posterior (the a,b parameters of the Gamma distribution) given this new data-point. Add the updated posterior to the plot you made above.

3

Important distributions

These distributions are used quite frequently in Bayesian data analyses, especially in psychology and linguistics applications. The Binomial and Poisson are discrete distributions, the rest are continuous. Each distribution comes with a family of `d-p-q-r` functions in R which allow us to compute the PDF/PMF, the CDF, the inverse CDF, and to generate random data. For example, the normal distribution's PDF is `dnorm`; the CDF and the inverse CDF are `pnorm` and `qnorm` respectively; and random data can be generated using `rnorm`.

The table below is adapted from https://github.com/wzchen/probability_cheatsheet, which is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

to-do: check that the notation is consistent with the main text's.

Distribution	PMF/PDF and Support	Expected Value	Variance
Binomial $Binomial(n, \theta)$	$P(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ $k \in \{0, 1, 2, \ldots n\}$	$n\theta$	$n\theta(1-\theta)$
Poisson $Pois(\lambda)$	$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ $k \in \{0, 1, 2, \ldots\}$	λ	λ
Uniform $Unif(a, b)$	$f(x) = \frac{1}{b-a}$ $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal $Normal(\mu, \sigma)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{(2\sigma^2)}}$ $x \in (-\infty, \infty)$	$\mu = \frac{\sum_{i=1}^n x_i}{n}$	$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$
Log-Normal $LogNormal(\mu, \sigma)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x - \mu)^2/(2\sigma^2)}$ $x \in (0, \infty)$	$\theta = e^{\mu + \sigma^2/2}$	$\theta^2(e^{\sigma^2} - 1)$
Beta Beta(a, b)	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x \in (0, 1)$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$
Exponential $Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}$ $x \in (0, \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma $Gamma(a, \lambda)$	$f(x) = \frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x \in (0, \infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$
Student-t $t(n)$ Cauchy is $t(1)$	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1 + x^2/n)^{-(n+1)/2}$ $x \in (-\infty, \infty)$	0 if $n > 1$	$\frac{n}{n-2}$ if $n > 2$

Bibliography

Blitzstein, J. K. and Hwang, J. (2014). *Introduction to probability*. Chapman and Hall/CRC.

Bürkner, P.-C. (2019). *brms: Bayesian Regression Models using 'Stan'*. R package version 2.8.0.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Fox, J. (2009). *A mathematical primer for social statistics*. Number 159. Sage.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, third edition.

Gill, J. (2006). *Essential mathematics for political and social research*. Cambridge University Press Cambridge.

Jäger, L. A., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Kolmogorov, A. N. (1933/2018). *Foundations of the Theory of Probability: Second English Edition*. Courier Dover Publications.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Morin, D. J. (2016). *Probability: For the Enthusiastic Beginner*. Createspace Independent Publishing Platform.

Nicenboim, B., Roettger, T. B., and Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, 70:39–55.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vasishth, S. and Engelmann, F. (2020). Sentence comprehension as a cognitive process: A computational approach. Under contract with Cambridge University Press.