Shravan Vasishth, Daniel Schad, Audrey Bürki, Reinhold Kliegl

Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction

Dedicated to ...

Contents

Preface **v**

0.1 Prerequisites v
0.2 How to read this book vi
0.3 Online materials vi
0.4 Software needed vi
0.5 Acknowledgements vii

About the Authors **ix**

1 Introduction **1**

1.1 Discrete random variables: An example using the Binomial distribution 1
1.1.1 The mean and variance of the Binomial distribution 3
1.1.2 What information does a probability distribution provide? 5
1.2 Continuous random variables: An example using the Normal distribution 8
1.3 Summary of useful R functions relating to distributions 13
1.4 *Summary of random variable theory 13
1.5 Summary of concepts introduced in this chapter 15
1.6 Further reading 15
1.7 Exercises 15
1.7.1 Practice using the `pnorm` function 15
1.7.2 Practice using the `qnorm` function 16
1.7.3 Practice using `qt` 16
1.7.4 Maximum likelihood estimation 1 17
1.7.5 Maximum likelihood estimation 2 17

2 Introduction to Bayesian data analysis **21**

2.1 Deriving the posterior using Bayes' rule: An analytical example . . . 22
2.1.1 Choosing a likelihood . . . 23
2.1.2 Choosing a prior for θ . . . 25
2.1.3 Using Bayes' rule to compute the posterior $p(\theta|n, k)$ . . . 28
2.1.4 Summary of the procedure . . . 30
2.1.5 Visualizing the prior, likelihood, and the posterior . . . 31
2.1.6 The posterior distribution is a compromise between the prior and the likelihood . . . 33
2.1.7 Incremental knowledge gain using prior knowledge . . . 34
2.2 Summary of concepts introduced in this chapter . . . 35
2.3 Further reading . . . 36
2.4 Exercises . . . 36
2.4.1 Deriving Bayes' rule . . . 36
2.4.2 Conjugate forms 1 . . . 37
2.4.3 Conjugate forms 2 . . . 38
2.4.4 Conjugate forms 3 . . . 38
2.4.5 Conjugate forms 4 . . . 38

3 Important distributions 41

Preface

This book is intended to be a relatively complete introduction to the application of linear mixed models in areas related to linguistics and psychology; throughout, we use the programming language R. Our target audience is cognitive scientists (e.g., linguists and psychologists) who carry out behavioral experiments, and who are interested in the foundational ideas behind modern statistical methodology from the ground up and in a principled manner.

Many excellent introductory textbooks already exist that discuss data analysis in great detail. Our book is different from existing books 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 complexities that are distinct from the problems one encounters when analyzing observational data. 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 experiment design analysis and model validation using simulated data.

0.1 Prerequisites

This book assumes high school arithmetic and algebra. We also expect that the reader already knows basic constructs in the programming lan-

guage R (R Core Team, 2019), such as writing for-loops. For newcomers to R, we provide a quick introduction in the appendix that covers all the constructs used in the book. For those lacking background in R, there are many good online resources on R that they can consult as needed. Examples are: R for data science[1], and Efficient R programming[2].

provide comprehensive book recommendations

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/frequentist_book. The exercise solutions are provided there as a pdf (to-do).

0.4 Software needed

Before you start, please install

[1] https://r4ds.had.co.nz/
[2] https://csgillespie.github.io/efficientR/

- R[3] (and RStudio[4] or any other IDE)
- 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).

```
set.seed(42)
library(MASS)
##be careful to load dplyr after MASS
library(dplyr)
library(purrr)
library(readr)
library(extraDistr)
library(ggplot2)
```

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) at the University of Potsdam. 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.

Shravan Vasishth, Daniel Schad, Audrey Bürki, Reinhold Kliegl,
Potsdam, Germany

[3] https://cran.r-project.org/
[4] https://www.rstudio.com/

About the Authors

Shravan Vasishth (http://vasishth.github.io) is professor of Psycholinguistics at the University of Potsdam, Germany. He holds the chair for Psycholinguistics and Neurolinguistics (Language Processing). 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 (2000-2002) and a PhD in Linguistics (1997-2002) from the Ohio State University, Columbus, USA, and an MSc in Statistics (2011-2015) from the School of Mathematics and Statistics, University of Sheffield, UK. He is a professional member of the Royal Statistical Society (GradStat ID: 128307), a member of the International Society for Bayesian Analysis, and a lifetime member of the Linguistic Society of America (LSA). He is on the editorial board of the Linguistic Society of America flagship journal Language as their statistics consultant for journal submissions. His research focuses on computational modeling of sentence processing in unimpaired and impaired populations, and the application of mathematical, computational, experimental, and statistical methods (particularly Bayesian methods) in linguistics and psychology. He runs an annual summer school, Statistical Methods in Linguistics and Psychology (SMLP): vasishth.github.io/smlp. He regularly teaches short courses on statistical data analysis (Bayesian and frequentist methods).

Daniel J. Schad (https://danielschad.github.io/) is an assistant professor in the department of Cognitive Science and AI at Tilburg University. He studied Psychology at the University of Potsdam, Germany, and at the University of Michigan, Ann Arbor, USA. He did a PhD in Cognitive Psychology at the University of Potsdam, working on computational models of eye-movement control and on mindless reading. He then did a five-year post-doc in the novel field of Computational Psychiatry at the Char-

ité, Universität Berlin, Germany (partly also at the University of Potsdam), with research visits at the ETH Zürich, Switzerland, and the University College London, UK, working on model-free and model-based decision-making and Pavlovian-instrumental transfer in alcohol dependence, and on the cognitive and brain mechanisms underlying Pavlovian conditioning. He has worked as a postdoctoral researcher at the University of Potsdam, conducting research on quantitative methods in Cognitive Science, including contrasts, properties of significance tests, Bayesian Workflow, and Bayes factor analyses.

Audrey Bürki (https://danielschad.github.io/) leads a research group at the University of Potsdam, Germany. Her research interests are… to-do

Reinhold Kliegl (https://www.uni-potsdam.de/de/trainingswissenschaft/mitarbeiter/rkliegl.html) is a senior professor in the Division of Training and Movement Science, at the University of Potsdam, Germany. His research interests are… to-do

1

Introduction

In linguistics and psychology, typical data-sets involve *continuous* dependent measures such as reading times or reaction times in milliseconds and EEG signals in microvolts, or *discrete* dependent measures such as acceptability ratings on a Likert scale (for example, ranging from 1 to 7), and binary grammaticality judgements. Whenever we fit a model to such data, we usually make some assumptions about how these measurements were *generated*. In particular, we usually assume that our observed measurements are coming from a particular *probability mass function* (when the data are discrete in nature) *probability density function* (when the data are continuous). Behind these PMFs and PDFs lies the concept of a random variable; as will become apparent in this chapter, it is extremely useful to be able to think about data in terms of the random variable assumed. It is important to consider the two cases, discrete and continuous, separately.

1.1 Discrete random variables: An example using the Binomial distribution

Imagine that our data come from a grammaticality judgement task, and that the responses from participants are a sequence of 1's and 0's, where 1 represents the judgment "grammatical", and 0 represents the judgement "ungrammatical". Assume also that each response, coded as 1 or 0, is generated independently from the others. We can simulate such a sequence of 1s and 0s in R. Here is a case where we run the same experiment 20 times (the sample size is 10 each time).

```
##  [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 $p(Y)$ called the **Binomial distribution**.[1]

For discrete random variable, the probability distribution $p(y)$ 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: $0, \ldots, 10$ successes. Which of these outcomes is most probable depends on a parameter in the Binomial distribution that represents the probability of success. We will call this parameter θ. The left-hand side plot in Figure 1.1 shows an example of a Binomial PMF with 10 trials and the parameter θ with value 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.

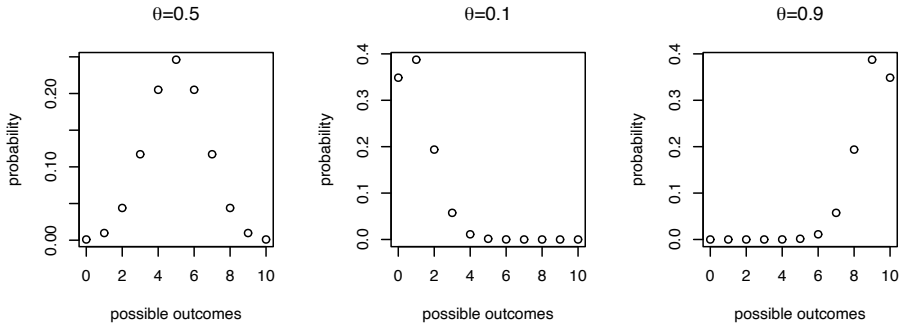


FIGURE 1.1: Probability mass functions of a binomial distribution assuming 10 trials, with 50%, 10%, and 90% probability of success.

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

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

$$\text{Binomial}(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{1.1}$$

[1] When an experiment consists of only a single trial (i.e., we can have a total number of only 0 or 1 successes), $p(Y)$ is called a **Bernoulli distribution**.

Here, n represents the total number of trials, k the number of successes, 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.

1.1.1 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 $Var(Y) = n\theta(1-\theta)$, respectively.

Of course, we always know n (because we decide on the number of trials ourselves), but 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, $\widehat{\theta} = k/n$. The quantity $\widehat{\theta}$ is the observed proportion of successes, and is called the **maximum likelihood estimate** of the true (but unknown mean). 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
$hattheta(1-\widehat{\theta})$. These estimates are then used for statistical inference.

What does the term "maximum likelihood estimate" mean? The term **likelihood** refers to the value of the Binomial distribution function for a particular value of θ, once we have observed some data. 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:

$$\text{Binomial}(k = 7 | n = 10, \theta) = \binom{10}{7} \theta^7 (1-\theta)^{10-7} \tag{1.2}$$

Once we have observed the data, 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:

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

The vertical bar notation above should be read as saying that, given some data y (which in the binomial case will be k "successes" in n trials), the function returns a value for different values of θ.

If we now plot this function for all possible values of θ, we get the plot shown in Figure 1.2.

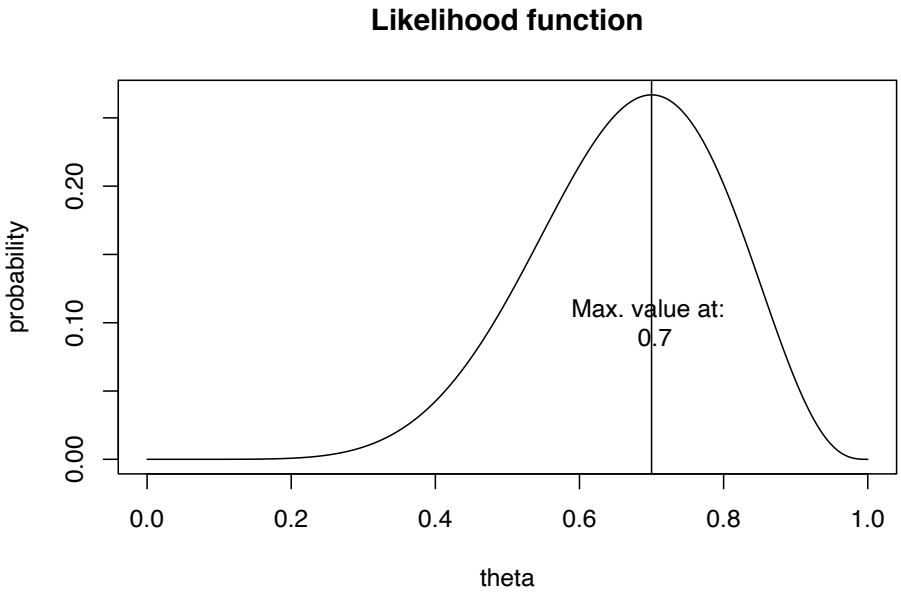


FIGURE 1.2: The likelihood function for 7 successes out of 10.

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 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.

1.1.2 What information does a probability distribution provide?

What good is a probability mass function? We consider this question next.

1.1.2.1 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 in the next section).

1.1.2.2 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:

$$\sum_{k=0}^{2} \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{1.3}$$

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.

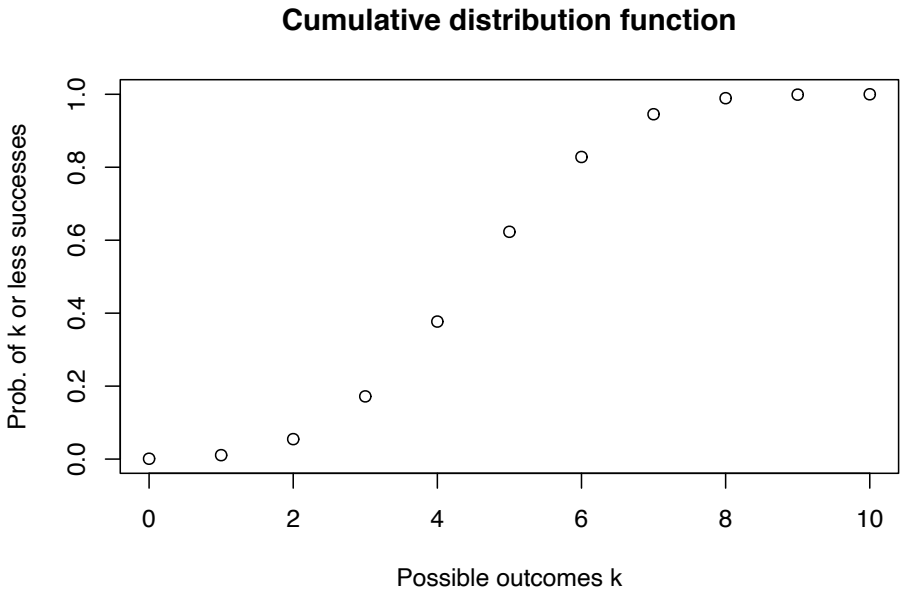


FIGURE 1.3: The cumulative distribution function for a binomial distribution assuming 10 trials, with 50% probability of success.

1.1.2.3 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`. 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:

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

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

1.1.2.4 Generate random data from a Binomial(n, θ) distribution

We can generate random 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(10,size=1,prob=0.5)
```

```
##  [1] 1 0 1 1 0 1 0 1 0 1
```

The above code generates a sequences of 1's and 0's. 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
```

1.2 Continuous random variables: An example using the Normal distribution

We will now revisit the idea of a 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 probability density function (PDF) of the Normal distribution is defined as follows:

$$Normal(y|\mu,\sigma) = f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{1.4}$$

Here, μ is the true mean, and σ is the true standard deviation of the Normal distribution that the reading times have been sampled from.

We can visualize 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`). See Figure 1.4. It is clear from the figure that these are three different ways of looking at the same information.

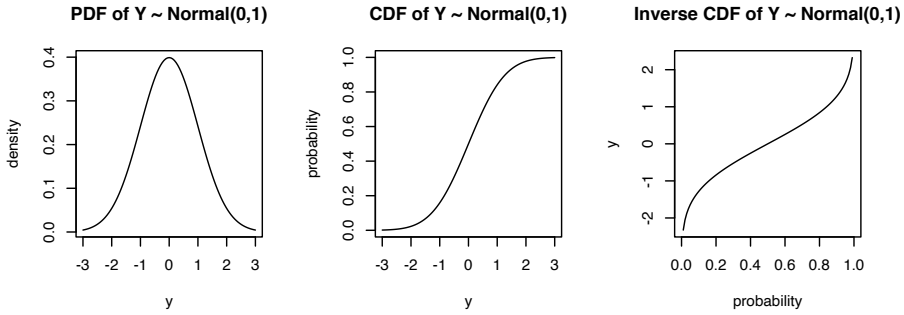


FIGURE 1.4: The PDF, CDF, and inverse CDF for the $Normal(\mu = 0, \sigma = 1)$.

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?** We can compute the probability of the random variable lying between 1 and minus infinity:

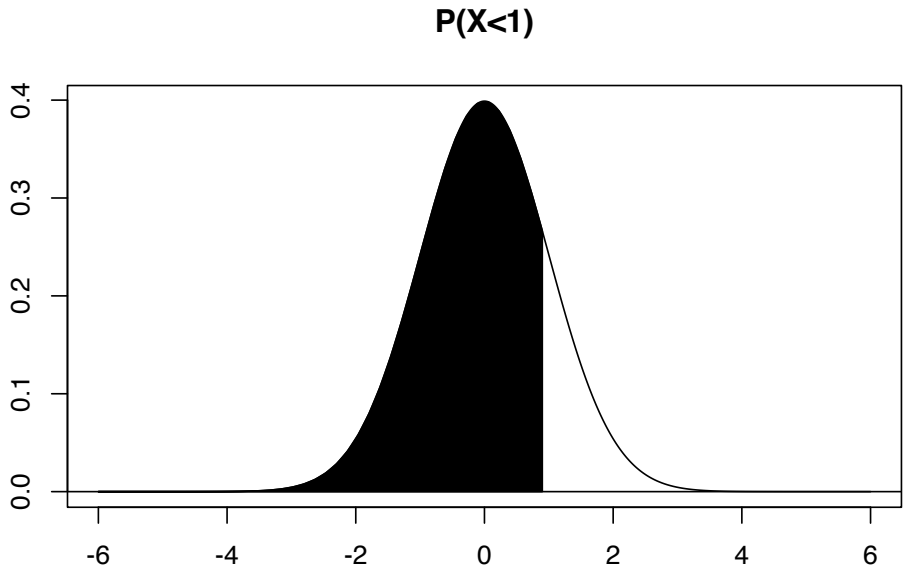
```
pnorm(1,mean=0,sd=1)-pnorm(-Inf,mean=0,sd=1)
```

```
## [1] 0.8413
```

```
## function for plotting area under curve:
plot.prob<-function(x,
                            x.min,
                            x.max,
                            prob,
                            mean,
                            sd,
                            gray.level,main){

        plot(x,dnorm(x,mean,sd),
                     type = "l",xlab="",
             ylab="",main=main)
        abline(h = 0)

## shade X<x
    x1 = seq(x.min, qnorm(prob), abs(prob)/5)
    y1 = dnorm(x1, mean, sd)

    polygon(c(x1, rev(x1)),
            c(rep(0, length(x1)), rev(y1)),
            col = gray.level)
}

shadenormal<-
function (prob=0.5,
          gray1="black",
          x.min=-6,
          x.max=abs(x.min),
          x = seq(x.min, x.max, 0.01),
          mean=0,
          sd=1,main="P(X<0)")
{

     plot.prob(x=x,x.min=x.min,x.max=x.max,
               prob=prob,
                      mean=mean,sd=sd,
```

```
     gray.level=gray1,main=main)
}
```

```
shadenormal(prob=0.84134,main="P(X<1)")
```



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 $P(a < Y < b)$. Also, notice that $P(a < Y < b)$ and $P(a \leq Y \leq b)$ will be the same probability, because of the fact that $P(Y = a)$ or $P(Y = b)$ both equal 0.

- **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 are

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.5}$$

$$Var(y) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} \tag{1.6}$$

For example, you could 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] 482.2
```

```
## [1] 13365
```

Again, depending on the sample size, the sample mean and sample variance may or may not be close to the true values of the respective parameters, despite the fact that these are maximum likelihood estimates.

This completes our informal discussion of random variable theory. We now summarize what we have learnt so far.

1.3 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(n,θ)	Normal(μ, σ)
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.

1.4 *Summary of random variable theory

We can summarize the above informal concepts very compactly if we restate them in mathematical form. A mathematical statement has the advantage not only of brevity but also that it reduces ambiguity.

Formally, a random variable Y is defined as a function from a sample space of possible outcomes S to the real number system: $Y : S \to \mathbb{R}$. The random variable associates to each outcome $\omega \in S$ exactly one number $Y(\omega) = y$. S_Y is all the y's (all the possible values of Y, the support of Y). I.e., $y \in S_Y$.

Every random variable Y has associated with it a probability mass (distribution) function (PMF, PDF). I.e., PMF is used for discrete distribu tions

and PDF for continuous distributions. The PMF/PDF maps every element of S_Y to a value between 0 and 1.

$$p_Y : S_Y \to [0, 1] \tag{1.7}$$

Probability mass functions (discrete case) and probability density functions (continuous case) are functions that assign probabilities or relative frequencies to all events in a sample space.

The expression

$$Y \sim f(\cdot) \tag{1.8}$$

will be used to mean that the random variable Y has pdf/pmf $f(\cdot)$. For example, if we say that $Y \sim Binomial(n, \theta)$, then we are asserting that the PMF is:

$$\text{Binomial}(k|n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \tag{1.9}$$

If we say that $Y \sim Normal(\mu, \sigma)$, we are asserting that the PDF is

$$Normal(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{yx - \mu)^2}{2\sigma^2}\right) \tag{1.10}$$

The **cumulative distribution function** or CDF is defined as follows:

For discrete distributions, the probability that Y is less than a is written:

$$P(Y < a) = F(Y < a) = \sum_{-\infty}^{a} f(y) \tag{1.11}$$

For continuous distributions, the summation symbol $\sum$ above becomes the summation symbol for the continuous case, which is the integral $\int$. The upper and lower bounds are marked by adding a subscript and a superscript on the integral. For example, if we want the area under the curve between points a and b for some function $f(y)$, we write $\int_b^a f(y)\, dy$. So, if we want the probability that Y is less than a, we would write:

$$P(Y < a) = F(Y < a) = \int_{-\infty}^{a} f(y)\, dy \tag{1.12}$$

The above integral is simply summing up the area under the curve between the points $-\infty$ and a; this gives us the probability of observing a or a value smaller than a.

A final point here is that we can go back and forth between the PDF and the CDF. If the PDF is $f(y)$, then the CDF that allows us to compute quantities like $P(Y < b)$ is just the integral $F(Y < b) = \int_{-\infty}^{b} f(y)\, dy$. If we differentiate the CDF, we get the PDF back: $d(F(y))/dy = f(y)$.

1.5 Summary of concepts introduced in this chapter

to-do: add summary

1.6 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.

1.7 Exercises

1.7.1 Practice using the `pnorm` function

1.7.1.1 Part 1

Given a normal distribution with mean `146` and standard deviation `101`, use the pnorm function to calculate the probability of obtaining values between `254` and `-61` from this distribution.

1.7.1.2 Part 2

Calculate the following probabilities. Given a normal distribution with mean `52` and standard deviation `2`, what is the probability of getting

- a score of `41` or less
- a score of `41` or more
- a score of `53` or more

1.7.1.3 Part 3

Given a normal distribution with mean `50` and standard deviation `4`, what is the probability of getting

- a score of `45` or less.
- a score between `47` and `53`.
- a score of `mu+1` or more.

1.7.2 Practice using the `qnorm` function

1.7.2.1 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 `0.01`.
- the area (the probability) to the left of the upper boundary is `0.72`.

1.7.2.2 Part 2

Given a normal distribution with mean `53.755` and standard deviation `1.014`. There exist two quantiles, the lower quantile q1 and the upper quantile q2, that are equidistant from the mean `53.755`, such that the area under the curve of the Normal probability between q1 and q2 is `95`%. Find q1 and q2.

1.7.3 Practice using `qt`

Take an independent random sample of size `145` from a normal distribution with mean `142`, and standard deviation `62`. 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 `140.391` and the sample standard deviation `56.407`.

- 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.

1.7.4 Maximum likelihood estimation 1

Given the data point `14.771`. 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 `14.771` assuming a mean of `12` and standard deviation 5.
- the likelihood of the data point `14.771` assuming a mean of `11` and standard deviation 5.
- the likelihood of the data point `14.771` assuming a mean of `10` and standard deviation 5.
- the likelihood of the data point `14.771` assuming a mean of `9` and standard deviation 5.

1.7.5 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] 492 484 502 497 503 487 490 511 504 506
```

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 `495.6`?
- Which value for the mean, the MLE or `495.6`, 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:

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

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. It is not a probability distribution as the area under the curve doesn't sum to one (but it could be made into a probability distribution through a normalization constant).

- The *Prior* is the initial probability distribution of the parameter, before seeing the data.
- The *Average Likelihood,* also called evidence or probability of the data, 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

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 onlt two possible outcomes, a success or a failure, and there is some true unknown probability θ of success that we want to estimate.

One way to empirically estimate this probability is to carry out a so-called cloze task. In a cloze task, participants are asked to complete a 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). If 8 out of 10 participants complete the context 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 prob-

ability of producing this word; we will designate the estimate with a "hat" on the parameter name: $\widehat{\theta} = 0.8$.

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, $n \times \widehat{\theta}(1 - \widehat{\theta}) = 10 \times 0.8 \times (1 - 0.8) = 1.6$), 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. By contrast, the Bayesian framework gives us the opportunity to think 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" was given as an answer, and n the total number of answers given. If $n = 10$ and $k = 8$, then the likelihood function is:

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

Notice that this is a function of the continuous value θ, which has possible values ranging from 0 to 1. Compare this to the Binomial distribution, which is a discrete distribution over the n+1 discrete values k, the possible number of successes. Because the likelihood function depends only on θ, we will sometimes abbreviate it as $\mathcal{L}(\theta)$. Notice also that although the likelihood function specifies a probability distribution for the different possible values of θ, is not a probability distribution in the sense that the area under the curve does not integrate to 1. We can quickly establish that this is true:

```
## Define the likelihood function:
LikFun<-function(theta){
  choose(10,8)*theta^8*(1-theta)^(10-8)
}
## compute the area under the curve:
integrate(LikFun,lower=0,upper=1)$value
```

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

Notice, however, that one can simply add a constant c to the function such that the area under the curve *does* integrate to 1. We just established above that:

$$\int_0^1 \mathcal{L}(\theta)\, d\theta = 0.09091 \tag{2.6}$$

We can simply solve for c in the following equation:

$$c \int_0^1 \mathcal{L}(\theta)\, d\theta = 1 \Rightarrow c \times 0.09091 = 1 \Rightarrow c = \frac{1}{0.09091} \tag{2.7}$$

This c is called a normalization constant because it ensures that the function sums to 1. This little exercise illustrates that, at least in the example above, it is possible to make a function become a proper probability distribution by adding a normalization constant.

We 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, we need to assume 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). In other words, we need a random variable with support in the range [0,1]. 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 has support over $[0, 1]$. The other reason for this choice will soon become apparent.

The Beta distribution has the following PDF.

$$p(\theta|a, b) = \frac{1}{B(a, b)}\theta^{a-1}(1 - \theta)^{b-1} \tag{2.8}$$

The term $B(a, b)$ expands to $\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta$, and is a normalizing constant, as discussed above. 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)!):

$$p(\theta|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1 - \theta)^{b-1} \tag{2.9}$$

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

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 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.

As in the Binomial and Normal distributions that we saw in chapter 1, one

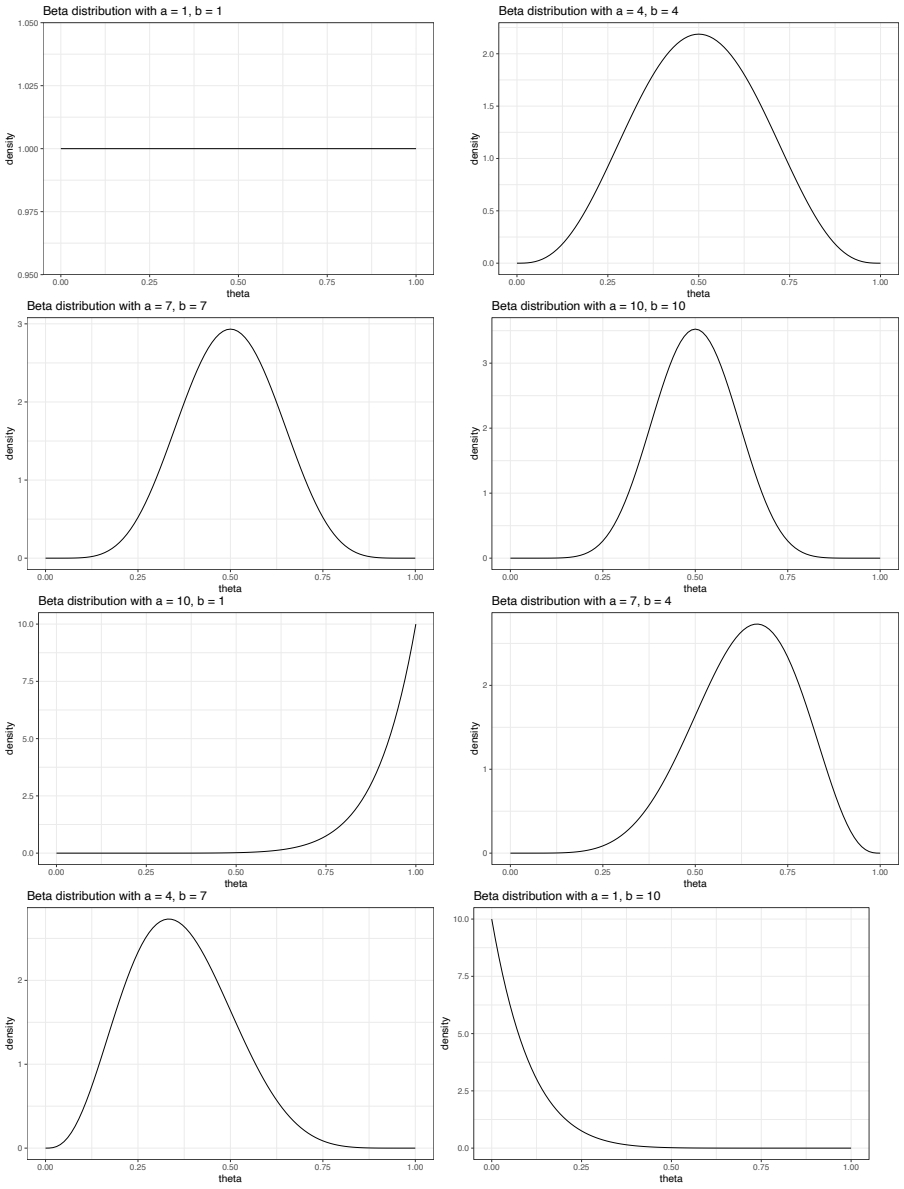


FIGURE 2.1: Examples of Beta distributions with different parameters.

can analytically derive the formulas for the expectation and variance of the Beta distribution. These are:

$$\mathrm{E}[X] = \frac{a}{a+b} \quad \mathrm{var}(X) = \frac{a \cdot b}{(a+b)^2(a+b+1)} \tag{2.10}$$

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: **uninformative prior**, **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 that the larger our param-

eters 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:

$$p(\theta) = \frac{1}{B(4,4)}\theta^3(1-\theta)^3 \tag{2.11}$$

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:

$$\text{Posterior} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Average Likelihood}} \tag{2.12}$$

Replace the terms for likelihood and prior into this equation:

$$p(\theta|n=10,k=8) = \frac{\left[\binom{10}{8}\theta^8 \cdot (1-\theta)^2\right] \times \left[\frac{1}{B(4,4)} \times \theta^3(1-\theta)^3\right]}{p(k=8)} \tag{2.13}$$

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:

$$p(\theta|n=10,k=8) = \left[\frac{\binom{10}{8}}{B(4,4) \times p(y)}\right] [\theta^8(1-\theta)^2 \times \theta^3(1-\theta)^3] \tag{2.14}$$

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 familiar 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 always compute this constant, as shown earlier), 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.

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

Notice that we are now stating the posterior as proportional to the right-hand side. This is a common way of writing the posterior:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{2.16}$$

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.

$$p(\theta|n = 10, k = 8) \propto [\theta^{8+3}(1-\theta)^{2+3}] = \theta^{11}(1-\theta)^5 \tag{2.17}$$

The expression on the right-hand side corresponds to a Beta distribution with parameters $a = 12$, and $b = 6$. This is because we can rewrite the right-hand side such that it represents the kernel of a Beta PDF:

$$\theta^{11}(1-\theta)^5 = \theta^{12-1}(1-\theta)^{6-1} \tag{2.18}$$

to-do: introduce the idea of an unnormalized posterior here?

Let's check that the area under the curve in this function does not sum up to one:

```
PostFun<-function(theta){
  theta^11 * (1-theta)^5
```

```
}
(AUC<-integrate(PostFun,lower=0,upper=1)$value)
```

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

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 (the marginal likelihood) we computed above.

$$p(\theta|n = 10, k = 8) = \frac{1}{B(12, 6)}\theta^{12-1}(1 - \theta)^{6-1} \tag{2.19}$$

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 $\theta|n, k\, Beta(12, 6)$. 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:

$$\theta^k(1-\theta)^{n-k}\theta^{a-1}(1-\theta)^{b-1} = \theta^{a+k-1}(1-\theta)^{b+n-k-1} \tag{2.20}$$

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.

```
## Warning: `mapping` is not used by stat_function()
```

```
## Warning: `mapping` is not used by stat_function()
```

```
## Warning: `mapping` is not used by stat_function()
```

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

$$\mathrm{E}[\widehat{\theta}] = \frac{12}{12+6} = 0.67 \tag{2.21}$$

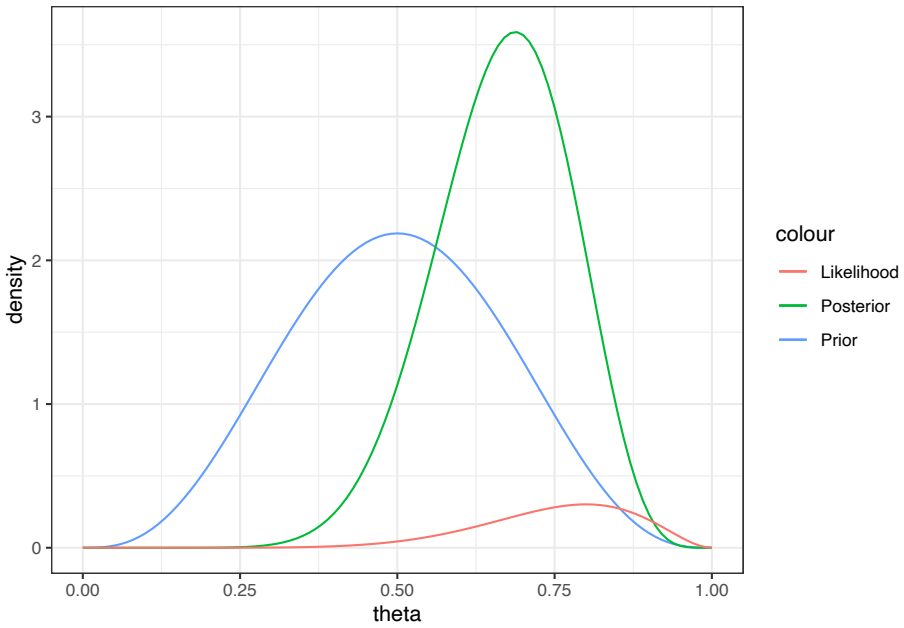


FIGURE 2.2: The likelihood, prior, and posterior in the Beta-Binomial example.

$$\text{var}[\hat{\theta}] = \frac{12 \cdot 6}{(12+6)^2(12+6+1)} = .01 \tag{2.22}$$

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.

```
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 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:

$$p(\theta|n,k) \propto p(k|n,\theta)p(\theta) \tag{2.23}$$

The four corresponding posterior distributios would be:

$$p(\theta \mid y,n) \propto [\theta^8(1-\theta)^2][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{10-1}(1-\theta)^{4-1} \tag{2.24}$$

$$p(\theta \mid y,n) \propto [\theta^8(1-\theta)^2][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{11-1}(1-\theta)^{5-1} \tag{2.25}$$

$$p(\theta \mid y,n) \propto [\theta^8(1-\theta)^2][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{14-1}(1-\theta)^{8-1} \tag{2.26}$$

$$p(\theta \mid y,n) \propto [\theta^8(1-\theta)^2][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{31-1}(1-\theta)^{23-1} \tag{2.27}$$

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 relatively vague 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 poste-

rior 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:

$$P(A|B) = \frac{P(A, B)}{P(B)} \text{ where } P(B) > 0 \tag{2.28}$$

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 a vector of data x consisting of 1's and 0's, coming from a Binomial(n,θ) distribution. 1 represents success, and 0 failure. Example data are shown below, generated with probability of success $\theta = 0.5$, just for illustration:

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

```
##  [1] 1 1 1 0 0 1 1 0 1 0
```

```
## k:
sum(x)
```

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

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:

$$Ga(a,b) = \frac{b^a \lambda^{a-1} \exp\{-b\lambda\}}{\Gamma(a)} \tag{2.29}$$

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

$$p(x_1, \ldots, x_n|\lambda) = \lambda^n \exp\{-\lambda \sum_{i=1}^{n} x_i\} \tag{2.30}$$

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:

$$f(x \mid \theta) = \frac{\exp(-\theta)\theta^x}{x!} \tag{2.31}$$

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.]

```
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.

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.

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.

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.